COMPUTATION
AND
PROBLEM SOLVING
IN
UNDERGRADUATE PHYSICS
Second Edition

• IDL
  MATLAB
  OCTAVE
• PYTHON
  MAXIMA
  MAPLE
• MATHEMATICA
  FORTRAN
  C
• NUMERICAL RECIPES
  LSODE
  MUDPACK
• \LaTeX
  TGIF

DAVID M. COOK
Department of Physics
Lawrence University
711 E Boldt Way, SPC 24
Appleton, Wisconsin 54911

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

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

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

Note: Usually with a second edition of a book, the preface to the first edition 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

---

1The 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 TEX while another may include the generic components and the components that focus on MATLAB, Mathematica, and FORTRAN (including Numerical Recipes). The table of contents and index contain only entries from the included chapters and sections. To facilitate communication among users of different versions, however, chapter and section numbers and the numbers identifying package-independent exercises are preserved in all versions. In a version that does not include FORTRAN, for example, the FORTRAN sections will be omitted from Chapters 9, 10, 11, 13, 14, and 15 and Chapters 12 and 16—for which FORTRAN is prerequisite—will be omitted altogether. In addition, FORTRAN-specific exercises will be omitted from the end-of-chapter exercises. Because of version-specific omissions such as those just described, there will therefore be gaps in the chapter, section, and exercise numbers in any version that does not include all options. In contrast, within each chapter, equation numbers, figure numbers, table numbers, and footnote numbers advance from one without gaps, and page numbers run continuously from the beginning of the book to the end. In consequence, the numbers assigned to identical equations, figures, tables, footnotes, and pages may differ from version to version, but the numbers assigned to chapters and sections with identical titles and the numbers assigned to identical exercises will be the same in all versions (and will have gaps reflecting omitted chapters, sections, and exercises). Such flexibility would be impossible were we not able to exploit features of \LaTeX, including the particular capabilities of the \texttt{ifthen} and \texttt{imakeidx} packages, to assemble the PDF files that permit the selected assemblage subsequently to be viewed on a screen or printed.

Even among sites that use the same spectrum of hardware and software, however, some aspects of local environments remain unique to individual sites. Local rules of citizenship; the features and elementary resources of the local operating system; local practices and policies governing structuring of public directories, assignment of accounts and passwords, backup schedules, and after-hours access; licensing restrictions on proprietary software; means to launch particular application programs, compile user-written FORTRAN and/or C programs, and access printers; and numerous other aspects are subject to considerable local variation. This book does not constrain local options in these matters. Instead, its users must draft a site-specific supplement, which we will refer to as the \textit{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 \texttt{\$HEAD}, \texttt{\$IDLHEAD}, and \texttt{\$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 \textit{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. Gehlke, 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. Stemmer, 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,

---

1In July 2001: Steve Adams (Widener University), Russel Kauffman (Muhlenberg College), Marylin Bell (Lake Forest College), Lawrence A. Mohar (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-Chiu Luk (University of California, Berkeley), Daryl Macon (Boise State University), Viktor Martisovits (Central College), Donald Miller (Central Missouri State University), Dorn Peterson (James Madison University), Robert Ragan (University of Wisconsin – Lacrosse), Shafiq Rahman (Allegeny College), Rahmatullah Syed (Norwich University), Jorge Talamantes (California State University – Bakersfield), Mark Timko (Elmhurst College), Paul Tjossem
provided a critical examination of the evolving manuscript and offered numerous suggestions for improvement—in addition to finding many previously undetected typographical glitches.

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

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

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

• Donald Knuth, originator of \TeX;
• Leslie Lamport, originator of \LaTeX;
• Han The Than, originator of pdf\TeX, which provides the engine underlying pdflatex for creating PDF files directly from \LaTeX\ source code.
• Pehong Chen and Nelson Beebe, originator and current maintainer of makeindex for preparing indices to \LaTeX\ documents;
• Enrico Gregorio, originator and current maintainer of the \LaTeX\ package 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;
ACKNOWLEDGEMENTS

- 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 lqusimp and lutrapzd and to distribute the source code for lqusimp 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 lqusimp.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), #891248 (1993), #011795 (2001)], the National Science Foundation for three grants [USE-8851685 (ILI–1988), DUE-9350667 (ILI–1993), and DUE-9952855 (CCLI–EMD–2000)], and Lawrence University for matching and other support.3 A small conference held at Lawrence in 1990 and supported by the Alfred P. Sloan Foundation stimulated thinking among the participants about

2Specifically flmoon, xflmoon, caldat, julday, xjulday, asevar, xavevar, rk4, xrk4, rksa, rkck, mmid, bsstep, rkdumb, odeint, trapzd, xtrapzd, qtrap, xqtrap, qsimp, qromb, point, rtbis, xrtbis, rtsnewt, xrtsnewt, xtsafe, xrtsafe, zbrak, gausoj, ludcmp, lubksb, tridag, svdcmp, svbkab, mnewt, newt, and brcydn (both in FORTRAN and in C).
3Any opinions, findings, and conclusions or recommendations expressed in this book are those of the author and do not necessarily reflect the views of any of these granting foundations or agencies.
uses of computers in upper-division undergraduate physics. All of these grants have contributed in many ways to the developments at Lawrence that have culminated in the writing of this book. In particular, the NSF CCLI-EMD grant made in February, 2000, supported my sabbatical while I finalized the text of (the first edition of) this book. That grant also supported four week-long summer faculty workshops that have, on the one hand, provided constructive feedback on a succession of drafts and, on the other hand, enhanced awareness nationally of this book and of the developments at Lawrence.

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

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

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

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

  In addition, these codings have been spot-checked with
  
  - MAXIMA Version 5.38.0 and wxMAXIMA 16.04.1,
  - MAXIMA 5.39.0 and wxMAXIMA 16.12.0, and
- MAXIMA 5.42.0 and wxMAXIMA 18.10.1

on a Hewlett-Packard platform running Windows 10.

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

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

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

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

- The OCTAVE codings have been fully verified with
  - 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{} 2e with the MiKTeX implementation on a Hewlett-Packard platform running Windows 7 and on a Hewlett-Packard platform running Windows 10. \LaTeX{} normally responds to the same source code on all platforms.

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

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

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

1 Preliminaries 1
   1.1 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

2 Introduction to IDL 15
   2.1 Beginning an IDL Session .......................................... 15
   2.2 Basic Entities in IDL ............................................... 16
      2.2.1 Data Types .................................................. 16
      2.2.2 Variable Names .............................................. 18
      2.2.3 Assignment of Values to Variables ......................... 18
      2.2.4 Commands .................................................... 19
   2.3 A Sampling of IDL Capabilities ................................. 19
      2.3.1 Creating and Examining Arrays .............................. 19
      2.3.2 Simple Mathematical Manipulations ....................... 23
      2.3.3 A Further Note on Array Indices; Matrix Multiplication .. 24
      2.3.4 Solving Linear Equations ................................ 25
      2.3.5 A First Graph ............................................... 26
   2.4 Saving and Retrieving the IDL Session .......................... 28
   2.5 Loops, Logical Expressions, and Conditionals ................. 31
   2.6 Reading Data from a File ......................................... 33
   2.7 On-line Help ..................................................... 34
   2.8 Pro-Files .......................................................... 35
   2.9 Eigenvalues and Eigenvectors ................................... 39
   2.10 Graphing Scalar Functions of One Variable .................... 40
      2.10.1 The Basic Strategy ......................................... 41
      2.10.2 Plotting Several Graphs on One Set of Axes ............. 41
      2.10.3 Polar Plots and Custom Axes ............................... 42
      2.10.4 Multiple Plots On a Page .................................. 44
      2.10.5 Plotting Experimental Data ................................ 44

xv
TABLE OF CONTENTS

2.11 Making Hard Copy ........................................... 47
  2.11.1 of Text ............................................... 47
  2.11.2 of Graphs .............................................. 48
2.12 Graphing Scalar Functions of Two Variables ..................... 48
  2.12.1 A Preliminary: The Procedure `lugen_grid' ................ 49
  2.12.2 Surface Plots: The Procedure `surface' .................. 50
  2.12.3 Contour Plots: The Procedure `contour' .................. 52
  2.12.4 Shaded Surfaces: The Procedure `shade_surf' ............. 53
  2.12.5 The Procedure `xsurface' ................................ 54
  2.12.6 The Procedure `show3' .................................. 55
  2.12.7 Functions of Two Variables in Polar Coordinates .......... 55
2.13 Graphing Scalar Functions of Three Variables ................... 56
  2.13.1 Reduction to Two-Dimensional Displays .................... 57
  2.13.2 Isosurfaces ............................................. 60
  2.13.3 Other IDL Routines ...................................... 61
2.14 Graphing Vector Fields ....................................... 62
  2.14.1 The Procedure `velovect' ................................ 63
  2.14.2 The Procedure `vel' ..................................... 63
  2.14.3 The Procedure `plot_field' .............................. 64
  2.14.4 More Elaborate Displays .................................. 65
  2.14.5 Three-Dimensional Vector Fields ......................... 65
2.15 Animation .................................................. 66
2.16 Advanced Graphing Features .................................... 68
  2.16.1 Fonts .................................................. 68
  2.16.2 Space Curves ............................................ 69
  2.16.3 System Variables ........................................ 71
  2.16.4 Using Multiple Windows .................................. 72
  2.16.5 Scaling so Circles are Circles ............................ 73
  2.16.6 Working with Color ...................................... 74
2.17 Miscellaneous Occasionally Useful Tidbits ....................... 76
  2.17.1 Specifying Directories .................................... 76
  2.17.2 IDL's Default Directory .................................. 76
  2.17.3 Customizing IDL .......................................... 77
  2.17.4 IDL's Search Path ....................................... 77
  2.17.5 Restoring IDL's Initial State ............................. 78
  2.17.6 Common Storage in IDL ................................... 78
  2.17.7 Reading Data from the Keyboard ........................... 78
  2.17.8 NaN and Inf .............................................. 79
2.18 References .................................................. 79
2.19 Exercises .................................................. 80
  2.19.1 Writing IDL Statements ................................... 80
  2.19.2 Finding Eigenvalues and Eigenvectors ..................... 81
  2.19.3 Graphing Scalar Functions of a Single Variable ........... 82
  2.19.4 Graphing Scalar Functions of Two Variables ............... 84
  2.19.5 Graphing Scalar Functions of Three Variables .............. 86
  2.19.6 Graphing Vector Fields ................................... 86
  2.19.7 Animation ............................................... 88
  2.19.8 Supplementary Exercises .................................. 89
2.A Listing of `lugen_grid' ........................................ 90

8 Introduction to *MATHEMATICA* .................................. 91
  8.1 Beginning a *Mathematica* Session ............................. 92
  8.2 On-Line Help ................................................ 93
  8.3 Basic Entities in *Mathematica* ............................... 94
  8.4 Variable Names in *Mathematica* ................................ 94
  8.5 Expressions in *Mathematica* ................................ 95
  8.6 Assigning Values to Variables; Defining Functions ............. 96
  8.7 Creating and Examining Lists .................................. 99
  8.8 Using *Mathematica* ......................................... 100
    8.8.1 ... for Arithmetic ..................................... 101
TABLE OF CONTENTS

8.8.2 for Algebra .......................................................... 102
8.8.3 for Complex Variables ............................................. 105
8.8.4 for Trigonometry ..................................................... 107
8.8.5 for Algebraic Equations ........................................... 108
8.8.6 for Manipulating Vectors and Matrices ........................... 110
8.8.7 for Basic Calculus ................................................... 111
8.8.8 for Laplace Transforms ............................................ 114
8.8.9 for Ordinary Differential Equations .............................. 115
8.8.10 for Vector Calculus ............................................... 119
8.9 Packages ................................................................. 121
8.10 Loops, Logical Expressions, and Conditionals ...................... 122
8.11 Command Files .......................................................... 124
8.12 High-Resolution Graphing ............................................ 126
8.13 Making Hard Copy ...................................................... 129
  8.13.1 from the NBI ....................................................... 130
  8.13.2 from the TBI ....................................................... 132
8.14 Output in \LaTeX Format .............................................. 132
8.15 Animation ............................................................... 133
8.16 Using the Notebook .................................................... 134
8.17 Customizations .......................................................... 135
8.18 Miscellaneous Occasionally Useful Tidbits ......................... 136
  8.18.1 Turning off Messages ........................................... 136
  8.18.2 Specifying Directories ......................................... 136
  8.18.3 Default Directory ............................................... 136
  8.18.4 Search Path ....................................................... 137
  8.18.5 Displaying and Changing Default Options ..................... 137
  8.18.6 Placing Text in Graphs ......................................... 138
  8.18.7 Space Curves ..................................................... 138
  8.18.8 Directing Plot to Separate Window ............................. 139
8.19 References .............................................................. 140
8.20 Exercises .............................................................. 140

9 Introduction to Programming ........................................ 151
  9.1 Components of a Programming Language ............................ 152
    9.1.1 Variables, Variable Names, and Use of Memory .............. 152
    9.1.2 Essential Action Statements ................................ 154
    9.1.3 Logical Expressions and Conditions .......................... 155
    9.1.4 Syntactic Wrinkles ............................................ 156
    9.1.5 Essential Control Structures ................................ 156
  9.2 Sample Short Algorithms ........................................... 160
  9.3 Two Larger Algorithms ............................................. 162
    9.3.1 Solving Laplace’s Equation ................................ 162
    9.3.2 File Output/Input .............................................. 164
  9.8 Solving Laplace’s Equation with IDL ................................ 167
  9.12 Creating and Storing Two-Dimensional Scalar Arrays ........... 168
    9.12.3 with IDL ....................................................... 168
  9.13 Creating and Storing Three-Dimensional Scalar Arrays .......... 169
    9.13.3 with IDL ....................................................... 170
  9.14 Creating and Storing Two-Dimensional Vector Arrays ............ 171
    9.14.3 with IDL ....................................................... 172
  9.15 Creating and Storing Three-Dimensional Vector Arrays .......... 173
    9.15.3 with IDL ....................................................... 174
  9.16 Reading Files ....................................................... 174
    9.16.3 with IDL ....................................................... 175
  9.17 References .......................................................... 175
  9.18 Exercises ........................................................... 175
  9.D Listing of lugen_list.pro (IDL) .................................. 182
  9.E Listing of lugen_list_event.pro (IDL) ............................ 184
  9.Q Listing of luread_list.pro (IDL) .................................. 186
# TABLE OF CONTENTS

## 11 Solving Ordinary Differential Equations

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.1 Sample Problems</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>11.2 Laplace Transforms</td>
<td>199</td>
</tr>
<tr>
<td>11.5 Solving ODEs Symbolically with Mathematica</td>
<td>200</td>
</tr>
<tr>
<td>11.5.1 Projectile in a Viscous Medium</td>
<td>201</td>
</tr>
<tr>
<td>11.5.2 Logistic Growth</td>
<td>203</td>
</tr>
<tr>
<td>11.5.3 Damped Harmonic Oscillator</td>
<td>204</td>
</tr>
<tr>
<td>11.5.4 Chain Radioactive Decay</td>
<td>204</td>
</tr>
<tr>
<td>11.5.5 Coupled Oscillators</td>
<td>207</td>
</tr>
<tr>
<td>11.5.6 Standing Waves in a String</td>
<td>208</td>
</tr>
<tr>
<td>11.5.7 Infinite Depth Quantum Well</td>
<td>211</td>
</tr>
<tr>
<td>11.5.8 Finding Solutions by Series Methods</td>
<td>211</td>
</tr>
<tr>
<td>11.6 Algorithms for Solving ODEs Numerically</td>
<td>214</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>218</td>
</tr>
<tr>
<td>11.6.4 Assessing Accuracy</td>
<td>220</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>11.7 Solving ODEs Numerically with IDL</td>
<td>223</td>
</tr>
<tr>
<td>11.7.1 Using Elementary Commands</td>
<td>223</td>
</tr>
<tr>
<td>11.7.2 Defining ODEs for IDL</td>
<td>226</td>
</tr>
<tr>
<td>11.7.3 The Procedures <strong>ri4</strong>, <strong>lsode</strong>, <strong>ludiffeq_23</strong>, and <strong>ludiffeq_45</strong></td>
<td>227</td>
</tr>
<tr>
<td>11.7.4 Chain Radioactive Decay</td>
<td>229</td>
</tr>
<tr>
<td>11.7.5 Damped Harmonic Oscillator</td>
<td>232</td>
</tr>
<tr>
<td>11.7.6 Planetary Orbits</td>
<td>234</td>
</tr>
<tr>
<td>11.7.7 Standing Waves in a String</td>
<td>237</td>
</tr>
<tr>
<td>11.7.8 The Quantum Harmonic Oscillator</td>
<td>240</td>
</tr>
<tr>
<td>11.7.9 Pro-files in the Public Library</td>
<td>244</td>
</tr>
<tr>
<td>11.13 Solving ODEs Numerically with Mathematica</td>
<td>246</td>
</tr>
<tr>
<td>11.16 Exercises</td>
<td>247</td>
</tr>
<tr>
<td>11.16.1 … using Symbolic Methods</td>
<td>247</td>
</tr>
<tr>
<td>11.16.2 … using Numerical Methods</td>
<td>251</td>
</tr>
</tbody>
</table>

## 13 Evaluating Integrals

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.1 Sample Problems</td>
<td>257</td>
</tr>
<tr>
<td>13.1.1 One-Dimensional Trajectories</td>
<td>257</td>
</tr>
<tr>
<td>13.1.2 Center of Mass</td>
<td>258</td>
</tr>
<tr>
<td>13.1.3 Moment of Inertia; Radius of Gyration</td>
<td>259</td>
</tr>
<tr>
<td>13.1.4 The Large Amplitude Simple Pendulum</td>
<td>260</td>
</tr>
<tr>
<td>13.1.5 Statistical Data Analysis</td>
<td>261</td>
</tr>
<tr>
<td>13.1.6 The Cornu Spiral</td>
<td>262</td>
</tr>
<tr>
<td>13.1.7 Electric and Magnetic Fields and Potentials</td>
<td>262</td>
</tr>
<tr>
<td>13.1.8 Quantum Probabilities</td>
<td>264</td>
</tr>
<tr>
<td>13.1.9 Expansion in Orthogonal Functions</td>
<td>264</td>
</tr>
<tr>
<td>13.4 Evaluating Integrals Symbolically with Mathematica</td>
<td>265</td>
</tr>
<tr>
<td>13.4.1 Relativistic Motion Under a Constant Force</td>
<td>266</td>
</tr>
<tr>
<td>13.4.2 Center of Mass</td>
<td>267</td>
</tr>
<tr>
<td>13.4.3 Moment of Inertia; Radius of Gyration</td>
<td>268</td>
</tr>
<tr>
<td>13.4.4 Electrostatic Potential of a Finite Line Charge</td>
<td>268</td>
</tr>
<tr>
<td>13.4.5 The Helmholtz Coil</td>
<td>270</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>13.4.6</td>
<td>Period of a Pendulum: Correction as Amplitude Grows</td>
</tr>
<tr>
<td>13.4.7</td>
<td>Fourier Coefficients for Half-Rectified Signal</td>
</tr>
<tr>
<td>13.5</td>
<td>Algorithms for Numerical Integration</td>
</tr>
<tr>
<td>13.5.1</td>
<td>Newton-Cotes Quadrature</td>
</tr>
<tr>
<td>13.5.2</td>
<td>Rearrangements for Computational Efficiency</td>
</tr>
<tr>
<td>13.5.3</td>
<td>Assessing Error</td>
</tr>
<tr>
<td>13.5.4</td>
<td>Iterative and Adaptive Algorithms</td>
</tr>
<tr>
<td>13.5.5</td>
<td>Gaussian Quadrature</td>
</tr>
<tr>
<td>13.6</td>
<td>Evaluating Integrals Numerically with IDL</td>
</tr>
<tr>
<td>13.6.1</td>
<td>Using Elementary Commands</td>
</tr>
<tr>
<td>13.6.2</td>
<td>The Function luqsimp</td>
</tr>
<tr>
<td>13.6.3</td>
<td>Moment of Inertia</td>
</tr>
<tr>
<td>13.6.4</td>
<td>Quantum Probabilities</td>
</tr>
<tr>
<td>13.6.5</td>
<td>Integrals as Functions of the Upper Limit</td>
</tr>
<tr>
<td>13.6.6</td>
<td>The Error Function</td>
</tr>
<tr>
<td>13.6.7</td>
<td>The Cornu Spiral</td>
</tr>
<tr>
<td>13.6.8</td>
<td>Integrals as Functions of an Internal Parameter</td>
</tr>
<tr>
<td>13.6.9</td>
<td>The Off-Axis Electrostatic Potential of Two Rings</td>
</tr>
<tr>
<td>13.12</td>
<td>Evaluating Integrals Numerically with Mathematica</td>
</tr>
<tr>
<td>13.12.1</td>
<td>Quantum Probability</td>
</tr>
<tr>
<td>13.12.2</td>
<td>The Error Function</td>
</tr>
<tr>
<td>13.12.3</td>
<td>The Off-Axis Electrostatic Potential of Two Rings</td>
</tr>
<tr>
<td>13.15</td>
<td>Exercises</td>
</tr>
<tr>
<td>13.15.1</td>
<td>... using Symbolic Methods</td>
</tr>
<tr>
<td>13.15.2</td>
<td>... using Numerical Methods</td>
</tr>
<tr>
<td>14</td>
<td>Finding Roots</td>
</tr>
<tr>
<td>14.1</td>
<td>Sample Problems</td>
</tr>
<tr>
<td>14.1.1</td>
<td>Classical Turning Points</td>
</tr>
<tr>
<td>14.1.2</td>
<td>A Max-Min Problem: Equilibrium Points</td>
</tr>
<tr>
<td>14.1.3</td>
<td>Natural Frequencies of Oscillating Systems</td>
</tr>
<tr>
<td>14.1.4</td>
<td>Range of Projectile in Viscous Medium</td>
</tr>
<tr>
<td>14.1.5</td>
<td>Energy Levels in a Quantum Well</td>
</tr>
<tr>
<td>14.2</td>
<td>Symbolic Approaches</td>
</tr>
<tr>
<td>14.5</td>
<td>Finding Roots Symbolically with Mathematica</td>
</tr>
<tr>
<td>14.5.1</td>
<td>Classical Turning Points</td>
</tr>
<tr>
<td>14.5.2</td>
<td>Equilibrium Points</td>
</tr>
<tr>
<td>14.5.3</td>
<td>Natural Frequencies</td>
</tr>
<tr>
<td>14.5.4</td>
<td>Range of Projectile in Viscous Medium</td>
</tr>
<tr>
<td>14.6</td>
<td>Algorithms for Finding Roots Numerically</td>
</tr>
<tr>
<td>14.6.1</td>
<td>The Method of Bisection</td>
</tr>
<tr>
<td>14.6.2</td>
<td>Newton’s Method</td>
</tr>
<tr>
<td>14.6.3</td>
<td>Other Methods</td>
</tr>
<tr>
<td>14.6.4</td>
<td>Assessing Accuracy</td>
</tr>
<tr>
<td>14.7</td>
<td>Finding Roots Numerically with IDL</td>
</tr>
<tr>
<td>14.7.1</td>
<td>Classical Turning Points</td>
</tr>
<tr>
<td>14.7.2</td>
<td>Natural Frequencies</td>
</tr>
<tr>
<td>14.7.3</td>
<td>Range of Projectile</td>
</tr>
<tr>
<td>14.7.4</td>
<td>Energy Levels in Quantum Well</td>
</tr>
<tr>
<td>14.13</td>
<td>Finding Roots Numerically with Mathematica</td>
</tr>
<tr>
<td>14.13.1</td>
<td>Classical Turning Points</td>
</tr>
<tr>
<td>14.13.2</td>
<td>Range of Projectile</td>
</tr>
<tr>
<td>14.13.3</td>
<td>Energy Levels in Quantum Well</td>
</tr>
<tr>
<td>14.16</td>
<td>Solving Simultaneous Equations</td>
</tr>
<tr>
<td>14.16.1</td>
<td>Systems of Linear Equations</td>
</tr>
<tr>
<td>14.16.2</td>
<td>Systems of Nonlinear Equations</td>
</tr>
<tr>
<td>14.17</td>
<td>Exercises</td>
</tr>
<tr>
<td>14.17.1</td>
<td>... using Symbolic Methods</td>
</tr>
<tr>
<td>14.17.2</td>
<td>... using Numerical Methods</td>
</tr>
<tr>
<td>14.17.4</td>
<td>Finding More than One Unknown</td>
</tr>
</tbody>
</table>
14.A Listing of IDL Procedure \texttt{lubisect.pro} ........................................ 352

15 Solving Partial Differential Equations ........................................ 353

15.1 Sample Problems ......................................................................... 355
  15.1.1 Motion of a String: The Wave Equation ................................ 355
  15.1.2 Heat Flow: The Diffusion Equation .................................... 358
  15.1.3 Steady State Heat Flow: The Laplace Equation .................... 359
  15.1.4 Other Situations ................................................................ 359
  15.1.5 Classification of Second-Order PDEs ................................. 360
  15.1.6 Equations of Fluid Dynamics: Navier-Stokes’ Equations .... 361
    15.1.6.1 Conservation of Mass: The Equation of Continuity .... 362
    15.1.6.2 Newton’s Second Law ........................................... 363
    15.1.6.3 A Special Case: Non-viscous Flow ......................... 365
    15.1.6.4 A Second Special Case: Sound Waves .................... 366
  15.1.7 A General 1D Equation ..................................................... 367
  15.1.8 A General 2D Equation ..................................................... 368

15.2 Finite Difference Methods (FDMs) in One Dimension ................. 368
  15.2.1 A General Coding ......................................................... 371
  15.2.2 An Example: Simple Harmonic Motion ............................ 372

15.3 Using IDL to Solve 1D PDEs via an FDM .................................... 371
  15.3.1 A General Coding ......................................................... 371

15.4 Finite Element Methods (FEMs) in One Dimension ....................... 375
  15.4.1 Discretizing the Domain: Preprocessing ........................... 376
  15.4.2 Selecting Interpolation or Shape Functions ...................... 377
  15.4.3 Formulating the Equations for a Single Element ............... 377
  15.4.4 Assembling the System of Equations ................................ 380
  15.4.5 Incorporating the Boundary Conditions ............................ 382

15.5 Using IDL to Solve 1D PDEs via FEMs ....................................... 384
  15.5.1 A General Coding ......................................................... 384
  15.5.2 An Example: Simple Harmonic Motion ............................ 386

15.6 Finite Difference Methods (FDMs) in Two Dimensions ............... 388
  15.6.1 The Wave Equation ....................................................... 389
  15.6.2 The Diffusion Equation ................................................ 391
  15.6.3 The Laplace Equation .................................................. 394

15.7 Using IDL to Solve 2D PDEs via an FDM .................................... 396
  15.7.1 The Wave Equation ....................................................... 397
  15.7.2 The Diffusion Equation ................................................ 402
  15.7.3 The Laplace Equation .................................................. 408

15.8 Finite Element Methods (FEMs) in Two Dimensions ..................... 414
  15.8.1 Discretizing the Domain: Pre-processing ........................ 414
  15.8.2 Selecting the Interpolation or Shape Functions ............... 416
  15.8.3 Formulating the Equations for a Single Element ............... 417
  15.8.4 Assembling the System of Equations ............................... 419
  15.8.5 Incorporating the Boundary Conditions ............................ 421

15.9 Using IDL to Solve 2D PDEs via FEM ....................................... 423
  15.9.1 A General Coding ......................................................... 423
  15.9.2 An Example: Isotropic Heat Flow .................................. 430

15.10 Using IDL to Solve 1D PDEs via FEMs ...................................... 433
  15.10.1 A General Coding ......................................................... 433
  15.10.2 An Example: Simple Harmonic Motion ............................ 434

15.11 Finite Difference Methods (FDMs) in Two Dimensions ............... 441
  15.11.1 The Wave Equation ....................................................... 441
  15.11.2 The Diffusion Equation ................................................ 444
  15.11.3 The Laplace Equation .................................................. 445

15.12 Using IDL to Solve 2D PDEs via an FDM .................................... 454
  15.12.1 A General Coding ......................................................... 444
  15.12.2 An Example: Isotropic Heat Flow .................................. 445

15.13 Using IDL to Solve 2D PDEs via FEM ....................................... 446
  15.13.1 A General Coding ......................................................... 446
  15.13.2 An Example: Isotropic Heat Flow .................................. 447

15.14 Exercises ................................................................................. 448
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLE OF CONTENTS</td>
<td>xxii</td>
</tr>
<tr>
<td>15.F.3  ... using Mathematica</td>
<td>448</td>
</tr>
<tr>
<td>15.G  Program(s) for FEM Approach to 2D Problems</td>
<td>451</td>
</tr>
<tr>
<td>15.G.1  Listing of fem2d.pro (IDL)</td>
<td>451</td>
</tr>
<tr>
<td>A  Introduction to \LaTeX\xspace</td>
<td>455</td>
</tr>
<tr>
<td>A.1  Creating a Simple Document</td>
<td>456</td>
</tr>
<tr>
<td>A.1.1  Structuring a \LaTeX\xspace Source File</td>
<td>456</td>
</tr>
<tr>
<td>A.1.2  Creating a PostScript File</td>
<td>458</td>
</tr>
<tr>
<td>A.1.3  Creating a PDF File</td>
<td>459</td>
</tr>
<tr>
<td>A.1.4  Displaying the Document on the Screen</td>
<td>460</td>
</tr>
<tr>
<td>A.1.5  Printing the Document</td>
<td>460</td>
</tr>
<tr>
<td>A.2  Specification of Global Style: The Preamble</td>
<td>461</td>
</tr>
<tr>
<td>A.3  In-Text Specification of Local Style</td>
<td>463</td>
</tr>
<tr>
<td>A.3.1  Type Style</td>
<td>463</td>
</tr>
<tr>
<td>A.3.2  Type Size</td>
<td>464</td>
</tr>
<tr>
<td>A.3.3  White Space</td>
<td>465</td>
</tr>
<tr>
<td>A.3.4  Environments</td>
<td>466</td>
</tr>
<tr>
<td>A.3.5  \LaTeX\xspace Packages</td>
<td>467</td>
</tr>
<tr>
<td>A.3.6  Miscellaneous Other Capabilities</td>
<td>467</td>
</tr>
<tr>
<td>A.4  Including Equations</td>
<td>468</td>
</tr>
<tr>
<td>A.5  Including Lists</td>
<td>469</td>
</tr>
<tr>
<td>A.6  Including Tables</td>
<td>470</td>
</tr>
<tr>
<td>A.7  Including Illustrations</td>
<td>472</td>
</tr>
<tr>
<td>A.7.1  Using Cut and Paste</td>
<td>472</td>
</tr>
<tr>
<td>A.7.2  Using the \texttt{graphicx} Package</td>
<td>473</td>
</tr>
<tr>
<td>A.7.3  Using the \texttt{tikz} Package</td>
<td>475</td>
</tr>
<tr>
<td>A.7.4  Using the \texttt{picture} Environment and \texttt{pict2e} Package</td>
<td>478</td>
</tr>
<tr>
<td>A.8  Including TOC, LOF, and LOT</td>
<td>478</td>
</tr>
<tr>
<td>A.9  Including an Index</td>
<td>479</td>
</tr>
<tr>
<td>A.10  Including Hyperlinks</td>
<td>481</td>
</tr>
<tr>
<td>A.11  Converting .\texttt{eps} and .\texttt{ps} Files to .\texttt{pdf}</td>
<td>482</td>
</tr>
<tr>
<td>A.12  Using Conditional Expressions in \LaTeX\xspace</td>
<td>484</td>
</tr>
<tr>
<td>A.13  Error Messages Generated by \LaTeX\xspace</td>
<td>485</td>
</tr>
<tr>
<td>A.14  The Page Previewer for .\texttt{dvi} Files</td>
<td>486</td>
</tr>
<tr>
<td>A.15  The Spell Checker in UNIX</td>
<td>487</td>
</tr>
<tr>
<td>A.16  A Sample Document</td>
<td>488</td>
</tr>
<tr>
<td>A.17  Miscellaneous Other Features</td>
<td>489</td>
</tr>
<tr>
<td>A.18  References</td>
<td>496</td>
</tr>
<tr>
<td>A.19  Exercises</td>
<td>497</td>
</tr>
<tr>
<td>A.A  Listings</td>
<td>500</td>
</tr>
<tr>
<td>A.A.1  ... for Windows</td>
<td>500</td>
</tr>
<tr>
<td>A.A.1.1  Batch File \texttt{ceps2pdf}\texttt{.bat}</td>
<td>500</td>
</tr>
<tr>
<td>A.A.1.2  Batch File \texttt{cps2pdf}\texttt{.bat}</td>
<td>500</td>
</tr>
<tr>
<td>A.A.1.3  Batch File \texttt{rdfileeps}\texttt{.bat}</td>
<td>500</td>
</tr>
<tr>
<td>A.A.1.4  Batch File \texttt{rdfileeps}\texttt{.bat}</td>
<td>500</td>
</tr>
<tr>
<td>A.A.2  ... for UNIX</td>
<td>501</td>
</tr>
<tr>
<td>A.A.2.1  Shell Script \texttt{ceps2pdf}</td>
<td>501</td>
</tr>
<tr>
<td>A.A.2.2  Shell Script \texttt{cps2pdf}</td>
<td>501</td>
</tr>
<tr>
<td>A.A.2.3  Shell Script \texttt{rdfileeps}</td>
<td>501</td>
</tr>
<tr>
<td>A.A.2.4  Shell Script \texttt{rdfileps}</td>
<td>501</td>
</tr>
<tr>
<td>A.A.3  ... for Windows and UNIX</td>
<td>502</td>
</tr>
<tr>
<td>A.A.3.1  Python Script \texttt{ExtractFileName.py}</td>
<td>502</td>
</tr>
</tbody>
</table>

Z  Contacting Software Vendors  

Index  

507
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

1. an array processing program (e.g., IDL®, MATLAB®, OCTAVE, PYTHON, ...);
2. a computer algebra system (e.g., MAPLE®, Mathematica®, MAXIMA, ...);
3. 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
4. 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

1. be acquainted with the main capabilities of at least one operating system (e.g., UNIX, Windows, Macintosh OS, ...),
2. 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
3. of a publishing package (e.g., TpXlive, LpXeX, MiKpXeX, OzpX, ...) 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

1Many 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 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 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 Local Guide. This paragraph, however, is the only point in the book at which the wisdom of being logged in is explicitly mentioned.

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.
1.1.1 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 \(\text{SHIFT}/5\)), 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 \(\langle SP\rangle\), has been “printed”.) Pressing the key labeled RETURN “prints” the carriage return character (denoted \(\langle\text{CR}\rangle\)), 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

\[
\text{LOOP} \\
\quad \text{Read code from keyboard} \\
\quad \text{Send code to printer} \\
\text{END_LOOP}
\]

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 \(\langle SP\rangle\) 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 \(\langle\text{CR}\rangle\) 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 \(\langle\text{CR}\rangle\) 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 \(\langle\text{CR}\rangle\) from the keyboard must trigger the sending of the codes for the pair of characters \(\langle\text{CR}\rangle/\langle\text{LF}\rangle\) to the printer. To simulate a typewriter, we must embellish the above statements to

\[\text{3}\]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.

\[\text{4}\]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 \langle CR\rangle
    THEN Send code for \langle LF\rangle 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 \langle CR\rangle and \langle LF\rangle. Other control characters familiar to the user of an ordinary typewriter are the backspace \langle BS\rangle, which causes the printhead to back up one space; the horizontal tab \langle HT\rangle, which causes the printhead to advance to the next pre-set (horizontal) tab position; and the vertical tab \langle VT\rangle, 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.\(^5\)\(^6\) 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,

\[
1 \times 128 + 0 \times 64 + 0 \times 32 + 1 \times 16 + 1 \times 8 + 1 \times 4 + 0 \times 2 + 1 \times 1 = 157
\]

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

### 1.1.3 The ASCII Character Set

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

There are, however, a number of conventional associations of codes with characters. The most commonly used scheme is the American Standard Code for Information Interchange (ASCII, pronounced ass’-key). In this code, characters are associated with eight-bit binary patterns. While the second 128 of the 256 distinguishable patterns [i.e., characters 128–255 (decimal)] have a variety of assignments to characters, the first 128 patterns [i.e., characters 0–127 (decimal)] have the standard assignments enumerated in Table 1.1. The control characters (non-printing characters) all have (decimal) ASCII codes in the range 0–31. Further, the ASCII code for each uppercase letter is 32 less than the code for the corresponding lowercase letter; i.e., turning off the 32-bit in the code for a lowercase letter generates the code for the corresponding uppercase letter. Finally, the ASCII code for a control character is 64 less than the code for the associated uppercase letter; i.e., turning off the 64-bit in the code for an uppercase letter (say C) generates the code for the corresponding control character (\(\langle\text{CTRL/C}\rangle\)). 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>Code</th>
<th>Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>NUL</td>
</tr>
<tr>
<td>1</td>
<td>SOH</td>
</tr>
<tr>
<td>2</td>
<td>STX</td>
</tr>
<tr>
<td>3</td>
<td>EOX</td>
</tr>
<tr>
<td>4</td>
<td>EOT</td>
</tr>
<tr>
<td>5</td>
<td>ENQ</td>
</tr>
<tr>
<td>6</td>
<td>ACK</td>
</tr>
<tr>
<td>7</td>
<td>BEL</td>
</tr>
<tr>
<td>8</td>
<td>BS</td>
</tr>
<tr>
<td>9</td>
<td>HT</td>
</tr>
<tr>
<td>10</td>
<td>LF</td>
</tr>
<tr>
<td>11</td>
<td>VT</td>
</tr>
<tr>
<td>12</td>
<td>FF</td>
</tr>
<tr>
<td>13</td>
<td>CR</td>
</tr>
<tr>
<td>14</td>
<td>SO</td>
</tr>
<tr>
<td>15</td>
<td>SI</td>
</tr>
<tr>
<td>16</td>
<td>DLE</td>
</tr>
<tr>
<td>17</td>
<td>DC1</td>
</tr>
<tr>
<td>18</td>
<td>DC2</td>
</tr>
<tr>
<td>19</td>
<td>DC3</td>
</tr>
<tr>
<td>20</td>
<td>DC4</td>
</tr>
<tr>
<td>21</td>
<td>NAK</td>
</tr>
<tr>
<td>22</td>
<td>SYN</td>
</tr>
<tr>
<td>23</td>
<td>ETB</td>
</tr>
<tr>
<td>24</td>
<td>CAN</td>
</tr>
<tr>
<td>25</td>
<td>EM</td>
</tr>
<tr>
<td>26</td>
<td>SUB</td>
</tr>
<tr>
<td>27</td>
<td>ESC</td>
</tr>
<tr>
<td>28</td>
<td>FS</td>
</tr>
<tr>
<td>29</td>
<td>GS</td>
</tr>
<tr>
<td>30</td>
<td>RS</td>
</tr>
<tr>
<td>31</td>
<td>US</td>
</tr>
</tbody>
</table>

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.\(^9\)

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

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!
CHAPTER 1. PRELIMINARIES

1.2 Files and Directories

At (nearly) the most microscopic level, information in a computer is recorded in bytes stored in memory or, more permanently, on a hard drive. At the next level up, aggregations of these bytes into larger units that must be kept together are called files. Each file will have a name. Some of the files containing portions of or referenced in this book, for example, are named \texttt{assemble.tex}, \texttt{laplace.f}, \texttt{laplace.c}, \texttt{trapezoidal.xc}, and \texttt{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.\footnote{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/eps} and \texttt{.pdf} identify PostScript and PDF files.} Some files—called ASCII text files—contain nothing but printing ASCII characters (and perhaps such simple control characters as \texttt{\langle CR\rangle} and \texttt{\langle HT\rangle}) and can be displayed on the screen, printed on a printer, or examined and edited with a text editor. Though some of their bytes can be interpreted as printing characters, other files—called binary files—contain also (perhaps numerous) non-printing characters and cannot be displayed on the screen, examined in (ordinary) text editors, or printed on a printer. Files of this latter type may be special data files created by programs; more often, they are executable files which contain compiled \textit{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, \textit{...}. 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, \textit{...}. In the computer world, we need 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.\footnote{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.} 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 \texttt{/}; the forward slash is also used to separate directory paths in an extended path. Thus, to specify the location of a file buried several directories down from this universal starting point, we would have to supply an identifier like

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

which indicates that the file \texttt{intro.tex}—the \LaTeX{} source file for this chapter—will be found in the \texttt{intro} directory in the \texttt{CCLI} directory in the \texttt{cook} directory in the \texttt{people} directory in the \texttt{usr} directory in the \texttt{/} (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}.\footnote{In Windows, for example, the backslash character \texttt{\backslash} is used to separate directories in a path.} 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.

\texttt{\footnotesize{10}}\footnote{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/eps} and \texttt{.pdf} identify PostScript and PDF files.} 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/eps} and \texttt{.pdf} identify PostScript and PDF files.
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. The Local Guide explains how to translate these symbols if your mouse has fewer than three buttons.

---

13If 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

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

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 $\text{SQRT}$, for example, would be invoked with a statement like

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

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

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

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

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

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

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

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

Mathematically, we assume in this book that you understand the notions of derivatives, integrals, and ordinary differential equations and that you have some acquaintance with linear algebra (matrix operations, eigenvalues and eigenvectors, ...). Physically, we suppose that you have taken a couple of introductory, calculus-based courses in physics and are continuing with intermediate courses in physics. This book makes contact with many intermediate-level physics courses, but it is not focused on any particular one of those courses. It draws on topics covered in several such courses whenever appropriate.

Local Guide but is referred to throughout this book with the symbol $\$\text{HEAD}$. In any attempt to access these files, you must replace this symbol in the text with whatever is proper for your site. More specifically, files associated with the chapter on IDL will be found in the directory $\$\text{HEAD/IDL}$, those associated with the chapter on integration will be found in subdirectories of the directory $\$\text{HEAD/integ}$, etc.

- Less often, we will also refer to items in the directory tree associated with one or another application program. The heads of those directories are also site specific and are identified in the Local Guide. In this book, we refer to the head of the IDL directory tree with the symbol $\$\text{IDLHEAD}$, the head of the Numerical Recipes directory tree with the symbol $\$\text{NRHEAD}$, etc.

- Especially in the chapters on solving ordinary differential equations, evaluating integrals, and finding roots (but also occasionally in other chapters), the sequence of numbers assigned to the end of chapter exercises has gaps. Most of these gaps come about because those few exercises that are specific to a particular software package are omitted in versions that do not include that package. The gaps are placed so that the numbers assigned to identical exercises in various versions of this book will be the same, thereby facilitating the keying of solutions to representative exercises to the book itself.

- In a similar vein, to facilitate communication among users of different versions, chapters and sections focusing on a particular software package retain in all versions the same numbers. That decision, of course, means that the sequence of numbers assigned to chapters and sections in a particular version will have gaps corresponding to chapters and sections omitted. In contrast, equation numbers, figure numbers, and table numbers will run continuously within chapters and page numbers will run continuously from the beginning of the book to the end. Thus, references to chapter numbers and section numbers will be universal, the same in all versions, while references to equations, figures, tables, and pages will be version-specific.

- With regard to the index, please note the following:
  - Italic page numbers in the index convey that the page reference is to an end-of-chapter exercise, not to a textual discussion.
  - Because index commands in \LaTeX cannot be placed in \texttt{verbatim} environments and, for a few similar reasons, some items referred to in the index may actually occur towards the bottom of the previous page or near the top of the following page. Readers must therefore understand that an index entry pointing to page 215, for example, identifies an area that includes perhaps the bottom quarter of page 214 and the top quarter of page 216 as well. Straightening out this departure from the conventional understanding is difficult; it is much easier to adjust the understanding of the meaning of a page reference.
  - References to such symbols as .*, ::, ^=, */.*/, and ! are included in the index. Since many of these symbols are, however, extremely common, the index includes pointers to only some of their appearances within the text. Those included have been chosen to illustrate a representative spectrum of uses, so reviewing the few items identified in the index should convey the full range of uses of these symbols.
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
\[
n!! = n \times (n-2) \times (n-4) \times (n-6) \times \cdots \times \begin{cases} 4 \times 2 & \text{n even} \\ 3 \times 1 & \text{n odd} \end{cases} \tag{1.2}
\]
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
\[
10!! = 10 \times 8 \times 6 \times 4 \times 2 = 2^5 \times 5 \times 4 \times 3 \times 2 \times 1 = 2^5 \cdot 5! \Rightarrow (2n)!! = 2^n \cdot n! \tag{1.3}
\]
If \( n \) is odd, say 9, recasting its double factorial takes mildly more work, but is illustrated in the chain
\[
9!! = 9 \times 7 \times 5 \times 3 \times 1 = \frac{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{8 \times 6 \times 4 \times 2} = \frac{9!}{8!!} = \frac{9!}{2^4 \cdot 4!} \Rightarrow (2n+1)!! = \frac{(2n+1)!}{2^n \cdot n!} \tag{1.4}
\]

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
\[
\Gamma\left(\frac{1}{2}\right) = \int_0^\infty \frac{e^{-t}}{\sqrt{t}} \, dt = \int_0^\infty \frac{e^{-x^2}}{x} \, x \, dx = \int_0^\infty e^{-x^2} \, dx \tag{1.7}
\]
Then, we examine the square of the quantity of interest, change to polar coordinates, and find that
\[
\left[\Gamma\left(\frac{1}{2}\right)\right]^2 = \left(\int_0^\infty e^{-x^2} \, dx\right)^2 = \left(\int_0^\infty e^{-y^2} \, dy\right) \left(\int_0^\infty e^{-x^2} \, dx\right) \tag{1.8}
\]
\[
= \int_0^\infty \int_0^\infty e^{-(x^2+y^2)} \, dx \, dy = \int_0^\infty e^{-r^2} \, r \, dr \int_0^{2\pi} d\phi = \pi \tag{1.9}
\]
and we conclude that
\[ \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \] (1.10)

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

The Gamma function, however, has a particularly interesting property that we can deduce if we apply integration by parts to the definition. Provided \( \nu > 1 \), we find that
\[ \Gamma(\nu) = -\int_0^\infty t^{\nu-1} e^{-t} \, dt + (\nu - 1) \int_0^\infty t^{\nu-2} e^{-t} \, dt = (\nu - 1) \Gamma(\nu - 1) \] (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-integer arguments. Indeed, one often sees the notation \( \nu! \) as an alternative to the notation \( \Gamma(\nu + 1) \)—and the latter in fact provides a formal definition of the former.\(^{15}\) Note that, since we know quite explicitly that \( \Gamma(1) = 1 \), this connection between the Gamma and factorial functions supports what is sometimes an assertion of convenience, namely that \( 0! = 1 \).

1.5.2 The Laplace Transform

One tool used behind the scenes by symbolic solvers of ordinary differential equations is called the Laplace transform, which we describe here to avoid duplicating the discussion at several places in subsequent chapters. While we are not likely to make much use of the Laplace transform directly, knowing its properties may sometimes be valuable as we try to guide a symbolic manipulator that uses the technique. Defined for a function \( f(t) \) by the integral
\[ \mathcal{L}\left(f(t)\right) = \hat{f}(s) = \int_0^{\infty} e^{-st} f(t) \, dt \] (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,
\[ \mathcal{L}\left(a f(t) + b g(t)\right) = \int_0^{\infty} e^{-st} \left(af(t) + bg(t)\right) \, dt = a \int_0^{\infty} e^{-st} f(t) \, dt + b \int_0^{\infty} e^{-st} g(t) \, dt = a \mathcal{L}\left(f(t)\right) + b \mathcal{L}\left(g(t)\right) \] (1.15)

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

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

Table 1.2: A short table of Laplace transforms.

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

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

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

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

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

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 2

Introduction to IDL

Note: All program (*.pro) and data (*.dat) files referred to in this chapter are available in the directory \$HEAD/idl, 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 IDL’s default search path. (See Section 2.17.4.) If so, the files can be found by IDL 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.

IDL®—an acronym for Interactive Data Language—is an interactive software package whose primary—though certainly not only—capabilities are to manipulate numbers, which may be scalars (1 x 1 arrays), vectors (1 x n or n x 1 arrays), matrices (m x n arrays) or higher dimensional arrays, and to produce graphical visualizations. A typical interaction with the program will involve (1) creating one or more arrays in IDL’s workspace (either by computing them directly in IDL or by importing them into IDL 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 assorted IDL documentation (see Section 2.18) and the on-line help messages accessible from within IDL itself (see Section 2.7). We shall refer to this documentation collectively as the IDL manuals.

Two user interfaces to IDL are provided, a command-line interface (CLI) and a graphical user interface (GUI), called the IDL Development Environment, which facilitates working with files containing IDL programs. Both interfaces are available in all operating systems, though accessing the CLI in the Windows and Macintosh environments may require special contortions. With either interface, statements to IDL are structured and typed in the same way. Thus, in this chapter, we confine ourselves to describing the CLI.

2.1 Beginning an IDL Session

Detailed instructions for initiating a session with IDL will be found in the Local Guide. In general, IDL will be started either (1) by typing the command idl or the command idlde at the prompt from the operating system, or (2) by double-clicking the left mouse button on the IDL icon on the

1IDL Exelis Visual Information Solutions, formerly ITT Visual Information Solutions, and formerly Research Systems, Inc., all of Boulder, Colorado, and is registered in the United States Patent and Trademark Office. The acronym IDL also identifies a commercially available program produced and marketed by Exelis. (See Appendix Z for full contact information. More recently, Exelis has become a division of Harris Geospatial Solutions of Melbourne, Florida.) Use of IDL at any particular site is subject to the provisions of whatever license that site has negotiated with Harris or its predecessors. The terms of that license are explained in the Local Guide.

2In this command to the operating system, case may be important.
desktop, or (3) by selecting IDL from a menu. Presently, the IDL prompt IDL> will appear, either in the window in which the command was typed or in a text entry box at the bottom of the GUI. We must from the outset be aware that

1. Internally, IDL is not case sensitive. Except within quoted strings, A and a have the same meaning to IDL.

2. Typing the command `exit` at the IDL prompt or selecting ‘Exit’ from the File menu in the GUI ends the session.

3. (Control-C) (UNIX; Linux), (Control-Break) (Windows), or (Command-) (Macintosh) instructs IDL to abort its present activity. The activity is not aborted, however, until the statement that was active finishes. Thus, it may take a while before the IDL prompt returns.

4. (Control-D) (UNIX; Linux), (ALT-F4) (Windows), or (Command-Q) (Macintosh) effects an immediate—and computationally rough—exit from IDL to the operating system (UNIX; Linux) or desktop (Windows, Macintosh).

5. Every command line in IDL must end with a (RETURN).

6. Arguments on a line must be separated by commas. If a statement is too long to fit comfortably on one line, the dollar sign ‘$’ at the end of one line can be used to tell IDL that the statement is continued on the next line.

7. The ampersand ‘&’ can be used to separate independent statements on a single line.

8. In any command line, the semicolon ‘;’ can be used to introduce a comment. The semicolon and all following characters in the line will be ignored.

9. An initial dollar sign ‘$’ allows temporary escape from IDL to the operating system. Thus, for example, the statement `$lp file.ps` (UNIX; Linux) or `$copy file.ps lpt1:` (Windows) issued to IDL will be executed not by IDL but rather by the operating system. Control is then returned to IDL. In the distant past when working on a time-shared terminal, this feature was particularly valuable. In contemporary desktop environments, however, this feature is unnecessary, since we can easily use resources of the desktop to execute other programs without exiting from IDL.

10. Within IDL, 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 illustrative examples of IDL’s capabilities, type `demo` at the IDL prompt and then surf through the multitude of menus provided. To exit from the IDL Demo Library window, select ‘Quit’ from the File menu in that window.

### 2.2 Basic Entities in IDL

#### 2.2.1 Data Types

While individual items of data can be aggregated into quite complicated data structures, twelve fundamental data types are available in the most recent versions of IDL: 8-bit (unsigned) bytes,
2.2. BASIC ENTITIES IN IDL

16-bit signed and unsigned integers, 32-bit signed and unsigned integers, 64-bit signed and unsigned integers, 32-bit (single precision) floating-point real numbers, 64-bit (double precision) floating-point real numbers, 32-bit (single precision) floating-point complex numbers, 64-bit (double precision) floating-point complex numbers, and (character) strings. The most important types for our purposes are

- **Bytes**, which occupy 8-bits (1 byte) and store unsigned integer values ranging from 0 to 255. They are also used for logical flags, where odd values (e.g., 1) are treated as ‘true’ and even values (e.g., 0) are treated as ‘false’.

- **Integers**, which occupy 16-bits (2 bytes) and store signed integer values ranging from \(-32,768\) through 0 (whose presence explains the asymmetry) to +32,767. Integers can also be used as logical flags, with the same interpretation as described for bytes.

- **Single-precision floating-point (real) numbers**, which have real decimal values with six or seven significant digits and range in absolute value from \(10^{-38}\) to \(10^{38}\). Each such number requires 32 bits (4 bytes), 24 of which store the (binary form of the) significant digits of the value (including a sign) and 8 of which store the power of two (again including a sign) by which the significant digits must be multiplied to obtain the actual value.

- **Double-precision floating-point (real) numbers**, which have real decimal values with about fourteen significant digits and range in absolute value from \(10^{-308}\) to \(10^{308}\). Each such number requires 64 bits (8 bytes), 53 of which store the (binary form of the) significant digits of the value (including a sign) and 11 of which store the power of two (again including a sign) by which the significant digits must be multiplied to obtain the actual value.

- **Complex numbers**, which are real-imaginary pairs of either single- or double-precision real floating point values.

- **(Character) strings**, which store text. While each individual character requires 8 bits (1 byte), the variable may store a character string of length ranging from zero up to 2,147,483,647 (\(2^{31} - 1\)) characters.

Although variables can be declared to have a specific data type, we need take no particular action to define the data type of any variable. If no declaration is made, IDL will determine an appropriate data type automatically and dynamically each time a variable is assigned a new value.

We must, however, remain particularly alert to one consequence of dynamic data typing. When using IDL to perform mathematical operations, an explicit decimal point forces floating point representations. Without a decimal point, integer representations will be used. For example, \(5/2\) effects an integer division (and yields the result 2) whereas \(5.0/2.0\) effects a floating point division (and yields the result 2.5). Likewise, the expression \(200^2\) to square 200 will be evaluated using integer operations and will give an incorrect result (−25536) because its actual value exceeds 32767 while \(200.0^2\) will be evaluated using floating point operations and will give the correct result (40000.0) stored as a floating point value. To effect floating point operations, we must be certain to include a decimal point or some other cue to prevent IDL from using integer operations by default; failure to include a decimal point may result in unexpected (and incorrect) answers. If an expression uses compatible mixed data types, however, the value of the expression will (usually) be computed after converting all values to the least restrictive data type in the expression and then stored with that data type. For example, the statement \(x = 5.1/2\) will be processed by converting the 2 to 2.0, doing a floating division to obtain the result 2.55, and storing the value 2.55 as x. By way of urging caution, however, we note that the statement \(x = 5.1 + 3/2\) will return the value 6.1, not 6.6!4

---

4Because division has higher priority than addition, the integer division of 3 by 2 is evaluated before the program realizes that another piece of the expression is clearly a floating point number.
With one exception, the IDL functions string, byte, fix, long, float, double, complex, and dcomplex can be used to convert a value of any type to a string, 8-bit unsigned integer, 16-bit signed integer, 32-bit signed integer, single-precision (32-bit) real, double-precision (64-bit) real, single-precision complex, or double-precision complex value, respectively. During some conversions, e.g., the conversion of the floating point value 3.14 to the integer 3 with the expression fix(3.14) or the conversion of a complex value to a floating value (which yields simply the real part of the original complex value), information will be lost and cannot be retrieved by invoking the reverse function.

2.2.2 Variable Names

The simplest variables in IDL represent either scalars (single values) or arrays (several values, all of the same data type, referred to by a single name). Variable names can have from one to 255 characters and must start with a letter. The rest of the characters may be letters, numbers, the underscore character, or the dollar sign. Variable names are insensitive 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.

2.2.3 Assignment of Values to Variables

In IDL, the equal sign = plays the role of the assignment operator so, for example,

- the statement alpha = 3.79 creates a single-precision floating-point variable named alpha and assigns the value 3.79 to that variable, i.e., stores in a four-byte memory location named alpha the single-precision floating-point binary representation of the value 3.79,

- the statement val = 5 creates a (signed) integer variable named val and assigns the value 5 to that variable, i.e., stores in a two-byte memory location named val the 16-bit binary representation of the value 5.

- the statement name = 'David' and the statement name = "David" both create a string variable named name and assign the value David to that variable, i.e., stores the codes for each of the five characters in five consecutive bytes. Character strings can be enclosed either in single or in double quotation marks.

These assignments only place the indicated value in the identified memory location; they do not automatically display the value on the screen.

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

\[ x = \alpha + \beta \]

will add the values currently stored in the variables alpha and beta and store the result in the variable x, creating x if it does not already exist and overwriting whatever is currently stored in x—and perhaps changing its data type—if the variable already exists.

---

5 A string value (e.g., 'cook') that does not represent a valid number cannot be converted to a number.
2.2.4 Commands

In IDL 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 IDL's commands, the behavior of which is frequently influenced by the value or values of one or more keywords. Some commands—those that invoke what are called IDL procedures—simply accomplish some task (printing an array on the screen, plotting a graph, ...) and do not return a value that must be stored. These commands will be invoked with a statement having the general format

\[ \text{command, argument, argument, keyword=value, keyword=value, ...} \]

They begin with the command name, which will be followed by one or more arguments, some of which are “free” and others of which are specified by using a named keyword. Other commands—those that invoke what are called IDL functions—return a value determined from their arguments. These commands will be invoked with a statement having the general format

\[ q = \text{command}( \text{argument, argument, keyword=value, keyword=value, ... }) \]

which differs from the first-described format (1) by the assignment of the returned value to a variable—here q—and (2) by the enclosing of the arguments in parentheses. In both statements, commas are used to separate arguments. When invoking either a function or a procedure, the order of the “free” arguments is mandatory, since the position of each such argument identifies its role. Arguments specified by keywords, however, can be presented in any order, since the keyword identifies the role of the specified value. Further, arguments specified by keywords can be intermingled in any way with “free” arguments. IDL will not be confused.

Most keywords have appropriate default values, so their explicit stipulation is necessary only if the default value is unacceptable. For keywords that control logical (Boolean) values, only the values ‘true’ (1) and ‘false’ (0) are significant. For these keywords, the default is almost always ‘false’, and the shorthand specification ‘/keyword’ rather than the longer specification ‘keyword=1’ can be used to set the keyword to ‘true’.

2.3 A Sampling of IDL Capabilities

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

2.3.1 Creating and Examining Arrays

Arrays can be created within IDL or—see Section 2.6—read in from files. Individual elements of arrays can be accessed by specifying the column and row, as in `array_name[column, row]`, where the column and row indices begin at zero (repeat: indices begin at 0), or by specifying the element number, as in `array_name[element number]` where the elements in the array are numbered sequentially starting with zero and then continuing from left to right, top to bottom. The syntax

6. Note that, in IDL, the first index identifies the column and the second index identifies the row in which the element resides. IDL's convention is the reverse of the more common assignment. Note also that, starting with Version 5.0 of IDL, array indices are enclosed in square brackets to avoid ambiguity. In earlier versions (and still acceptable but not recommended), array indices and arguments of functions were both enclosed in ordinary parentheses, leaving the notation `data(3)`, for example, meaning either the fourth entry in an array `data` or the value of the function `data` at argument 3—an ambiguity that the recommended new notation avoids.

7. We here limit our discussion to arrays of one and two dimensions. See the IDL manuals for ways to create arrays of three or more dimensions.
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\(^8\)\(^9\)

\[
\begin{align*}
\text{IDL} & \geq \text{alpha} = 3.79 \\
\text{IDL} & \geq \text{print, alpha} \\
\text{IDL} & \geq \text{val} = 5 \\
\text{IDL} & \geq \text{print, val} \\
\text{IDL} & \geq \text{name} = \text{'David'} \\
\text{IDL} & \geq \text{print, name}
\end{align*}
\]

Assign floating value to \text{alpha}. Display value on screen. Assign integer value to \text{val}. Display value on screen. Assign string value to \text{name}. Display value on screen.

Note that we have here introduced the IDL command \text{print}, which invokes a \textit{procedure}. The command itself is separated from its argument by a comma.

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

\[
\begin{align*}
\text{IDL} & \geq \text{b} = [1, 0, 3, 7, 10] \\
\text{IDL} & \geq \text{print, b} \\
\text{IDL} & \geq \text{a} = [ [1,2,3], [4,5,6], [7,8,9] ] \\
\text{IDL} & \geq \text{print, a} \\
\text{IDL} & \geq \text{data} = \text{fltarr}(100, 100)
\end{align*}
\]

Create a (row) vector \text{b} of \textit{integers} and print it to the screen. To create a floating point vector, at least one value would have to have an explicit decimal point. Entries in the row are separated by commas. Assign (integer) values to a \(3 \times 3\) matrix and print it to the screen. Values in the inner sets of brackets define the rows. Note the positions of commas and brackets. Create a \(100 \times 100\) array of zeros. Remember, however, that the indices range from 0 to 99. Note also that we have introduced the command \text{fltarr}, which invokes a \textit{function}. In effect, this statement preallocates space for the array \text{data}.

Many IDL functions that operate on or produce particular data types are complemented by analogous functions that will perform the corresponding operations on the other data types. Thus, complementing the function \text{fltarr} (\textit{floating array}), the functions \text{bytarr}, \text{intarr}, \text{lonarr}, \text{dblarr}, \text{complexarr}, \text{dblcomplexarr}, and \text{strarr} create arrays of some of the other main data types. Each of these functions creates an array of the specified dimensions full of zeros or—in the case of the string array—of nulls, into which data of the specified type can then be placed.

To remind ourselves of the variables IDL knows at any particular moment and to find out something about those variables we could use the command \text{help} at the IDL prompt. Given the state created in IDL’s workspace by the commands in this section, we might invoke \text{help} with no arguments to obtain the report\(^{10}\)

\(^{8}\)IDL statements are shown on the left; comments describing the statements are shown on the right. Further, to save space, we will routinely compact IDL’s output by omitting blank lines and extra spaces.

\(^{9}\)In contrast to some of its competitors, statements in IDL do not automatically display the results of their execution and no special terminating character is necessary to suppress that output. An explicit \text{print} statement must be used if explicit output on the screen is desired.

\(^{10}\)For simplicity, we have deleted additional information about features we have not yet introduced. Note also that variable names have been displayed entirely in upper case, even though we entered them in lower case.
2.3. A SAMPLING OF IDL CAPABILITIES

IDL> help
A   INT = Array[3, 3]
ALPHA FLOAT = 3.79000
B   INT = Array[5]
DATA FLOAT = Array[100, 100]
NAME STRING = 'David'
VAL   INT = 5

which lists all known variables, their data types, and either their values (for scalars) or their dimensions (for arrays). The statement help, variable_name will display the same information about the specified variable. Many keywords—all of them identified in the IDL manuals—may be used with the help command.

Information about individual variables can also be obtained with the function size, which returns a value that must be assigned to a variable or printed. For example, the statements

IDL> q=1.5 & ln='Cook' & v=intarr(5) & m=fltarr(6,8)
IDL> print, size(q), size(ln), size(v), size(m)

create a floating scalar, a string scalar, an integer vector, and a floating array and then display information about each in order. In all cases, the first value in the (row) vector returned by size is the dimension of the entity, with 0 for scalars, 1 for vectors, and 2 for two-dimensional arrays. Also in all cases, the next to last element codes the data type with, for example, 1 for byte, 2 for integer, 3 for long integer, 4 for single-precision floating point, and 7 for string. Further, the last element in each case is the total number of elements in the variable itself (1 for scalars, 5 for a 5-element vector, 48 for a 6 x 8 array). When present, additional elements (which will be positioned between the first and the penultimate elements) convey the dimensions of the variable (second element 5 for the 5-element vector; second and third elements 6 and 8 for the 6 column x 8 row array).

Should we wish to delete some of the variables to free memory for other purposes, we could use a statement like

IDL> delvar, q, ln, v, m

which removes the variables from IDL’s list and releases the memory their values had occupied.

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

IDL> print, a[0,1]

Print the value in the first column, second row of a.

IDL> print, a[3]

Print the fourth element of a. Note that this element is the same as the element printed by the previous statement, but it is being addressed by element number rather than by its column and row indices.

IDL> a[3] = 10
IDL> print, a

Assign the value 10 to the fourth element (first column, second row) of a and print the new matrix. The alternative statement a[0,1] = 10 would achieve the same end.
Print the second row of \( a \). The asterisk is a wild card and stands for all values of the corresponding index.

Extract a \( 2 \times 2 \) matrix made from the four elements in the upper left hand corner of \( a \) and print it. The colon between two values tells IDL to use a range of values running in integer steps from the first to the last, inclusive.

Similar to the function `fltarr`, which returns an array of zeroes, the function `findgen` (floating index generator) returns a floating point array of the specified dimensions, each element of the array being the element number of that element. For example,

```
IDL> y = findgen(3)
IDL> print, y
  0.00000  1.00000  2.00000
```

Create a three-element vector with each element set equal to its element number. Note that (1) the values in the array are floating point values and (2) they start at zero.

```
IDL> z = findgen(3,3)
IDL> print, z
  0.00000  1.00000  2.00000
  3.00000  4.00000  5.00000
  6.00000  7.00000  8.00000
```

Create a \( 3 \times 3 \) matrix (nine elements total), with each element set to its element number. Again, (1) the values in the array are floating point values and (2) they start at zero.

The counterparts of the function `findgen` for some of the other data types are `bindgen`, `cindgen`, `dcindgen`, `dindgen`, `indgen`, `lindgen`, and `sindgen`.

Further, once an array has been created, it can be massaged in a variety of ways and various properties of the array can be evaluated. Note, for example the statements

```
IDL> z = indgen(3,3)
IDL> print, z
  0  1  2
  3  4  5
  6  7  8
```

Create and display a simple array.

```
IDL> print, transpose(z)
  0  3  6
  1  4  7
  2  5  8
```

Evaluate and display the transpose of \( z \).

```
IDL> print, reverse(z,1)
  2  1  0
  5  4  3
  8  7  6
```

Evaluate and display \( z \) after reversing the order of the columns (second argument = 1).

```
IDL> print, reverse(z,2)
  6  7  8
  3  4  5
  0  1  2
```

Evaluate and display \( z \) after reversing the order of the rows (second argument = 2).
2.3. A SAMPLING OF IDL CAPABILITIES

2.3.2 Simple Mathematical Manipulations

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

Further, IDL 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, ALOG10 and ALOG; the exponential exp; the hyperbolic functions cosh, sinh, and tanh; and the Bessel functions of integer order beselj, beseli, and besely. Except for the Bessel functions (see the IDL manuals), each is invoked with a statement of the form

q = function_name(x)

after the execution of which the variable q will contain the value of the corresponding function at the argument contained in the variable x. 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 atan has an alternative two-argument form,

q = atan( y, x )

which will return 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 2.3.1, the following record of an IDL session demonstrates a few mathematical manipulations:

IDL> eq1 = 8.4*sin(30.0*!pi/180.0)*sqrt(3.0)  Assign the value of an expression to the variable eq1. The variable !pi is a system variable and has the value π = 3.14159.... Other system variables are listed in the IDL manuals.

IDL> print, eq1  Print the value of eq1 to the screen. Note that it is a floating point value.
    7.27461

IDL> print, byte(eq1)  Convert eq1 to a byte value. Note the loss of information in this conversion.
    7

IDL> print, complex(eq1)  Convert eq1 to a complex number. Note that the real and imaginary parts of the answer are separated.
    ( 7.27461, 0.000000 )
24 CHAPTER 2. INTRODUCTION TO IDL

IDL> print, c
    1  2  
    10  5
IDL> d = [ [3,4], [7,4] ]
IDL> print, d
    3  4  
    7  4
IDL> e = c + d
IDL> print, e
    4  6  
    17  9
IDL> e = c * d
IDL> print, e
    3  8  
    70  20
IDL> f = c / d
IDL> print, f
    0  0  
    1  1

View again the value assigned to c in Section 2.3.1.

Create a matrix d.

Add c and d element by element.

Multiply c and d element by element.

Divide c and d element by element. Note that integer division has been done. To achieve floating division, the elements would be entered with explicit decimal points. Alternatively, at least one array might be “floated” with the function float.

2.3.3 A Further Note on Array Indices; Matrix Multiplication

Most of the time, we need not be particularly conscious of IDL’s unusual convention in which the first index in the symbol a[i,j] for an element of a matrix is regarded as the column index and the second index as the row index. With matrix multiplication, however, that unusual convention makes a difference. Both for IDL and for the rest of the world, the elements of a matrix product are defined by

\[ e_{ij} = \sum_k d_{ik} c_{kj} \] (2.1)

With this definition, the summation for IDL involves migrating down a column (i.e., from row to row in a fixed column) of the first matrix as one migrates across a row (i.e., from column to column in a fixed row) of the second matrix. That process is specified by the operator # and, for the matrices d and c defined above, yields the result

IDL> e = d # c
IDL> print, e
    17  12  
    65  60

Find the matrix product dc using IDL’s convention.

This result is not what most people will find in simply writing out the matrices and evaluating the matrix product by hand. That operation—migrating from column to column in a fixed row of the first matrix as we migrate from row to row in a fixed column of the second matrix—yields the result

\[ e_{ij} = \begin{bmatrix} d_{ij} & c_{ij} \end{bmatrix} \] (2.2)

which, in IDL, is effected by the operator ##, as shown by the statements

IDL> e = d ## c
IDL> print, e
    43  26  
    47  34

Find the matrix product dc with conventional row-column interpretation.

The second of these results can in fact also be found by transposing the result of IDL-multiplying the transposes of the two factors,
2.3. A SAMPLING OF IDL CAPABILITIES

IDL> print, transpose( transpose(d) # transpose(c) )
    43   26
    47   34

but the operator ## is distinctly simpler. IDL's unusual convention can readily cause confusion. Beware!

If (in the usual convention) we (matrix) multiply a row vector into a column vector, we obtain a number that is called the dot product of the n-dimensional vectors. That dot product can be quickly evaluated with the statements

IDL> a = [1,2,3,4,5]  
IDL> b = [6,7,8,9,10]  
IDL> dotprod = a ## transpose(b)  
IDL> print, dotprod

Create two vectors.

Evaluate dot product.

Display result.

130

Since the two vectors a and b are created as row vectors, the second factor in the dot product must be transposed to convert it into a column vector before the matrix product is evaluated.11

2.3.4 Solving Linear Equations

One particular context in which IDL's row-column switching must be respected emerges when solving simultaneous linear equations. Suppose, for example, we had the equations

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

where the matrix expression uses the usual convention for matrix multiplication. Though the approach ultimately proves to be incorrect, were we not aware of IDL's convention, we might naively be tempted to enter and verify the coefficient matrix and the vector of inhomogeneities with the statements

IDL> a = [ [2.0,5.0,3.0], [-1.0,3.0,-4.0], [1.0,-1.0,0.0] ]
IDL> print, a

2.00000 5.00000 3.00000
-1.00000 3.00000 -4.00000
1.00000 -1.00000 0.00000

IDL> b = [3.0, -4.0, 1.0]
IDL> print, b

3.00000 -4.00000 1.00000

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

IDL> ainv = invert(a)
IDL> x = ainv # b
IDL> print, x

-0.0588236 -0.294118 2.82353

We might even—apparently—confirm this solution with the statement

IDL> dotprod = transpose(a) # b.

11We have used the operator ## for multiplication. We might alternatively have used IDL’s conventions with the statement dotprod = transpose(a) # b.
IDL> print, a # x
3.00000  -4.00000  1.00000

The solution appears to be correct but, in fact, it is not, as explicit hand substitution into the original equations will reveal. Because of IDL’s internal conventions, we have in fact found and verified the solution to the equations

\[
\begin{align*}
2x_1 - x_2 + x_3 &= 3 \\
5x_1 + 3x_2 - x_3 &= -4 \\
3x_1 - 4x_2 &= 1
\end{align*}
\] (2.4)

To obtain the correct solution to the original equations, we must express the coefficient matrix with the coefficients in the first equation running down the first column, etc. That is, IDL needs a coefficient matrix that is the transpose of the one we might at first blush imagine. Starting with the matrix \(a\) already entered, we would find the correct solution with the statements

IDL> atran = transpose(a)
IDL> atraninv = invert(atran)
IDL> x = atraninv # b
IDL> print, x
0.852941  -0.147059  0.676471

This solution correctly solves the original equations, as hand substitution will confirm.

Beyond introducing another feature of IDL, this subsection underscores the importance of being careful. Even things that look right can be wrong for very subtle reasons.

### 2.3.5 A First Graph

Perhaps the most common use of the function `findgen` is to support the graphing of known univariate functions. First, we need a vector of values spanning the domain of the independent variable. In IDL, the procedure for assigning an array of values equally spaced over some interval to a variable entails (1) deciding the number of segments, \(N\), between the minimum and maximum values that are necessary to achieve a smooth plot, (2) calculating the segment size

\[
dx = \frac{(x_{\text{max}} - x_{\text{min}})}{N}\] (2.5)

and (3) generating the needed vector \(x\) containing the values \(x_{\text{min}}, x_{\text{min}} + dx, x_{\text{min}} + 2dx, \ldots, x_{\text{min}} + Ndx = x_{\text{max}}\) with the expression

\[
x = x_{\text{min}} + dx * \text{findgen}(N + 1)\] (2.6)

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 would use the IDL statements

IDL> dx = 6.0 / 100.0  \quad \text{Determine segment size when an interval of length 6 is divided into 100 segments.}
IDL> x = -3.0 + dx*findgen(101)  \quad \text{Create a vector of 101 equally spaced values ranging from \(-3.0\) to 3.0. Each element of \(x\) differs from adjacent elements by \(dx\).}
IDL> y = cosh(x)  \quad \text{Assign to each element of a vector \(y\) the hyperbolic cosine of the corresponding element of \(x\).}
Figure 2.1: Graph of the hyperbolic cosine function.

\[ \text{IDL}\geq\text{plot, } x, y \]

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. Note also that arguments are separated not only from the command but from one another by commas.

The graph created by these statements is shown in Fig. 2.1.

Additional capabilities of the command \texttt{plot} are invoked by adding one or more keywords to the command. Four of the most commonly used keywords are

- \texttt{charsize} (default 1.0), which controls the size of the characters used in labels and titles;
- \texttt{thick} (default 1.0), which controls the thickness of the line drawn for the graph;
- \texttt{ticklen} (default 0.02; 1.0 makes marks that span the range of the axes, i.e., draws a full grid), which controls the length of the major tick marks; and
- \texttt{title} (default null), which specifies a string value that will title the graph.

Thus we might produce the more readable and useful graph shown in Fig. 2.2 by embellishing the above \texttt{plot} command to read

\[ \text{IDL}\geq\text{plot, } x, y, \text{charsize}=1.25, \text{thick}=3.0, \text{title}='\text{Hyperbolic Cosine Function}' \]

Additional capabilities of the \texttt{plot} command are discussed in Section 2.10.

Note, incidentally, the power of the statement \( y = \cosh(x) \). The argument of the function \( \cosh \) is a vector and contains several values. In response to this statement, IDL 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—an explicit loop—to instruct the computer to work on each element individually. IDL automatically understands that intention with this simpler statement. Beyond simpler coding, IDL also achieves faster execution than would be achieved in more traditional coding.

2.4 Saving and Retrieving the IDL Session

The command `save` in IDL will save some or all of the variables. For example, the statement

```
save, filename='example.dat', a, b, c, /verbose
```

saves the variables `a`, `b`, and `c` in a file named `example.dat` in the default directory.\(^\text{12}\)\(^\text{13}\) If no variables are specified, then the command `save` saves all of the variables currently in the IDL workspace; if no filename is specified, the information is saved in the file `idlsave.dat`. The keyword `/verbose`, if present, causes IDL also to save an informative message about each variable.

Saved variables can be retrieved using the `restore` command. To restore the file just saved, for example, one would use the statement\(^\text{14}\)

```
restore, filename='example.dat', /verbose
```

which loads into the IDL workspace all of the variables saved in the file `example.dat` and restores their original names. The optional keyword `/verbose` prints on the screen an informative message about each variable.

The procedure `save` saves IDL variables in a compact, binary, IDL-specific, but machine-independent format. Sometimes we need to save IDL variables as ASCII text files so that other programs such as a text editor can read them. The statements

\(^{12}\)The default directory is also referred to as the current directory.

\(^{13}\)See Section 2.17 for a fuller discussion of the default directory and of means to change it.

\(^{14}\)If the default directory is unclear or ill-defined, a specification of the full path to the file may be needed.
2.4. SAVING AND RETRIEVING THE IDL SESSION

```idl
openw, 1, 'textexample.dat'
printf, 1, a
close, 1
```

write the value(s) in the variable `a` in an ASCII text file. Here, the command `openw` opens a file for writing in the default directory, gives it the name `textexample.dat`, and associates with it a Logical Unit Number (LUN), which is used to facilitate subsequent reference to this file.\(^\text{15}\) The variable `a` is then written to the file with the procedure `printf`. Finally, the file associated with LUN 1 is closed. This text file, whose contents is identical to what the statement `print, a` would have displayed on the screen, may then be read by any application that respects the structure IDL gives to the file.

Though one must be careful to have the same default directory, the file created in the previous paragraph can then be read and displayed in a later session with IDL by invoking the statements

```idl
openr, 1, 'textexample.dat'
readf, 1, a
close, 1
print, a
```

One can, of course, include the full path in the quoted string containing the file name and, thereby, avoid concern over the identity of the default directory.

In creating and reading the file `textexample.dat`, we have exploited IDL’s capacity to select a format that reflects the character of the data. If we wished to be more explicit about the formatting, we could include the keyword `format` in the `printf` statement. For definiteness, suppose `a` stores a $5 \times 5$ array of integers. Then, we might write that array to a file with the statement

```idl
openw, 1, 'textexample.dat'
printf, 1, a, format = '(5I4)'
close, 1
```

IDL uses a FORTRAN-like style for specifications.\(^\text{16}\) Here, because the data in `a` are integers, the format string uses `I` to indicate integer format and the specification `5I4` conveys that five values are to be written in each line and that each value should occupy four character positions. Since more than five values are to be written, the format specification is reused until all values have been transferred to the output. In the example, the resulting file will have five lines, each containing five values. Additional options in the format string include features to output

- floating point values, e.g., `5f10.4`, which specifies output of five floating point numbers, each occupying a field that is 10 characters wide with four characters after the decimal point.
- varying the format for the numbers, as in

  ```idl
  printf, 1, a, format = '(I4,I6,I8,I10,I12)'  
  ```

which will output five integers with fields 4, 6, 8, 10, and 12 characters wide, respectively.

\(^{15}\)The LUN must be between one and ninety-nine, since other possible values have predetermined applications, and it cannot be the same as any other user-assigned LUN. Note that the string `textexample.dat` is in quotation marks (single or double) to distinguish it from a variable. Note also that we have explicitly selected a specific LUN for this statement. It would actually be safer for us to have used the statement

```idl
openw, unit, 'textexample.dat', /get_lun
```

so that IDL would assign an available LUN, to which we would refer in subsequent statements with the variable `unit`.

\(^{16}\)Newer versions of IDL also can use C-style format specifications, but the FORTRAN style is IDL’s native style.
including text in the output, as in

\[
\text{printf, 1, a, format = '}'first ", I4, " second", I6,I8,I10,I12)'
\]

In stipulating the field width, remember to allow a character position for a possible minus sign and a character position for a decimal point (if any).

A file produced as described in the previous paragraph can, of course, be examined in any text editor. If the contents of the file are to be imported for processing in some other program, whatever statements import the file must respect its format. For example, to import this \(5 \times 5\) array into a fresh instance of IDL, one would use the statements

\[
\begin{align*}
\text{a = intarr(5,5)} & \quad \text{Define a with proper dimensions} \\
\text{openr, 1, 'textexample.dat'} & \quad \text{Open file for reading} \\
\text{readf, 1, a} & \quad \text{Read array from file} \\
\text{close, 1} & \quad \text{Close file} \\
\text{print, a} & \quad \text{Display array}
\end{align*}
\]

Note particularly that the variable \(a\) must be defined before it can be referenced in the \textbf{readf} statement. Note also that white space between values in the file triggers the separation of values when the file is read; the \textbf{readf} statement need not invoke the keyword \textbf{format}. Note finally that, had \(a\) in this most recent example been defined as a \(3 \times 3\) array, only the first nine values in the original \(5 \times 5\) array would appear in the created array.

Finally, the command \textbf{journal} saves to a file everything that is printed in the IDL command window. The one argument to this command names the file and must be a string enclosed in single quotes. For example, to save a session to show the differences among the several matrix products, we might execute the statements\(^{17}\)

\[
\begin{align*}
\text{IDL> journal, 'matrix_journal.pro'} & \quad \text{Open journal file in default directory.} \\
\text{IDL> journal, 'Showing matrix products'} & \quad \text{Write comment to the file.} \\
\text{IDL> a = [[1,3,0], [3,7,2], [0,2,4]]} & \quad \text{Create and display two matrices.} \\
\text{IDL> print, a} & \quad \text{Create and display two matrices.} \\
\text{IDL> print, b} & \quad \text{Create and display two matrices.} \\
\text{IDL> print, a * b} & \quad \text{Multiply matrices element by element.} \\
\text{IDL> print, a # b} & \quad \text{Multiply matrices using IDL's convention.} \\
\text{IDL> print, a ## b} & \quad \text{Multiply matrices using usual convention.}
\end{align*}
\]

\(^{17}\)If you are working in a Windows or Macintosh environment and you haven’t yet set the default directory to something outside the IDL directory structure, you probably should make that change before proceeding. See Section 2.17 for information about how to set the default directory.
As illustrated in the second and the last statements, (1) once journaling is in effect the statement `journal, 'text'` merely writes the quoted text to the file and (2) the statement `journal` with no arguments turns journaling off and closes the file.

The file produced by `journal` can easily be edited in a text editor. As described in Section 2.8, this file can also be executed as a command file to IDL.

### 2.5 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 IDL. The simplest is the for loop. Storing the squares of the integers from 1 to 10 in a vector, for example, is readily accomplished with the statements:

```
IDL> x = fltarr(10)
IDL> for i = 0, 9 do x[i] = (i+1)^2
IDL> print, x
```

```
1.00000 4.00000 9.00000 16.0000 25.0000
36.0000 64.0000 81.0000 100.0000
```

The first statement establishes the vector, the second steps through all values—integers from 0 through 9, inclusive—of the loop index `i`, calculating their squares and storing each in the proper element of `x`, and the third simply displays the end result. With an embellishment of the index, the loop can even be stepped by something other than one, as in

```
IDL> x = fltarr(10)
IDL> for i= 0, 9, 2 do x[i] = (i+1)^2
IDL> print, x
```

```
1.00000 9.00000 25.0000 81.0000 100.0000
```

In this version, only elements 0, 2, 4, 6, and 8 are computed; elements 1, 3, 5, 7, and 9 are left at the values established by the function `fltarr`.

If we wanted not only to fill the vector `x` but also to see the values as they are computed, we would have to incorporate both the calculation and the `print` statement in the loop, i.e., we must incorporate two statements in the loop. The proper syntax uses a `begin-end` block and is illustrated in the statements:

```
IDL> x = fltarr(10)
IDL> for i = 0, 9, 2 do begin x[i] = (i+1)^2 & print, i, x[i] & endfor
```

```
0 1.00000
2 9.00000
```

---

18Note that the vector `x` is a row vector. Its display is folded to keep the values on the screen.

19Note also that the floating point data type of `x` established by the first statement is preserved in the loop, even though the expression `(i+1)^2` will be evaluated with integer operations. Writing `x[i]=float(i+1)^2` would be marginally “cleaner”.

20Alternatively, the statement `x=findgen(10)` followed by the statement `x=(x+1.0)^2` could replace the first two statements. This alternative would execute slightly faster but would not illustrate the explicit construction of a loop.

21All `begin-end` blocks can be terminated with the simple word `end`. To aid in keeping track of which loop is being terminated at a particular point, however, one can add reminders after the word `end`, as in `endfor` used here.
Here, we use the character & to separate individual statements on a single line and the words begin and end for to “bracket” the block of instructions to be executed within the loop. Note that there is no ampersand after the word begin but there is an ampersand before the word end for. Further, for interactive entry of a for loop, IDL must see the entire loop as a single line. For especially long loops, we may find it convenient to present the instructions on several physical lines, using a $-sign at the end of each save the last to tell IDL that the multiline statement constitutes a single logical line. Note, however, that the $-signs will not replace ampersands or commas, which are needed just as if the statement were on a single physical line.

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

```idl
IDL> x = fltarr(10)
IDL> i = 0
IDL> while i lt 10 do begin i=i+1 & x[i-1] = i^2 & endwhile
IDL> print, x
     1.00000  4.00000  9.00000  16.0000  25.0000  36.0000
     49.0000  64.0000  81.0000 100.000
```

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 IDL’s operator lt (less than) to express our first logical condition. IDL, of course, possesses the standard six such operators, specifically lt (less than), gt (greater than), eq (equal to), ne (not equal to), le (less than or equal to), and ge (greater than or equal to). In addition, the operator not toggles logical values between ‘true’ and ‘false’, so, for example, the conditions i lt 10 and not (i ge 10) are equivalent. Finally, Boolean algebra on logical expressions is facilitated by the IDL operators and (logical and), or (logical or), and xor (logical exclusive or).

Logical conditions appear not only in controlling loops but also in structuring branches in a sequence of statements. As with most programming languages, IDL 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

```idl
for i = 0, 9 do if (x[i] lt 0.0) then x[i] = -x[i]
```

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

```idl
for i = 0, 9 do if (x[i] gt 10.0) then x[i] = 10.0
```

will replace all values greater than 10.0 with the value 10.0, and the statement

```idl
if a ge 0 then b = a 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 the \texttt{then} clause will be separated by ampersands and bracketed by \texttt{begin} and \texttt{endif} while multiple statements in the \texttt{else} clause will be separated by ampersands and bracketed by \texttt{begin} and \texttt{endelse}, as in

\[
\text{if condition then begin statement \& statement \& endif} \quad \text{\&}
\text{else begin statement \& statement \& endelse}
\]

2.6 Reading Data from a File

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

We illustrate the process with the file\textsuperscript{23} \texttt{radio.dat}, which was created by a FORTRAN program that simulates the decay of a radioactive material into a material that is itself unstable. We must know ahead of time that the file begins with a single line containing labels and continues with 201 additional lines, each containing four numerical values, each separated from the next with one or more spaces.\textsuperscript{24} The first value in each line is a time, and the second, third, and fourth values are the quantities of the initial material, the intermediate material, and the final (stable) material, respectively, at that time.\textsuperscript{25} We read the data from this file with the statements

\[
\begin{align*}
\text{IDL } & \ln = "" & \text{Define null string for first line.} \\
\text{IDL } & \text{quantity} = \text{fltarr} ( 4, 201 ) & \text{Create array for data.} \\
\text{IDL } & \text{openr, 1, 'radio.dat'} & \text{Open the file, assigning logical unit number (lun)=1.} \\
\text{IDL } & \text{readf, 1, ln} & \text{Read first line.} \\
\text{IDL } & \text{readf, 1, quantity} & \text{Read data, storing it in prepared array.} \\
\text{IDL } & \text{close, 1} & \text{Close file.}
\end{align*}
\]

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

\[
\begin{align*}
\text{IDL } & \text{plot, quantity[0,\ast], quantity[1,\ast], thick=3.0, charsize=1.25, } \\
& \text{xstyle=8, ystyle=8} \\
\text{IDL } & \text{oplot, quantity[0,\ast], quantity[2,\ast], thick=3.0} \\
\text{IDL } & \text{oplot, quantity[0,\ast], quantity[3,\ast], thick=3.0} \\
\text{IDL } & \text{xyouts, 2.5, 850.0, 'A'} \\
\text{IDL } & \text{xyouts, 2.5, 300.0, 'B'} \\
\text{IDL } & \text{xyouts, 40.0, 800.0, 'C'}
\end{align*}
\]

The first of these statements plots the quantity of the initial material (\texttt{quantity[1,\ast]}, the second column of the array) versus time (\texttt{quantity[0,\ast]}, the first column of the array), while the second and third statements overplot the quantities of the second and third materials versus time. The remaining three statements, in which locations are expressed in the coordinates of the graph, place the labels \( A, B, \) and \( C \) at appropriate points on the graph. The resulting plot is shown in Fig. 2.3.

\textsuperscript{23} We assume that the file \texttt{radio.dat} is stored in a directory in IDL’s search path (see Section 2.17.4). If that is not the case, then the first statement in the code below must specify the full path to the file, perhaps \$\texttt{HEAD/idl/radio.dat}. See the note at the beginning of this chapter.

\textsuperscript{24} Commas or tabs can also be used to separate values in a line.

\textsuperscript{25} Do not fail to copy this file into your directory and examine its structure with a text editor.
Four previously unmentioned features of IDL have been introduced in this example: (1) The keywords \texttt{xstyle} and \texttt{ystyle} to the \texttt{plot} command provide several possibilities. Here, with the value 8, they simply suppress the “box” style axes in favor of axes drawn on only one side of the graph. Frequently useful other values include 1 (to suppress “nice” scaling on the axis), and 4 (to suppress the axis altogether). A combination of effects results from a sum of values. (2) The command \texttt{oplot}, whose arguments are similar to those of \texttt{plot} (though many keywords available to \texttt{plot} are forbidden to \texttt{oplot}), allows plotting additional graphs without erasing the existing graph. (3) The command \texttt{xyouts}, whose arguments specify a location in the units of the graph and a text string, allows for the placement of text at any point in a graph. (4) 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.

2.7 On-line Help

IDL provides several ways for users to request help while in an IDL session. As we have already seen, we can obtain information about the current status of the IDL workspace or specific information regarding one or more of the variables stored in that workspace with the command \texttt{help}, perhaps with additional arguments specifying the particular variables about which information is desired, and with the command \texttt{size}. For information about procedures, functions, and keywords in IDL, we can type a question mark (?) at the IDL prompt or select ‘Help Contents’ from the HELP menu in the menu bar of the GUI. IDL responds with a pop-up window with two panels, the left-hand panel providing two tabs (Contents and Index) and the right-hand panel displaying text. Select the appropriate tab by clicking ML on the desired selection and then click ML on the desired entry to bring up the corresponding text in the right-hand panel. If you select the index tab, you can use the search box to locate the desired item. Hot keys to help you narrow your search abound in this help screen. Alternatively, if you know the name of the command on which you seek information, you can type the statement ‘?command’ at the IDL prompt to bring up the help window displaying information on that command directly. However you get there, the information
displayed will describe the syntax of the command—including information on allowed keywords and sample uses. The question mark provides the best route to obtain the most recent documentation on any aspect of IDL.

To display a brief description of routines in the IDL User's Library (such as the ordinary differential equation or numerical integration routines, or any Lawrence-generated files), type the statement

```
doc_library, 'NameOfRoutine'
```

where NameOfRoutine is the name of the routine whose description is desired.

## 2.8 Pro-Files

Any command file called by IDL is saved as a pro-file, so named because the file itself by convention is named with a file type `.pro`. A command file may be a procedure file, which defines a procedure controlled by arguments and effecting isolation of all variables internal to the procedure, a function file, which defines a function that returns a value to a variable (and also isolates internal variables), or simply a batch execution file or batch file, which contains a succession of statements that one might alternatively type directly at the command line (and does not isolate variables used within the file). IDL makes available many built-in procedures and functions. We have already introduced several procedures (print, printf, plot, openw, openr, ...) and functions (fltarr, invert, findgen, sin, cosh, sqrt, transpose, ...). Others include procedures for drawing error bars, invoking filters, selecting color tables, and producing animated displays. All are described in the IDL manuals.

IDL's built-in procedures and functions can be supplemented with user-written batch files, procedures or functions. The batch file is the simplest 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\(^{26}\)

```
ln = "" ; Define null string for first line
quantity = fltarr( 4, 201 ) ; Create array for data
openr, 1, 'radio.dat' ; Open the file, assigning logical unit number (lun)=1
readf, 1, ln ; Read past first line
readf, 1, quantity ; Read data, storing it in prepared array.
close, 1 ; Close file
plot, quantity[0,*], quantity[1,*], thick=3.0, charsize=1.25, $
     xstyle=8, ystyle=8
oplot, quantity[0,*], quantity[2,*], thick=3.0
oplot, quantity[0,*], quantity[3,*], thick=3.0
xyouts, 2.5, 850.0, 'A'
xyouts, 2.5, 300.0, 'B'
xyouts, 40.0, 800.0, 'C'
```

(Introduced by a semicolon, comments can be included in the file and will be ignored by IDL.) Then, after this file has been stored with the name `plotradio.pro` (say), either in the default directory or in a directory that is in IDL's search path (see Section 2.17 and the Local Guide), the simple statement\(^{27}\)

```plaintext
We again assume that the file radio.dat resides in a directory in IDL's search path. For more detail, see footnote 23.

\(^{27}\)When no file type is specified, the file type `.pro` is assumed by default.
CHAPTER 2. INTRODUCTION TO IDL

IDL> @plotradio

will instruct IDL to execute all these statements in turn and ultimately produce the graph we have already shown in Fig. 2.3. Especially when we are doing several similar tasks or developing and debugging an extended sequence of statements, the creation of a batch file, 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 PRO or FUNCTION on the first non-blank line of the file informs IDL that a user-defined procedure or function follows. As with batch files, every user-defined procedure or function file should be named with the file type ‘.pro’. For example, the statements

```idl
PRO luplot, funct=tempfunct, N, start, stop
  ;+ Procedure luplot is passed a function identified by the
  ; keyword 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.
  ;-
  dx = ( stop - start ) / float(N) ; Calculate increment
  x = start + dx*findgen(N+1) ; Create vector of independent variable
  y = call_function( tempfunct, x ) ; Evaluate dependent variable
  plot, x, y, charsize=1.25, thick=3.0, ticklen=1.0 ; Plot graph
RETURN
END
```

define a procedure that plots a function passed as a string value `tempfunct` assigned to a user-declared keyword `funct`, with the graph extending over an interval determined by the “free” arguments `start` and `stop` and the interval broken into a number of segments specified by the “free” argument `N`. The procedure is invoked by typing `luplot` followed by the requisite arguments. The procedure will be automatically compiled when it is first called.\(^{28}\) For example, if the file `luplot.pro` is stored in the default directory or in a directory in IDL’s search path, the statement

IDL> luplot, funct='cos', 200, 0.0, 10.0

is a legal call to the procedure and would generate a graph of the cosine function \(\cos(x)\) over the interval \(0 \leq x \leq 10\), dividing the interval into 200 segments (i.e., plotting 201 points uniformly distributed over the interval).

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

\[
y(x) = \frac{a^2}{b^2 + (x - x_0)^2}
\] (2.7)

for particular values of \(a\), \(b\), and \(x_0\). We would begin by writing a function pro-file defining the Lorentz lineshape, perhaps using the coding

\(^{28}\)Note, however, that we will have to be more explicit about compiling `luplot` if our first execution reveals an error. Our natural instinct would be to fix the error in the file, save the file, and then rerun `luplot`. Unfortunately, at this point, IDL already has knowledge of the (erroneous) version and will not automatically recompile the new version. We will need to invoke the statement `.compile luplot` explicitly to correct IDL’s understanding of the procedure.
2.8. PRO-FILES

Figure 2.4: The Lorentz lineshape.

FUNCTION lineshape, x
  ; 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
  RETURN, a^2/(b^2 + (x-x0)^2)
END

Once this file has been stored with the name lineshape.pro (say) in the default directory or in
IDL’s search path, we simply use a statement like

IDL> luplot, funct='lineshape', 200, -10.0, 10.0

to call luplot once again. In a second or two, the graph of the Lorentz lineshape shown in Fig. 2.4
will appear on the screen.

With the approach of the previous paragraph, we would have to edit the function pro-file and
recompile it if we wanted to graph the Lorentz line shape for a different set of parameters. Because
of the way luplot is written, however, we cannot simply add these parameters as arguments to
lineshape. Were we to adopt that route, we would want to set the parameters a, b, and x0 at IDL’s
command level and have them known not within luplot itself but within a routine that luplot
calls. The call statement in luplot does not include the passing of parameters to lineshape. To
achieve this passing of a parameter from IDL’s command level into a routine that is not directly
called from the command level, we must make use of named common. In essence, at the command
level, we define a fixed and named area of memory into which we place the desired values of the
parameters at the command level and from which the parameters can be read by any routine that
also identifies the same named area. To this end, we might decide to name the common area `param` (say) and then rewrite `lineshape` to be

```idl
FUNCTION lineshape1, x
  ;+  ; The function lineshape returns the value of the Lorentz
  ; distribution for given argument x. The parameters a, b, and
  ; x0 must be placed in the common area named param prior to invoking
  ; lineshape1, which then reads the values from that same common
  ; area.
  ;-
  common param, a, b, x0
  RETURN, a^2/(b^2 + (x-x0)^2)
END
```

and store it in the file `lineshape1.pro`. Here, the `common` statement tells the function that it is to find the values of the variables `a`, `b`, and `x0` in the common area named `param`. Then, we invoke `luplot` with this function using the statements

```
IDL> common param, a, b, x0         Define common area.
IDL> a = 1.0 & b = 1.0 & x0 = 2.0    Set parameters.
IDL> luplot, funct='lineshape1',      Graph function
   200, -10.0, 10.0
```

and produce again the graph shown in Fig. 2.4. This time, however, our approach is more flexible, since we can more easily change the values of the parameters and produce another graph.

All of the above pro-files have made use of IDL's capacity for self-documentation. Beyond including comments introduced by semicolons, we have judiciously placed the two lines `;+` and `;->` before and after the descriptive comments at the beginning of each file. Those special lines bracket comments that will be printed when the procedure `doc_library` is invoked with a statement like

```
IDL> doc_library, 'luplot'
```

to extract these comments as help messages. _The wisdom of documenting user-written procedures and functions thoroughly cannot be overstressed, and this particular feature of IDL makes it easy to keep the documentation coordinated with—and in the same file as—the coding itself._

Finally, journal files produced as described in Section 2.4 may also be executed as batch files. When IDL creates a journal file, it automatically puts a semicolon before every line that does not contain a statement. Thus, the file acts as a command file and can be executed using the symbol `@` before the file name.

One distinction between batch pro-files on the one hand and procedure and function pro-files on the other hand must be carefully noted. In Section 2.5, we described the `interactive` statement

```
for i = 0, 9, 2 do begin x[i] = (i+1)^2 & print, i, x[i] & endfor
```

which we might have spread over more than one line by writing

```
for i = 0, 9, 2 do begin $  
  x[i] = (i+1)^2 & $  
  print, i, x[i] & $  
endfor
```
2.9. EIGENVALUES AND EIGENVECTORS

In a batch pro-file, this statement would have to appear exactly this way, including the dollar signs and ampersands. If this statement were to appear in a procedure or function pro-file, however, it would be written in the form

\[
\text{for } i = 0, 9, 2 \text{ do begin}
\]
\[
x[i] = (i+1)^2
\]
\[
\text{print, } i, x[i]
\]
\[
\text{endfor}
\]

Neither the dollar signs nor the ampersands would be included. In function and procedure pro-files, starting a new line is sufficient to start a new statement, though the dollar signs and ampersands can still be used to separate individual statements on a single line and indicate the continuation of a single statement on a second line. In the last form above, the keywords do and begin must be on the same line (or a dollar sign will be necessary, even in the form for a procedure or function pro-file).

2.9 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 Lawrence-written procedure, \texttt{lueigen}, takes as arguments the name of the (symmetric) matrix whose eigenvalues and eigenvectors are desired and the names of the two arrays in which the eigenvalues and eigenvectors are to be stored. Thus, to find eigenvalues and eigenvectors, we need only create the matrix and invoke \texttt{lueigen} with statements like\textsuperscript{29}

\[
\begin{align*}
\text{IDL} & \text{> } a = \begin{bmatrix} 1,3,0 \end{bmatrix}, \begin{bmatrix} 3,7,2 \end{bmatrix}, \begin{bmatrix} 0,2,4 \end{bmatrix} \\
\text{IDL} & \text{> print, } a \\
& \quad 1 \quad 3 \quad 0 \\
& \quad 3 \quad 7 \quad 2 \\
& \quad 0 \quad 2 \quad 4 \\
\text{IDL} & \text{> lueigen, } a, \text{ evals, evecs} \\
\text{IDL} & \text{> print, evals} \\
& \quad -0.389315 \quad 3.44685 \quad 8.94246 \\
\text{IDL} & \text{> print, evecs} \\
& \quad -0.891227 \quad 0.310657 \quad 0.330465 \\
& \quad 0.412732 \quad 0.253377 \quad 0.874902 \\
& \quad -0.188062 \quad -0.916129 \quad 0.354035 \\
\end{align*}
\]

Each column of the array \texttt{evvecs} is one of the normalized eigenvectors of the symmetric matrix \(a\). To make sure we entered everything properly, we can check the result using the eigenvector equation

\[
a x = \lambda x
\]

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

\[
\begin{align*}
\text{IDL} & \text{> lambda } = \text{ evals}[0] \\
\text{IDL} & \text{> x } = \text{ evecs}[0, \star] \\
\text{IDL} & \text{> lhs } = a \ #\ # x \\
\text{IDL} & \text{> rhs } = \text{ lambda } * x \\
\text{IDL} & \text{> print, } [ \text{ lhs, rhs } ] \\
& \quad 0.346966 \quad 0.346966 \\
& \quad -0.160683 \quad -0.160683 \\
& \quad 0.0732154 \quad 0.0732154 \\
\end{align*}
\]

\textsuperscript{29}See the note at the beginning of this chapter.
CHAPTER 2. INTRODUCTION TO IDL

Table 2.1: The procedure lueigen.pro

```idl
PRO lueigen, a, eigenvalues, eigenvectors
>+ ; Procedure lueigen uses the IDL routines trired and triql to find
+ ; the eigenvalues and eigenvectors of a real, symmetric matrix, a.
+ ; Only argument a is supplied as input. The arguments eigenvalues
+ ; and eigenvectors identify the variables into which lueigen places
+ ; its results. Each column of the array eigenvectors is a normalized
+ ; eigenvector of the matrix a. The eigenvector in the ith column
+ ; corresponds to the eigenvalue in the ith element of eigenvalues.
+ ;-
dummy = a ; Protect a
trired, dummy, diagelem, offdiagelem ; Find tridiagonal form
triql, diagelem, offdiagelem, dummy ; Find eigenvectors, values
eigenvalues = diagelem ; Place eigenvalues for output
eigenvectors = transpose(dummy) ; Make eigenvectors columns
RETURN
END
```

Inspection shows that the two columns are the same within differences attributable to internal roundoff.

Behind the scenes, lueigen uses two of IDL’s built-in routines for finding eigenvalues and eigenvectors. First, lueigen invokes trired to reduce the symmetric matrix to an equivalent tridiagonal form which has non-zero elements only in positions on and immediately adjacent to the main diagonal and which, more significantly, has the same eigenvalues and eigenvectors as the original matrix. Then, lueigen invokes triql to compute the eigenvalues and eigenvectors for the transformed matrix. The coding for lueigen, which is implemented as a procedure pro-file, is shown in Table 2.1. In this procedure, trired takes the original matrix—moved to dummy to prevent a from being lost—as input and returns two vectors, diagelem and offdiagelem, containing the diagonal and off-diagonal elements of the equivalent tridiagonal matrix. Then triql takes the diagonal and off-diagonal elements as input and produces the eigenvalues, which are returned in diagelem, and eigenvectors, which are returned in the rows of dummy.

Other IDL routines for dealing with eigenvalues and eigenvectors include eigenql (for real, symmetric arrays), eigenvect (for real, non-symmetric arrays), elmhes (to reduce a real, non-symmetric array to a special form known as the upper-Hessenberg form) and hqr (for finding eigenvalues—but not eigenvectors—of matrices in upper-Hessenberg form). Details about these routines can be found in the IDL manuals.

2.10 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. The basic command for producing such a graph is plot, which—in its simplest form—has the syntax

```
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.
2.10.1 The Basic Strategy

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

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

(2.9)

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 IDL statements

IDL> dz = 8.0/100.0
IDL> z = -4.0 + dz * findgen(101)
IDL> B = (1.0 + z^2)^(-1.5)
IDL> plot, z, B, title='Magnetic Field on z axis', $ 
IDL> xtitle='Dimensionless Position', ytitle='Magnetic Field', $ 
IDL> thick=3.0, charsize=1.25, ticklen=1.0

where the keywords title, thick, charsize, and ticklen have the effects described in Section 2.3.5 and the newly introduced keywords xtitle and ytitle label the x and y axes, respectively. This sequence of statements will produce the graph shown in Fig. 2.5.

2.10.2 Plotting Several Graphs on One Set of Axes

IDL’s procedure oplot provides the capacity 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. 2.6 is produced with the statements

IDL> plot, z, sin(z), title='Undamped Sine Wave', $ 
IDL> xtitle='Dimensionless Position', ytitle='Magnetic Field', $ 
IDL> thick=3.0, charsize=1.25, ticklen=1.0

IDL> plot, z, sin(2*z), title='Damped Sine Wave', $ 
IDL> xtitle='Dimensionless Position', ytitle='Magnetic Field', $ 
IDL> thick=3.0, charsize=1.25, ticklen=1.0
Table 2.2: Available line styles for use with the keyword linestyle.

<table>
<thead>
<tr>
<th>Value for linestyle</th>
<th>Line Style Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Solid</td>
</tr>
<tr>
<td>1</td>
<td>Dotted</td>
</tr>
<tr>
<td>2</td>
<td>Dashed</td>
</tr>
<tr>
<td>3</td>
<td>Dash Dot</td>
</tr>
<tr>
<td>4</td>
<td>Dash Dot Dot Dot Dot</td>
</tr>
<tr>
<td>5</td>
<td>Long Dashes</td>
</tr>
</tbody>
</table>

Figure 2.6: Damped and Undamped sine waves.

IDL> dx = 20.0/100.0
IDL> x = dx*findgen(101)
IDL> sine = sin(x)
IDL> dampsine = exp(-x/10.0) * sin(x)
IDL> plot, x, sine, title='Damped and Undamped Sine Waves', thick=3.0, charsize=1.25
IDL> oplot, x, dampsine, linestyle=4, thick=3.0

The keyword linestyle, which can be used both in the command plot and in the command oplot, specifies a line style other than the default (a solid line). Table 2.2 enumerates the available line styles.

2.10.3 Polar Plots and Custom Axes

IDL is also capable of producing polar plots and of using different axis styles instead of the standard box style. Polar plots are accomplished by specifying the radial (r) and polar (θ) coordinates in the execution of the plot command along with the keyword /polar.\(^\text{30}\) For example, to graph the

\(^{30}\)The keyword polar is a Boolean (logical) variable, and the specification /polar is equivalent to the specification polar = 1; either sets the keyword to ‘true’.
cardioid defined in polar coordinates by the equation
\[ r(\theta) = a(1 - \cos \theta) \] (2.10)
and shown in Fig. 2.7, we might execute the statements

IDL> dt = 2.0*!pi/100.0
IDL> theta = dt*findgen(101)
IDL> a = 3.0
IDL> r = a*(1.0 - cos(theta))
IDL> plot, r, theta, /polar,
   xstyle=4, ystyle=4,
   subtitle='The Cardioid',
   thick=3.0, charsize=1.25
IDL> axis, 0.0, 0.0, xaxis=0
IDL> axis, 0.0, 0.0, yaxis=0

Here, in the command plot, the radial coordinate comes first and the azimuthal coordinate comes second, the keywords xstyle and ystyle have been set to the value 4 to suppress axes that would otherwise be drawn (inappropriately) as a box around the entire figure, and we have used the keyword subtitle, which differs from the keyword title because it uses a slightly smaller font and it places the title at the bottom of the graph. Having suppressed the automatically drawn axes in the command plot, we then invoke the command axis to draw a new x axis, centered at the origin, and a new y axis, also centered at the origin. In the first invocation of axis, the keyword xaxis is set to indicate that an x axis is to be drawn and set specifically to zero so that the labels are placed below the axis and the ticks are drawn upwards from the axis. In the second invocation, the keyword yaxis is set to indicate that a y axis is to be drawn and set specifically to zero so that the labels are placed to the left of the axis and the ticks are drawn to the right from the axis.\(^{31}\)

The graph of Fig. 2.7 is, however, distorted because, despite the equal ranges on the two axes, equal coordinate increments in the two directions do not correspond to equal lengths on the two axes. The graph is stretched horizontally by the aspect ratio of the display. Adjusting this aspect ratio is discussed at some length in Section 2.16.5, where several additional keywords available in some of the graphics commands, including the plot command, are introduced.

\(^{31}\)Setting xaxis to 1 places the labels above and draws the tick marks down from the x axis; setting yaxis to 1 places the labels to the right of and draws the tick marks to the left from the y axis.
2.10.4 Multiple Plots On a Page

Sometimes we wish to plot several graphs on one page. IDL gives us substantial flexibility in formatting the layout of these graphs by providing the system variable \texttt{!p.multi}, which controls how many plots appear, how they are formatted, and in what order they are drawn. \texttt{!p.multi} is a five-element integer vector. For our purposes, the first three elements are the important ones. These are

- \texttt{!p.multi[0]}, which keeps track of which segment of the graph window will be used next,
- \texttt{!p.multi[1]}, which specifies the number of plot columns per page,\footnote{If \texttt{!p.multi[1]} \leq 0, one column of plots is assumed.} and
- \texttt{!p.multi[2]}, which specifies the number of plot rows per page.\footnote{If \texttt{!p.multi[2]} \leq 0 one row of plots is assumed.}

If more than two plots are placed in either the horizontal or vertical directions, then the character size used to annotate the graph is decreased by fifty percent.

As a quick example using \texttt{!p.multi}, suppose we wish to plot the sine and cosine functions along with their analogous hyperbolic functions. If we set \texttt{!p.multi[0]=0}, \texttt{!p.multi[1]=2} and \texttt{!p.multi[2]=2}, we will be able to plot two columns and two rows of graphs on one page. The first plot is drawn in the upper left-hand corner of the plot window, and the following plots are placed left to right, top to bottom. The plot shown in Fig. 2.8 is produced with the statements

\begin{verbatim}
IDL> dz = 8.0/100.0
IDL> z = -4.0 + dz*findgen(101)
IDL> v = sin(z) & w = cos(z) & x=sinh(z) & y=cosh(z)
IDL> plot, z, v, title='Sine', thick=3.0
IDL> plot, z, w, title='Cosine', thick=3.0
IDL> plot, z, x, title='Hyperbolic Sine', thick=3.0
IDL> plot, z, y, title='Hyperbolic Cosine', thick=3.0
\end{verbatim}

Note (1) that the ampersand ‘&’ separates individual statements placed on a single line, (2) that \texttt{!p.multi} is reset to specify one plot per page at the end of this sequence of statements, and (3) that the variable \texttt{!p.multi[0]} is adjusted automatically as the graphing proceeds but can be reset to zero at any time so that the next invocation of \texttt{plot} will erase the current display and restart the graphing at the upper left segment of the graph window.

2.10.5 Plotting Experimental Data

As a final example of two-dimensional graphing, we describe IDL’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. IDL 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 \texttt{rcdata.dat} containing the data was created with a text editor and stored in the default directory.\footnote{See footnote 23.} 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).\footnote{Do not fail to copy this file from \texttt{$HEAD/idl} into your directory and examine its structure with a text editor.} The data are read into the IDL workspace with the statements
Figure 2.8: Graphs of sine, cosine, hyperbolic sine, and hyperbolic cosine on the interval $-4 \leq x \leq 4$.

```
IDL> openr, 1, 'rcdata.dat'
IDL> data = fltarr(2,21)
IDL> readf, 1, data
IDL> close, 1
```


Next, we plot the data using the statement

```
IDL> plot, /xlog, data[0,*], data[1,*], psym = 4, charsize=1.25, $
IDL> yrange=[0.0, 1.1], ystyle=1, $
IDL> title='Experimental Data for High Pass Filter', $
IDL> xtitle='Frequency', ytitle='Gain (Vout/Vin)'
```

This statement plots 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) and specifies with the keyword `xlog` that a logarithmic scale be used on the $x$ axis but preserves a linear scale on the $y$ axis. Using the coding in Table 2.3, the keyword `psym` specifies which of several symbols is to be placed at each data point. We have also used the keyword `yrange`, which is set to a two-element vector, to wrest control of the range plotted on the $y$ axis from IDL by specifying the minimum (first element) and maximum (second element) of the desired range. Further, we have set the (already introduced) keyword `ystyle` to 1 to force IDL not to round the specified limits to “nice” values. 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. 2.9.

A graph of experimental data, however, is not complete without error bars. IDL supplies a routine called `errplot` to overplot error bars on a graph. The statements used to add the error bars are

```
IDL> unc = data[1,*]*0.05
IDL> low = data[1,*] - unc
IDL> high = data[1,*] + unc
IDL> errplot, data[0,*], low, high, width=0.02
```

Compute uncertainties at 5%. Compute lower values. Compute upper values. Draw error bars.

---

35 The keyword `ylog` can be set to true to specify a logarithmic scale on the $y$ axis. The older commands `plot_o1` for log $x$ and linear $y$, `plot_l1` for linear $x$ and log $y$, and `plot_oo` for log $x$ and log $y$ axes have been superseded with the keywords `xlog` and `ylog`.

37 For a more complete description of the `errplot` procedure, see the IDL manuals.
Table 2.3: Available symbols for use with the keyword psym.

<table>
<thead>
<tr>
<th>Value of psym</th>
<th>Symbol Drawn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Plus sign</td>
</tr>
<tr>
<td>2</td>
<td>Asterisk</td>
</tr>
<tr>
<td>3</td>
<td>Period</td>
</tr>
<tr>
<td>4</td>
<td>Open diamond</td>
</tr>
<tr>
<td>5</td>
<td>Open triangle</td>
</tr>
<tr>
<td>6</td>
<td>Open square</td>
</tr>
<tr>
<td>7</td>
<td>Times sign</td>
</tr>
<tr>
<td>8</td>
<td>User-defined symbol; see <code>usersym</code> procedure in the IDL manuals</td>
</tr>
</tbody>
</table>

Procedure `errplot` requires as input three arguments, the first a vector containing the horizontal—yes, horizontal—coordinates of the plotted points, the second a vector containing the lower bound of the value at each data point, and the third a vector containing the upper bound of the value at each data point. Assuming a 5% uncertainty, we calculate first the uncertainty and then the upper and lower values for each data point. (These vectors might alternatively be read from a properly expanded input file.) Then, we invoke `errplot`. A third optional keyword, `width`, is included to specify the length of the horizontal bars drawn at the top and bottom of the vertical error bars as a fraction of the width of the plot window. (The default value is 0.01.)

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

$$G = \frac{V_{out}}{V_{in}} = \frac{1}{\sqrt{1 + \left(\frac{f_0}{f}\right)^2}}$$

where $f_0$ is defined as $f_0 = \frac{1}{2\pi RC}$ (2.11)

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 `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 `temp`, i.e.

$$f = 10^{\text{temp}} = (e^{2.3026})^{\text{temp}} = e^{2.3026*\text{temp}}$$

(2.12)

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

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

```
IDL> r = 15.0e3                      Set value of $R$.
IDL> c = 0.0442e-6                   Set value of $C$.
IDL> f0 = 1/(2.0*pi*r*c)             Calculate implied reference frequency.
IDL> temp = findgen(76)/25.0 + 1.0   Create vector with values from 1.0 to 4.0.
IDL> f = exp(2.3026 * temp)          Create frequencies from $10^1$ to $10^4$.
IDL> gtheory = 1.0/sqrt(1.0 + f0^2/f^2) Calculate theoretical curve.
IDL> oplot, f, gtheory, linestyle=2, thick=3.0 Overplot theoretical curve.
```

Here, we have used the values $R = 15$ kΩ and $C = 0.0442$ µF 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

IDL> plots, [500.0, 900.0], [0.3, 0.3], linestyle = 2, thick=3.0
IDL> xyouts, 1000.0, 0.3, 'Theoretical'
IDL> plots, 700.0, 0.25, psym = 4
IDL> xyouts, 1000.0, 0.24, 'Experimental'

The first of these statements uses plots to draw a line from data coordinate (500.0, 0.3) to the data coordinate (900.0, 0.3). The second then uses xyouts to apply the label “Theoretical” 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. 2.9 have been compiled into the file plotrc.pro, which is available where all other files mentioned in this chapter are found.38

2.11 Making Hard Copy

2.11.1 ... of Text

To print mathematical results from IDL when IDL is being run on a device with X-window output, it is easiest to use the cutting and pasting capabilities of the X-window system/clipboard to copy the information from the window running IDL to the input window of the text editor in use. Alternatively (and necessarily on devices without X-window capability), the command journal described in Section 2.4 will create a file containing everything written to the screen in an IDL session, which file can then be edited to delete the unwanted portions.

38See footnote 23.
2.11.2 ... 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 in IDL requires two steps. First, we must create a file containing a PostScript description of the plot. 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 IDL and is accomplished with the statements

```idl
set_plot, 'ps'
device, filename='filename.ps'
```

Specify PostScript output.

Identify file for output. The file will be created in the default directory.

Create graph, which goes to file, not to screen.

Close file.

Restore X-window output.

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 keywords `charsize` and `thick` in the plot command may change when a new output device is selected. Thus, when these (and perhaps other) keywords are invoked, some experimentation may be necessary to produce a printed graph that exactly reflects what may have been displayed on the screen. Further, be particularly aware that, once IDL has been told to produce output in PostScript format, *nothing* will be directed to the plot window on the screen—even as statements are being executed in the creation of the file. The last statement in the above sequence restores normal interaction between the program and plot window on the screen.

For example, to obtain a file (in the default directory) containing a PostScript description of the first graph in Section 2.10.1, we would execute the statements

```idl
IDL> set_plot, 'ps'
IDL> device, filename='bfield.ps'
IDL> plot, z, B, title='Magnetic Field on z-axis', 
   \$ xtitle='Dimensionless Position', ytitle='Magnetic Field'
IDL> device, /close
IDL> set_plot, 'x'
```

Note that, here—and even more—when the desired graph requires many statements, we should adopt a strategy of constructing and debugging the graphical statements in a batch file. That file can be edited and re-executed until the on-screen display is as we wish. Then, we simply change the device, define an output file, and re-execute the batch file to produce the PostScript file, saving the labor of and potential errors in retyping the statements that produce the display.

2.12 Graphing Scalar Functions of Two Variables

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

\[ z(x, y) = \sin(2\pi x) \sin(3\pi y) \]  \hspace{1cm} (2.13)

---

39 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 document.

40 This statement, of course, supposes that IDL is being run on a device with X-window output. The proper statement to return to screen display in Microsoft windows is `set_plot, 'win';` on a Macintosh computer, the statement is `set_plot, 'mac'.` Other less frequently used possibilities are described in the IDL manuals.
which defines the shape of the $[2, 3]$ mode of oscillation for a square membrane, or

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

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

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

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 IDL processes arrays element by element, a single statement, which would have the forms

$$z = \sin(2.0*\pi*x) \cdot \sin(3.0*\pi*y)$$
$$I = (\sin(x_i)/x_i)^2 \cdot (\sin(\eta)/\eta)^2$$

for the functions in Eqs. (2.13) and (2.14), 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 surface, contour, shade_surf, xsurface, and show3.

### 2.12.1 A Preliminary: The Procedure lugen_grid

Similar to the IDL procedure findgen but functioning in two dimensions, the Lawrence-written procedure lugen_grid 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. This procedure has six arguments. The two “free” arguments name the arrays into which lugen_grid should place values of $x$ and $y$; the four other arguments, which are specified by the keywords xrange, nx, yrange, and ny, supply the interval over which each independent variable is to extend and the number of segments into which each interval is to be divided. Thus, for example, the statement

IDL> lugen_grid, x, xrange=[-3.0, 3.0], nx=3, y, yrange=[-1.0, 1.0], ny=4

creates the arrays

---

41With 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.

42A full listing of lugen_grid.pro is included in Section 2.A.

43The procedure lugen_grid admits a third “free” argument (z) and two additional keywords (zrange, nz). These entities are optional and will be introduced more explicitly when we first need them.

44Note that the division of the $x$ interval $[-3.0, 3.0]$ into three segments results in four values in the array (i.e., $-3, -1, 1, 3$). In the same manner, the division of the $y$ interval $[-1.0, 1.0]$ into four segments results in five values ($-1, -0.5, 0, 0.5, 1$).
CHAPTER 2. INTRODUCTION TO IDL

IDL> print, x
-3.00000 -1.00000 1.00000 3.00000
-3.00000 -1.00000 1.00000 3.00000
-3.00000 -1.00000 1.00000 3.00000
-3.00000 -1.00000 1.00000 3.00000

IDL> print, y
-1.00000 -1.00000 -1.00000 -1.00000
-0.50000 -0.50000 -0.50000 -0.50000
0.00000 0.00000 0.00000 0.00000
0.50000 0.50000 0.50000 0.50000
1.00000 1.00000 1.00000 1.00000

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 a single statement, say

IDL> z = x^2 + y^2

IDL> print, z
10.0000 2.00000 2.00000 10.0000
9.25000 1.25000 1.25000 9.25000
9.00000 1.00000 1.00000 9.00000
9.25000 1.25000 1.25000 9.25000
10.0000 2.00000 2.00000 10.0000

Here, each entry in the array z is the value of the function z 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 lugen_grid and the nature of its output, we above chose an example that is far too small to produce a useful display of any function. To prepare for displaying the functions of Eqs. (2.13) and (2.14), then, let us create rather larger arrays z and I with the statements

IDL> lugen_grid, x, xrange = [ 0.0, 1.0 ], nx = 40, $
IDL> y, yrange = [ 0.0, 1.0 ], ny = 40
IDL> z = sin( 2.0*pi*x ) * sin( 3.0*pi*y )
IDL> lugen_grid, xi, xrange=[-3.0*pi, 3.0*pi], nx=49, $
IDL> eta, yrange=[-3.0*pi, 3.0*pi], ny=49
IDL> I = (sin(xi)/xi)^2 * (sin(eta)/eta)^2

(The use of 49 divisions—rather than 50—in the third statement is designed to avoid the appearance of the explicit value 0 in either xi or eta, thereby avoiding several divisions by zero in evaluating I.\footnote{IDL actually handles division by zero in a sensible way. Rather than terminate execution, IDL returns the value \texttt{Inf} and goes on to the next statement presented to it. Since the command \texttt{plot}, for example, is able to ignore \texttt{Inf}, division by zero normally has no more significant impact than the creation of a gap in the resulting graph.})

2.12.2 Surface Plots: The Procedure surface

The simplest IDL procedure for displaying two-dimensional arrays is \texttt{surface}, and the statements

IDL> surface, z
IDL> surface, I
2.12. GRAPHING SCALAR FUNCTIONS OF TWO VARIABLES

Figure 2.10: A surface representation of the function $z$.

Figure 2.11: A surface representation of the function $I$.

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 keywords, however, will create a more refined display. Executing the more elaborate statements\(^6\)

IDL> surface, z, x, y, xtitle='x', ytitle='y', ztitle='z', $\text{title}=[2,3] \text{ Mode}$, charsize=2.0
IDL> surface, I, xi, eta, $\text{title}='!4n!3', \text{ytitle}='!4g!3', \text{ztitle}='I'$, $\text{title}='\text{Diffraction\!C\!C (Square\!C\!C Aperture)}', \text{charsize}=2.0

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 second and third “free” (and optional) arguments ($x$ or $xi$; $y$ or $eta$) tell IDL to use the elements of these arrays to scale the two axes. Thus, the axes will be labeled to show a range of $0.0 \leq x,y \leq 1.0$ or $-3\pi < \xi,\eta < 3\pi$. The more complete graphs produced by these last statements are shown in Figs. 2.10 and 2.11. Note that, in the perspective shown, the $x$ and $\xi$ axes run positively to the east northeast, the $y$ and $\eta$ axes run positively to the west northwest, and the $z$ and $I$ axes run positively to the north (vertical).

The procedure `surface` also admits two optional keywords that control the angle from which the surface is viewed. The keyword `ax` specifies a rotation, in degrees, about the $x$ axis and the keyword `az` specifies a rotation, also in degrees, about the $z$ axis. The default values are $ax=az=+30^\circ$. To change the rotation about the $z$ axis to $-40^\circ$, we would use the statements

IDL> surface, z, x, y, az = -40, xtitle='x', ytitle='y', ztitle='z', $\text{title}=[2,3] \text{ Mode}$, charsize=2.0
IDL> surface, I, xi, eta, az = -40, $\text{title}='!4n!3', \text{ytitle}='!4g!3', \text{ztitle}='I'$, $\text{title}='\text{Diffraction\!C\!C (Square\!C\!C Aperture)}', \text{charsize}=2.0

Figures 2.12 and 2.13 show the resulting rotated views for comparison with Fig. 2.10 and Fig. 2.11.

\(^6\)In the second of these statements, we have invoked IDL’s capability to change fonts (e.g., $!4$ and $!3$ in `xtitle` and `ytitle`) and start a new line (`!C` in `title`). These features will be more fully described in Section 2.16.1.
Examination of Figs. 2.11 and 2.13 suggests that the diffraction 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 yet another keyword to specify a vertical scale that ignores the central peak. For example, the statement

\[
\text{IDL> surface, I, xi, eta, zrange=[0.0,0.1], }
\text{IDL> xtitle='!4n!3', ytitle='!4g!3', ztitle='I', }
\text{IDL> title='Diffraction!C!C(Square!C!C Aperture)', charsize=2.0}
\]

which forces the \( I \) axis to be confined to the range \( 0.0 \leq I \leq 0.1 \) (and the central peak simply goes off scale). The resulting display, shown in Fig. 2.14, reveals the secondary peaks in the diffraction pattern and even hints at some of those not on the main axes.

Other keywords that can be used with the \texttt{surface} procedure are described in the IDL manuals.

### 2.12.3 Contour Plots: The Procedure \texttt{contour}

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

\[
\text{IDL> contour, z}
\]

Exploiting a few keywords, however, we could produce a more meaningful contour map with the statement

\[
\text{IDL> contour, z, x, y, charsize=1.25, xtitle='x', ytitle='y', nlevels=16}
\]

where, as with \texttt{surface}, the second and third “free” arguments define the scales for the two axes. The keyword \texttt{nlevels} overrides the default drawing of contour lines at six values, here stipulating
16 values instead. Alternatively, we could use the keyword `levels` to stipulate the values at which contour lines are to be drawn, the keyword `c_linestyle` to stipulate the linestyle to be used for each contour line, and the keyword `c_labels` to stipulate which contour lines should be labeled with their values. The statement

```
IDL> contour, z, x, y, charsize=1.25, xtitle='x', ytitle='y',
    IDL> levels = [-0.8,-0.6,-0.4,-0.2,0.2,0.4,0.6,0.8],
    IDL> c_linestyle = [ 2, 2, 2, 2, 0, 0, 0, 0 ],
    IDL> c_labels = [ 0, 1, 0, 1, 1, 0, 1, 0 ],
    IDL> thick=3.0
```

produces the display shown in Fig. 2.15. Facilities available in IDL to make the aspect ratio in this figure 1:1 are described in Section 2.16.5. Details about other keywords that affect the command `contour` can be found in the IDL manuals.

### 2.12.4 Shaded Surfaces: The Procedure `shade_surf`

The procedure `surface` displays a two-dimensional array using a “wire-mesh” technique in which data points are connected to adjacent points using lines. IDL can also generate an alternate display in which a smooth but shaded surface conveys the values in a two-dimensional array. The procedure `shade_surf` uses the same syntax as `surface` and accepts essentially all of the keywords available to `surface`. To produce shaded surface representations of our two examples, we might execute the statements

```
IDL> device, decomposed=0
IDL> loadct, 3
IDL> shade_surf, z, x, y, charsize=1.5,
    IDL> xtitle='x', ytitle='y', ztitle='z',
```

47Behind the scenes, shaded surfaces require fairly large byte arrays. Because of limitations on the sizes of arrays, only quite small shaded surfaces can be produced with the student version of IDL.
Figure 2.15: Contour plot of the function $z$. 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.

IDL> title='[3,2] Mode'
IDL> shade_surf, I, xi, eta, charsize=1.5, 
IDL> xtitle='!4n!3', ytitle='!4g!3', ztitle='z', 
IDL> title='Diffraction!C!C(Square Aperture)'

The statement `loadct, 3` loads a color table (specifically, color table 3; the default is color table 0, which defines a grey scale) for the shaded surface display. Figures 2.16 and 2.17 show these plots, though the printed version using a grey scale pales by comparison with the color plot on the screen. Note that execution of the statement `loadct` will not affect any displays already on the screen and must precede the plotting statements that are to use the new color table.

2.12.5 The Procedure `xsurface`

Exploration of the nature of a surface in two dimensions is greatly facilitated by the widget-based procedure `xsurface`. Statements like

IDL> xsurface, z
IDL> xsurface, I

will bring up a new window that includes not only a display of the surface produced with `surface` but also several buttons and widgets that allow a variety of adjustments in the size, angle of view, and nature of the display. Further, and of particular usefulness, the window contains a box in which the statement that will produce the current display is presented. Thus, we can use this tool to identify the detailed statement to place in a developing pro-file so as to produce the most effective display of the surface.

---

48 The command `loadct` with no arguments will produce a list of the possible color tables and prompt for the entry of the desired number.
2.12.6 The Procedure show3

A final way to represent two-dimensional arrays uses the procedure show3. In a single three-dimensional box, show3 displays a wire mesh surface, a contour plot on the top of the box, and a shaded contour on the bottom of the box. The statements

IDL> show3, z
IDL> show3, I

will produce this composite display for each of our sample functions. The procedure show3 accepts only a few keywords, one of which stipulates keywords to be passed to surface and a second of which stipulates keywords to be passed to contour. The details of structuring more elaborate statements using show3 are fully described in the IDL manuals.

2.12.7 Functions of Two Variables in Polar Coordinates

IDL also makes available the function polar_surface and the procedure polar_contour for plotting functions of the radial and azimuthal polar variables rather than of Cartesian variables. The function polar_surface takes as input

- a two-dimensional array containing values of the surface at a grid of points spanning the region $0 \leq r \leq a$, $0 \leq \phi \leq 2\pi$,
- a vector containing the values of $r$ in that interval, and
- a vector containing the values of $\theta$ (in radians) in that interval

and returns as output a two-dimensional array giving values of the surface on a Cartesian grid of points spanning $-a \leq x, y \leq a$. This latter array can then be displayed using the routines already described for displaying surfaces expressed in Cartesian coordinates. For example, the statements

49We have here placed the arguments to lugen_grid in a different order than that used previously. Remember that IDL distinguishes between “free” arguments and arguments labeled with keywords. The free arguments must appear in the right order, since their position in that order communicates their significance. Keywords can be presented in any order, and they can be mixed in any way with “free” arguments. IDL will not be confused.
IDL> lugen_grid, r, phi, xrange=[0.0,5.0], nx=25, yrange=[0.0,360.0], ny=36
IDL> phi_rad = 2.0*!pi*phi/360.0
IDL> x = findgen(51)/5.0 - 5.0
IDL> z = (1.0 - r^2)*exp(-r)*cos(phi_rad)

create arrays \(r\) and \(\phi\) radial conveying a grid of points in the intervals \(0.0 \leq r \leq 5.0\) and \(0 \leq \phi \leq 2\pi\) (in radians), create a vector \(x\) containing the values of \(x\) and \(y\) that \textit{polar} surface will use,\(^{50}\) and then evaluate the function

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

over that grid. With these variables now ready, the single statement

IDL> surface, polar_surface( z, r[*,0], phi_rad[0,*], /grid ), $ x, x, charsize=1.5

exploits \textit{polar} surface to convert the values of \(z\) from a polar grid to a Cartesian grid and plot a mesh surface conveying that function. A small number of keywords to \textit{polar} surface allows some control over the relationship between the Cartesian and polar grids. When set to true, the keyword \textit{grid} informs the procedure that the radial and azimuthal grids are regular (points uniformly spaced); another keyword (quintic) allows selection of a slower but more accurate method of interpolating in the conversion. Still others are identified in the IDL manuals. The resulting graph is shown in Fig. 2.18.

The procedure \textit{polar} contour takes the same input as \textit{polar} surface (though it expects the “free” variables in a different order). As output, \textit{polar} contour produces a contour map directly from the polar representation of the surface; no conversion to a Cartesian representation is necessary. For example, given the arrays already evaluated in the previous paragraph, the single statement

IDL> polar_contour, z, phi_rad, r, nlevels=20, thick=3.0, /isotropic

produces the contour map shown in Fig. 2.19. Available keywords, which are fully described in the IDL manuals, include almost all those accepted by the procedure \textit{contour}. In particular, the keyword \textit{isotropic} forces equal scaling on both axes.

The purpose of this brief section is to draw attention to the existence of these two routines. Details of their use can be found in the IDL manuals.

### 2.13 Graphing Scalar Functions of Three Variables

Functions of three variables also occur regularly in physics. Displaying them graphically, however, is complicated because there are \textit{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, the \(ijk\) element of which 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. In \textit{Cartesian} coordinates \((x, y, z)\) with the nucleus located at the origin, this probability density is given by the expression

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

\(^{50}\)By default, the routine will divide the interval \(-5.0 \leq x, y \leq 5.0\) into 50 segments.
Figure 2.18: Mesh representation of the function $z$ in Eq. (2.15).

Figure 2.19: Contour map of the function $z$ in Eq. (2.15).

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

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

(2.17)

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 30 segments, which will entail evaluating $p(x, y, z)$ at 31 values of $x$, 31 values of $y$, and 31 values of $z$—a total of $31 \times 31 \times 31 = 29791$ values.

Using IDL to generate a three-dimensional array containing values of this probability density at a grid of points in the desired region is straightforward, since we built `lugen_grid` so that it would also provide the arrays for the independent variables in three dimensions. The array representing the probability density in Eq. (2.16) can be created using the statements

IDL> `lugen_grid`, x, xrange=[-10.0, 10.0], nx=30, $\$\$ Calculate grid.
IDL> y, yrange=[-10.0, 10.0], ny=30, $\$\$ Calculate premultiplier.
IDL> z, zrange=[-10.0, 10.0], nz=30
IDL> pfactor = 1.0/(2.0*!pi*27.0^3)
IDL> zs = z^2
IDL> rhos = x^2 + y^2 + zs
IDL> rho = sqrt(rhos)
IDL> p = pfactor * exp(-2.0*rho/3.0) * $\$\$ Calculate $p(x, y, z)$.
IDL> (9.0*zs^2 - 6.0*zs*rhos + rhos^2)

The first statement—which appears distributed over three physical lines—produces a grid of values for the evaluation of the scalar field.\(^51\) 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.

### 2.13.1 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

\(^51\) We here use the optional arguments to `lugen_grid` for the first time. These arguments specify ranges on the $z$ coordinate. Their presence triggers the creation of three-dimensional arrays containing the $x$, $y$, and $z$ coordinates of a grid of points in three dimensions.
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, $z = z_0 + 5 \ast dz$). Thus, the statements

```
IDL> surface, p[*,*,5], x[*,*,5], y[*,*,5]
IDL> contour, p[*,*,5], x[*,*,5], y[*,*,5]
IDL> shade_surf, p[*,*,5], x[*,*,5], y[*,*,5]
```

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

```
IDL> px = fltarr(31,31)
IDL> for i = 0 , 30 do $
IDL> \quad$ for j = 0 , 30 do $
IDL> \quad\quad$ px[i,j] = p[15,i,j]
```

for example, will extract data in a plane parallel to the $yz$ plane. Then, the statements

```
IDL> xx = findgen(31)/1.5 - 10.0
IDL> surface, px, xx, xx, charsize=1.5, $
IDL> \quad$ xtitle='x', ytitle='y'
IDL> contour, px, xx, xx, nlevels=20, $
IDL> \quad$ xtitle='x', ytitle='y', /isotropic
```

will prepare a vector for labeling the axes and generate both a mesh surface and a contour map representing the probability density in the chosen plane. The figures produced by these commands are presented in Fig. 2.20 and Fig. 2.21.

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 displaying only a very small portion of the data. IDL’s widget-based tool `slicer3` is more powerful, since it allows easy relocation of the plane that “slices” through the three-dimensional volume and also permits oblique slices. Data are supplied to `slicer3` not by passing a variable name but by specifying a pointer as an argument. When the data to be displayed already exist in IDL’s workspace, we must first create a pointer to the data with the statement

```
IDL> ptr_p = ptr_new( p )
```

Then, we invoke `slicer3` with the statement

```
IDL> slicer3, ptr_p
```

A new window containing three menus (FILE and TOOLS along its top; MODE in its control panel), a number of buttons, and a graph window will be created. As displayed initially, the tool is in ‘slice’ mode and is set to draw orthogonal sections perpendicular to the $x$ axis, though the orientation of the cut-plane can be changed by buttons labeled ‘X’, ‘Y’, and ‘Z’ in the control panel. After selecting the orientation of the cut plane, we place the cursor in the larger graph window, drag the plane to

---

52 We must extract two-dimensional arrays from all three arguments, since these statements will not behave properly with three-dimensional arrays in any position.

53 The statement `ptr_p = ptr_new( p, /no_copy )` would be less consumptive of memory, and may sometimes be necessary. Unfortunately, the more elaborate statement will render the original variable $p$ undefined.
Figure 2.20: Mesh representation of the function \( p \) in Eq. (2.16) in the plane \( x = x_0 + 15 * dx = -10.0 + 15 * 20.0/32.0 = -0.625 \).

Figure 2.21: Contour map of the function \( p \) in Eq. (2.16) in the plane \( x = x_0 + 15 * dx = -10.0 + 15 * 20.0/32.0 = -0.625 \).

the desired position with ML and, finally, release ML to trigger the drawing of the section. Further sections can be selected in the same way and will be either merged with ('draw' button) or displayed on top of ('expose' button) all previously drawn sections. If the ‘oblique’ button is selected, then ML can be used to drag the normal to the cut-plane into the desired direction and a click of ML on the ‘Display’ button will trigger the creation of the section. To erase the display and start over, we select ‘erase’ from the TOOLS menu. Many other features of slicer3 are described in the IDL manuals. Note particularly that the value of ptrp is destroyed on exit from slicer3 and must be recreated before re-entering slicer3.
2.13.2 Isosurfaces

Another method of viewing three-dimensional scalar fields is to look at isosurfaces, or sets of locations at which the function has a fixed—and specified—value. In our present example, we locate and display a surface of constant probability density with the statements

IDL> `isosurface`, `p`, `1.2e-4`, `vertices`, `polygons`  
Find isosurface.

IDL> `scale3`, `xrange`=[0.0,30.0], `yrange`=[0.0,30.0], `$  
Set transformation matrices.

IDL> `zrange`=[0.0,30.0]  
Set white background, which must occur prior to the invocation of `polyshade`.

IDL> `!p.background=255`  
Create image.

IDL> `img = polyshade( vertices, polygons, /t3d )`  
Display the image.

IDL> `tvscl`, `img`  

The first statement—`isosurface`—searches the array `p` to generate in `vertices` and `polygons` lists of the vertices and polygons that together define a three-dimensional isosurface at which \( p(x,y,z) \) has the value \( 1.2 \times 10^{-4} \). The second statement—`scale3`—converts the specified ranges of the data into the appropriate transformation matrices for projection of the three-dimensional image into the two-dimensional computer screen.\(^\text{55}\) The third statement—`polyshade`—uses the information in `vertices`, `polygons`, and—because the keyword `t3d` is set—the transformation matrices to construct an image of the isosurface, and the fourth statement—`tvscl`—displays that image. The result is shown in Fig. 2.22.

The widget-based procedure `slicer3` also has options that permit the drawing of isosurfaces. We bring up the widget again with the statements

IDL> `ptr_p = ptr_new( p )`  

IDL> `slicer3`, `ptr_p`

\(^{54}\)Other values could be used. We should, of course, be familiar with the range of the data so that the chosen value actually occurs within the data. IDL’s functions `max` and `min` can be used to determine those extreme values.

\(^{55}\)By setting more parameters in `scale3`, such as angles, we can display different views of the image.
2.13. GRAPHING SCALAR FUNCTIONS OF THREE VARIABLES

Then, we select ‘surface’ from the Mode menu, select the value at which the isosurface is to be drawn by dragging the vertical line in the small window (which shows a histogram of the values in the database) with any mouse button or entering a value explicitly in the text-entry window below that histogram and, finally, clicking ML on the ‘Display’ button.

An even more useful approach to examining the character of a three-dimensional isosurface exploits the IDL routine xobjview. After the quantities vertices and polygons have been created in the above dialogue, we replace calls to scale3, polyshade, and tvscl with an invocation of obj_new to create a suitable object from the data returned by isosurface followed by an invocation of xobjview to create a view of that object in a window detached from the IDL command window. Specifically, the statements

\[
\begin{align*}
\text{IDL}> & \text{wavefct} = \text{obj\_new('IDLgrPolygon', vertices, } \\
& \quad \text{polygons=polygons, color=[255,0,0]}) \\
\text{IDL}> & \text{xobjview, wavefct}
\end{align*}
\]

achieve the desired end. (Here, IDLgrPolygon is the reserved name for the specific class of object we need here to create.) Once the Xobjview window has been created and the object has been drawn in it, we can use the mouse and the window’s buttons and menus to rotate, resize, or translate the object, to zoom in or out, to restore the original view, and perform many other adjustments—all in real time. Keywords allowed in the arguments of obj_new are influenced by the particular class of object to be created, and there are many keywords for each class; here we have illustrated two (polygons, which is essential to the creation of the desired object, and color (default black), which uses the RGB designation to specify the color for drawing the object). Keywords allowed in the arguments of xobjview include scale (default 1/√3), title (default null), and [x,y]size (default 400 pixels), which specify the initial scaling, a title for the display, and the size of the display window. Numerous other keywords are identified and described in the IDL manuals.

2.13.3 Other IDL Routines

Additional IDL routines for examining functions of three variables include

- **extract_slice**, which passes a plane of user-specified position and orientation through the three-dimensional data set and constructs a two-dimensional array containing values from which the character of the data in that two-dimensional slice can be examined.

- **project_vol**, which converts a three-dimensional array of values, e.g., the arrays representing scalar functions of three variables, into a two-dimensional array by projecting the data in a way similar to that of an x-ray picture. The value at each point in the two-dimensional array conveys information about the nature of the three-dimensional data along a particular path from the source of the x-ray through the data to the viewing screen.

- **voxel_proj**, whose function is similar to that of project_vol, except that the variety of projections is smaller and execution time is therefore likely to be faster.

- **xvolume**, whose function is similar to that of xsurface except that it supports interactive manipulation of isosurfaces in three-dimensional scalar fields.

Details on all four of these routines are described in the IDL manuals.

---

56We are here simply identifying what is, in fact, an enormous topic: IDL’s capabilities for creating, processing, and displaying objects. Even though we shall in one or two places open a window into the topic, even a superficial treatment would go well beyond the scope of this book.
2.14 Graphing Vector Fields

Graphical display of vector functions of two or three variables is more complicated than display of scalar functions, but IDL has several routines 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. 2.23 when the electric field has only a $z$ component, the magnetic field has only $x$ and $y$ components, and none of these components depends on $z$. In this case, the only non-zero components of the electric field $\mathbf{E}$ and the magnetic field $\mathbf{H}$ are given at time $t = 0$ by

$$
E_z = \cos \kappa_x b \pi \sin n\pi \gamma \quad H_x = \sin \kappa_x b \pi \cos n\pi \gamma \quad H_y = -\frac{\kappa_x b}{n\pi} \cos \kappa_x b \pi \sin n\pi \gamma
$$

(2.18)

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

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

(2.19)

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. (2.19) by $\omega$, $b$, $c$, and $n$. 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, we find that the fields we seek to display are given by

$$
E_z = \cos 2\pi \pi \sin 2\pi \gamma \quad H_x = \sin 2\pi \pi \cos 2\pi \gamma \quad H_y = -\cos 2\pi \pi \sin 2\pi \gamma
$$

(2.20)

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

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 in a Dover reprint after January, 2003, but that edition is also now out of print.
Each component of these fields can now be represented by a two-dimensional array. The \( \mathbf{H} \) field, which has two non-zero components, then is translated into two such arrays; the \( \mathbf{E} \) field, which has only one non-zero component, requires only one such array. These arrays are readily created with the IDL statements\(^\text{58}\)

\[
\begin{align*}
\text{IDL} & > \text{lugen\_grid, } x, y, \text{xrange}=[0.0,1.0], \text{nx}=30, \text{yrange}=[0.0,1.0], \text{ny}=30 \\
\text{IDL} & > \text{xp} = 2.0*\text{pi}*x \text{ & yp} = 2.0*\text{pi}*y \\
\text{IDL} & > \text{Ez} = \text{cos(xp)} \ast \sin(yp) \\
\text{IDL} & > \text{Hx} = \text{sin(xp)} \ast \cos(yp) \\
\text{IDL} & > \text{Hy} = -\text{cos(xp)} \ast \sin(yp)
\end{align*}
\]

### 2.14.1 The Procedure velovect

The IDL procedure \textit{velovect} produces a two-dimensional vector field plot. At each grid point \textit{velovect} draws an arrow which conveys both the direction and the magnitude of the field at that point. The calling sequence for this procedure has the form\(^\text{59}\)

\[
\textit{velovect, u, v, x, y}
\]

where \( u \) and \( v \) are the \( x \) and \( y \) components of the two-dimensional field, respectively, and \( x \) and \( y \) (which are optional) are \textit{vectors} containing the actual coordinates of the grid points on the two axes. Thus, for example, the two components of \( \mathbf{H} \) in IDL’s memory can be displayed in this form by executing the statement

\[
\text{IDL} > \text{velovect, Hx, Hy, x[*,0], y[0,*], xtitle='x', ytitle='y'}
\]

(Here, \( x[*,0] \) and \( y[0,*] \) extract the first row and first column, respectively, from the two-dimensional arrays \( x \) and \( y \) to create the vectors required by \textit{velovect}.) The resulting display is shown in Fig. 2.24.

As with most IDL commands, \textit{velovect} can be embellished with many keywords. We have already used \textit{xtitle} and \textit{ytitle}. Among the others are

- \textit{length} (default 1.0), which specifies the length of the longest vector as a multiple of the length of a cell, i.e., the distance from one grid point to the next.

- \textit{color}, which specifies the color index used for the plot.

and all of the keywords accepted by \textit{plot}.

### 2.14.2 The Procedure vel

Invoked with a statement of the form

\[
\textit{vel, u, v}
\]

where \( u \) and \( v \) again contain the \( x \) and \( y \) components of the field,\(^\text{60}\) the procedure \textit{vel} draws a vector field with \textit{curved} arrows that follow the field and are proportional in length to the field strength. Each arrow is composed of a number of short segments and is positioned at a random point on the \( xy \) grid. Thus, the statement

\(^{58}\)From here on in this section, we drop the overbars.

\(^{59}\)This statement merely illustrates the syntax. Since \( u \) and \( v \) have not yet been defined, this specific statement is not itself executable.

\(^{60}\)The procedure \textit{vel} does not accept additional “free” arguments to specify actual coordinates nor does it accept the keywords \textit{xtitle} and \textit{ytitle}.
Figure 2.24: Magnetic field at $t = 0$ in a rectangular wave guide using `velovect`. The illustrated structure, of course, propagates along the guide as time unfolds.

```idl
IDL> vel, Hx, Hy, nvecs=500, $ IDL> title = 'Magnetic field in rectangular wave guide'
```

which uses the keyword `nvecs` to set the number of vectors drawn to 500 (default 200) and the keyword `title` to specify a title (default ‘Velocity Field’), produces the plot shown in Fig. 2.25. Other useful keywords include

- `length` (default 0.1), which specifies the length of each line segment used to draw the curved arrows. It is expressed as a fraction of the longest vector divided by the number of steps.
- `nsteps` (default 10), which specifies the number of line segments used to draw each arrow.

### 2.14.3 The Procedure `plot_field`

The procedure `plot_field` also plots a two-dimensional field with a statement of the form

```idl
plot_field, u, v
```

where `u` and `v` are defined above. This procedure places each arrow at randomly picked points on the plane. In addition to the keyword `title` (default ‘Velocity Field’) as described for the `vel` procedure, some of the keywords accepted by this command are

- `n` (default 200), which specifies the number of arrows to be drawn.
- `length` (default 0.1), which specifies the length of the longest vector as a fraction of the plotting area.

The display produced by `plot_field` is similar to that produced by `vel` and is not shown.
Figure 2.25: Magnetic field at $t = 0$ in a rectangular wave guide using \texttt{vel}. The illustrated structure, of course, propagates along the guide as time unfolds.

2.14.4 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 \texttt{velovect} are superimposed on a contour map of the $z$ component of the electric field. The statements

```idl
IDL> velovect, Hx, Hy, x[*,0], y[0,*], xtitle='x', ytitle='y'
IDL> contour, /overplot, Ez, x, y, thick=3.0, $
IDL> levels=[-0.9,-0.7,-0.5,-0.3,-0.1,0.1,0.3,0.5,0.7,0.9], $
IDL> c_linestyle=[2,2,2,2,2,0,0,0,0,0]
```

produce the display shown in Fig. 2.26. Here, we have used the keywords \texttt{levels} and \texttt{c_linestyle} to specify, first, that contours at particular values are to be drawn and, second, that contours for negative values should be drawn with a dashed line (linestyle 2) while contours for positive values should be drawn with a solid line (linestyle 0). 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 dashed line.

2.14.5 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 the procedures already discussed to draw pictures of the projections of the field in any given plane into that plane. In addition, IDL provides the procedure \texttt{flow3} (see the IDL manuals) for drawing a picture similar to those produced by \texttt{vel} and \texttt{plot_field}, but drawing three-dimensional space curves to represent each field line. The primary input to \texttt{flow3} is a trio of three-dimensional arrays giving the $x$, $y$, and $z$ components of the vector field at a grid of points spaced through the volume of interest. The exploration of these capabilities is left to an exercise.
Figure 2.26: Plot showing both magnetic and electric fields at \( t = 0 \) in wave guide. The magnetic field lines are closed curves in the plane of the paper. The electric field is directed perpendicular to the paper and has at each point a component conveyed by the contour lines, where solid lines indicate a positive component and dashed lines a negative component. The illustrated structure, of course, propagates along the guide as time unfolds.

2.15 Animation

Along with all of the image processing and graphing capabilities, IDL 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 a properly sized byte array, and then displays each of those bit-mapped images in turn.\(^{61}\) To illustrate this process, we use a data file containing a succession of 41 33 × 33 arrays, each of which conveys the displacement of a square drumhead at a particular moment in time.\(^{62}\) The statements creating a three-dimensional byte array containing the 41 images in the animation are\(^{63}\)

\[
\begin{align*}
& \text{IDL} > \text{openr, 1, 'ldrum.dat'} \quad \text{Read data file.} \\
& \text{IDL} > \text{disp = fltarr( 33, 33, 41 )} \\
& \text{IDL} > \text{readf, 1, disp} \\
& \text{IDL} > \text{close, 1}
\end{align*}
\]

\(^{61}\)Limitations on array sizes in the student version of IDL constrain the extent to which animated displays can be produced with that version.

\(^{62}\)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 drumhead.

\(^{63}\)We assume that the file \texttt{ldrum.dat} is stored somewhere in IDL’s search path. See footnote 23 if that is not the case.
2.15. ANIMATION

IDL> window, /free, title='Animation', $
IDL> xsize = 600, ysize = 600

IDL> frames = bytarr(600,600,41)
IDL> for i = 0, 40 do begin $
IDL> shade_surf, disp[*,*], i, zrange=[-0.25,0.25] & $
IDL> frames[0,0,i] = tvrd(0,0,600,600) & $
IDL> endfor

At this point, the array frames contains all of the images stored as bit maps so they can be displayed one after the other in quick succession. In this coding, note the following: (1) the byte array frames, in which bit-mapped representations of images are stored, requires 600 × 600 × 41 bytes ≈ 14.8 MB of memory and (2) the arguments of the command tvrd, which reads the contents of the window and stores it in the byte array, give the coordinates of the bottom left and the top right corners of the window (in pixels).

Several options for displaying the images can be invoked. The command tvsc1, for example, displays a single image, scaling the colors to use the full spectrum available on the display device. Thus, the statement

IDL> tvsc1, frames[*,*],0

will display the first frame and the statement

IDL> for i = 0, 40 do tvsc1, frames[*,*],i

will display all 41 frames in quick succession. In older versions of IDL, we would have now used the procedure movie with a statement like

IDL> movie, frames, order=0

which would have displayed each image in the array frames one after another from beginning to end, repeatedly until stopped. This route offered control over the rate at which frames are displayed, but the procedure movie has been deleted from more recent versions of IDL.

In more recent versions of IDL, we would use a still more elaborate widget-based IDL procedure called xinteranimate, which offers substantial interactive control over the animation by allowing it easily to be stepped forward or backward frame by frame. Three statements are needed to set it up. First, we must initialize the animation tool with a statement like

IDL> xinteranimate, set = [ 600, 600, 41 ], /showload

This statement creates a window titled ‘XInterAnimate’, which in turn includes a display window that is 600 × 600 pixels in size and prepares the tool for receiving 41 images of that size. Then, the statement

IDL> for i = 0, 40 do xinteranimate, frame = i, image = frames[*,*],i

will load the images into the tool, one at a time, and—because the keyword showload was set to true in the previous statement—will display each image as it is loaded. Finally, the statement

IDL> xinteranimate
will initiate the animation and make all of the controls active so we can speed up or slow down the animation, pause it, change the direction of its progress, and select specific frames for closer examination. Clicking ML on the 'End Animation' button will exit from the animation tool.\footnote{Do not close the window by clicking on the $\times$ in its upper right corner. That action may hang the system in a way from which it will be difficult to recover.}

The procedures described in the previous two paragraphs ultimately display a succession of images, each of which is separately defined from 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. IDL procedures can then be used to generate the remaining images from that initial image. For example, generating an animated display showing the simple harmonic motion of a drumhead is especially easy if we have data on its initial shape. One simply creates 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.\footnote{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.}

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

```idl
IDL> lugen_grid, x, y, xrange=[0.0,1.0], $ Create array with initial shape.
IDL> nx=25, yrange=[0.0,1.0], ny=25
IDL> u = sin(2*!pi*x) * sin(3*!pi*y) Create values of time.
IDL> t = findgen(41)/40.0 Create array for images.
IDL> disp = fltarr(26,26,41) Calculate data for for each image.
IDL> for i = 0, 40 do $ Calculate data for for each image.
IDL> disp[*,*,i] = u*cos(2*!pi*t(i))
```

At this point, the necessary data are stored in the 3D array disp. From here, the creation of the animated display follows exactly the pattern exhibited in the second and third paragraphs of this subsection. We simply have generated the necessary input data by calculating it from the initial shape rather than by reading it from a file.

### 2.16 Advanced Graphing Features

In this section, we describe some of IDL’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.

#### 2.16.1 Fonts

Among the most useful of IDL’s auxiliary features is a capability to choose characters from a wide variety of fonts (including sans serif, italic, and bold characters, Greek letters, and common mathematical symbols) and to position characters as subscripts and superscripts. Unfortunately, the manner in which these features are invoked is awkward and non-intuitive. The example in this subsection is intended to introduce some of IDL’s capabilities but is in no way a substitute for careful reading of IDL’s own documentation on the topic.

Fonts and character positions are selected by commands embedded in quoted strings. Each such command begins with an exclamation point ‘!’ and is completed with a code, some of which are
Table 2.4: Some codes for use in character strings in IDL. Codes in the first and third columns identify fonts. Codes in the fifth column effect changes in position and size of the characters drawn. The simplex Roman font (!3) and the normal level (!N) are the defaults.

<table>
<thead>
<tr>
<th>Code</th>
<th>Font</th>
<th>Code</th>
<th>Font</th>
<th>Code</th>
<th>Shift to</th>
</tr>
</thead>
<tbody>
<tr>
<td>!3</td>
<td>Simplex Roman</td>
<td>!12</td>
<td>Simplex Script</td>
<td>!C</td>
<td>New line</td>
</tr>
<tr>
<td>!4</td>
<td>Simplex Greek</td>
<td>!13</td>
<td>Complex Script</td>
<td>!D</td>
<td>Subscript 0.62 size</td>
</tr>
<tr>
<td>!5</td>
<td>Duplex Roman</td>
<td>!14</td>
<td>Gothic Italian</td>
<td>!E</td>
<td>Superscript 0.44 size</td>
</tr>
<tr>
<td>!6</td>
<td>Complex Roman</td>
<td>!15</td>
<td>Gothic German</td>
<td>!I</td>
<td>Subscript 0.44 size</td>
</tr>
<tr>
<td>!7</td>
<td>Complex Greek</td>
<td>!16</td>
<td>Cyrillic</td>
<td>!N</td>
<td>Normal level</td>
</tr>
<tr>
<td>!8</td>
<td>Complex Italic</td>
<td>!17</td>
<td>Triplex Roman</td>
<td>!U</td>
<td>Superscript 0.62 size</td>
</tr>
<tr>
<td>!9</td>
<td>Math</td>
<td>!18</td>
<td>Triplex Italic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>!10</td>
<td>Special</td>
<td>!20</td>
<td>Miscellaneous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>!11</td>
<td>Gothic English</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

identified in Table 2.4. For example, a Greek character set is selected with the command !4 while raising characters to the position of a superscript and shrinking their size is effected by one of the commands !E or !U. Returning to the default character set is accomplished with the command !3 and to the original character position by the command !N. Further details on these commands and information about which characters are in which positions in which character sets can be found in the IDL manuals. Further, the statement

IDL> showfont, 3, "Arbitrary Title of User’s Choice"

will create a display of all the characters in font 3. Of course, reserving the exclamation point to introduce commands of this sort makes conscious printing of that character difficult; to include an exclamation point explicitly in the printed output, type two exclamation points '!!' in the character string.

In Section 2.12.2, we have already seen one use of this capability. We present one more example. Let us embellish the graph of a magnetic field first presented in Fig. 2.5 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

IDL> dz = 8.0/100.0  
Set increment between points.
IDL> z = -4.0 + dz * findgen(101)  
Create vector of 101 values in interval.
IDL> B = (1.0 + z^2)^(−1.5)  
Evaluate B, also a vector of 101 values.
IDL> plot, z, B, title='Magnetic Field on !8z!6 Axis',  
Plot graph.
IDL> xtitle='!6Dimensionless Position, !8z!3', $ 
Set xtitle.
IDL> ytitle='!8B!6(!8z!6)=1/(1+!8z!6!E2!N)!E3/2!N!3 ', $ 
Set ytitle.
IDL> thick=3.0, charsize=1.5, ticklen=1.0 
Set other plotting parameters.

will produce the revised graph of the magnetic field shown in Fig. 2.27. Here, the commands !6 and !8 in a character string select the complex roman and complex italic fonts, respectively. In ytitle, the command !E was used to raise the character to the exponent level and automatically reduce the character size and the command !N was used to return to the normal level and original character size. As always, restoring the font to the default font with !3 is prudent, since changes in font made in one statement will sometimes “bleed” into subsequent statements.

2.16.2 Space Curves

Sometimes it is desirable to view the trajectory of a particle in three-dimensional space. IDL has the ability to take vectors of x, y, and z coordinates, project these points onto the two-dimensional
screen, and then connect consecutive points with lines. To illustrate this feature of IDL, let α be a constant and consider the curve defined parametrically with the equations

\[ x = \cos t \quad ; \quad y = \sin t \quad ; \quad z = \alpha t \quad (2.21) \]

These equations describe 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 \(z\). The next step is to use the procedure `scale3` to calculate and store the transformation matrix for projecting these three-dimensional coordinates onto the two-dimensional screen. Procedure `scale3` has three mandatory “free” arguments giving the \(x\), \(y\), and \(z\) coordinates of points on the curve and keywords giving the appropriate \(x\), \(y\), and \(z\) ranges. Finally, the procedure `plots` (not to be confused with the plain vanilla procedure `plot`) is used to plot the graph. The statements

IDL> erase
Clear plot window (if current window shows a display).\(^{66}\)

IDL> t = 30.0*findgen(201)/200.0
Set time values.

IDL> alpha = 0.5
Set parameter (\(z\) velocity).

IDL> x = cos(t)
Evaluate \(x\).

IDL> y = sin(t)
Evaluate \(y\).

IDL> z = alpha*t
Evaluate \(z\).

IDL> print, min(x), max(x)
Find ranges of each variable.

-0.999965 1.00000

IDL> print, min(y), max(y)

-0.999988 0.999994

IDL> print, min(z), max(z)

0.000000 15.00000

IDL> scale3, xrange=[-1.0,1.0], yrange=[-1.0,1.0], zrange=[0.0,15.0]
Invoke `scale3` with rounded ranges to generate transformation matrix.

IDL> plots, x, y, z, /3d
Plot graph, using transformation matrix.

will produce the graph in Fig. 2.28. Note in particular that `scale3`, which also allows the keywords

\(^{66}\)The command `plots` to be used presently does not automatically erase the current display. Without this statement, the graph here being set up will ultimately overlay any existing display.
ax and az to override the default rotations about the x and z axes, creates the transformation matrix which is then invoked by the command plots with the inclusion of the keyword /t3d. Note also that a new invocation of plots to produce space curves will not automatically clear the plot window; to erase the plot window prior to calling plots a second time, use the erase command.

An even more useful approach to examining the character of a three-dimensional space curve exploits the IDL routine xplot3d. After, the three coordinates x, y, and z have been evaluated, we simply execute the statement

IDL> xplot3d, x, y, z

and wait for the display to be produced. Once the XPlot3D window has been created and the space curve has been drawn in it, we can use the mouse and the window’s buttons and menus to rotate, resize, or translate the image, to zoom in or out, to change the style of the axes, to restore the original view, and perform many other adjustments—all in real time. Numerous keywords, including thick, [x,y,z]range, title, [x,y,z]title, and linestyle can be included to modify the details of the display.

2.16.3 System Variables

Many of the features of IDL are controlled by the values of system variables, the most important of which for graphical operations are !p, !x, !y, and !z. These variables are structured variables with many components, each of which has its own (sub)structure and each of which has a default value. We have already seen how the system variable !p.multi, which is a component of !p, allows for possible placement of more than one graphic image in a single window. A full listing of the components of these system variables will be found in the IDL manuals. Frequently useful components of these system variables include

- !p.background (default 0), a scalar that sets the background color for the graph window. With the default color table, the values 0 and 255 give black and white backgrounds, respectively.
- !p.charsize (default 1.0), a scalar that sets the size of the characters drawn to annotate a graph.
• !p.color (default 255), a scalar that sets the color for the lines drawn in the graph window. With the default color table, the values of 0 and 255 give black and white lines, respectively. Note that black lines will be invisible on the screen unless the background color is changed away from its black default.

• !p.linestyle (default 0), a scalar that specifies which of several line styles should be used for the drawing of the lines of a graph. Possible values and their effects are listed in Table 2.2.

• !p.thick (default 1), a scalar that controls the width of the lines used to draw graphs (but has no effect on the width of lines used for axes).

• !p.psym (default 0), a scalar that—when different from zero—identifies the plotting symbol to be used by plot, oplot, and plots.

• !p.ticklen (default 0.02), a scalar that specifies the length of the (major) tick marks, expressed as a fraction of the plot size.

• !x.range, !y.range, and !z.range (default for each [0.0,0.0], which results in autoscaling), two-component vectors that specify the range of values on the $x$, $y$, and $z$ axes, respectively.

• !x.style, !y.style, and !z.style (default for each 0), scalars specifying the style of the axes. Full details, which are complicated, are described in the IDL manuals.

• !prompt (default IDL>), a scalar that identifies the prompt displayed when IDL is waiting for a statement.

Explicit change of the value of a system variable makes the new value “permanent” (i.e., until we change it again—or exit from IDL). Similarly named keywords in many of the graphics commands permit overriding the default values set by the system variables, but only within the statement in which the keyword is specified. Statements use the values of the system variable as they exist when the statement is invoked unless the statement itself contains an explicit stipulation of a keyword overriding the current value of the system variable. The IDL manuals also contain a listing of these keywords and indicate correspondences between keywords and system variables.

2.16.4 Using Multiple Windows

In the simple circumstances so far introduced, we have accepted IDL’s willingness to create a window (IDL 0) when a plot command is invoked or to use an existing window 0 if it had already been created by a previous plot command. We can manipulate additional windows with statements like

```
window, n
window, n, title='New Window'
window, n, xsize=400, ysize=400
window, n, xpos=400, ypos=200
```

Create window $n$, where $n = 1, 2, \ldots$.
Create window $n$ with title.
Create window $n$ with size 400 $\times$ 400 pixels.
Create window $n$, overriding default positioning to place lower left corner 400 pixels to the right of and 200 pixels above the lower left corner of the screen.

```
window, /free
wdelete, n
wset, n
wshow, n
```

Create window with an available free index.
Delete window $n$.
Select window $n$ as the default.
Bring window $n$ to foreground.

Note that execution of the statement `window, 2` (say) when window 2 already exists will delete the existing window before creating the new one. Note also that the system variable !d.window, whose value is $-1$ when no plot window is on the screen, always contains the index of the current plot window; the commands `window` and `wset` manipulate this system variable, which we can print if we forget its value but we should not change it except indirectly through using those commands.
that manipulate it behind the scenes. Plot commands always direct their output to the current plot window. Thus, the statements

```idl
IDL> x = findgen(101)/10.0 & sine = sin(x) & cosine = cos(x)
IDL> window, 0, title = 'Sine Curve'
IDL> window, 1, title = 'Cosine Curve'
IDL> wset, 0 & plot, x, sine
IDL> wset, 1 & 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.

### 2.16.5 Scaling so Circles are Circles

If we accept all of the system defaults, the graph produced by the statements

```idl
IDL> s = findgen(101)/10.0
IDL> x = cos(s) & y = sin(s)
IDL> plot, x, y
```

will not appear at all as the circle it should be. More recent versions of IDL, however, make available the keyword isotropic, which is legal with the commands plot and contour (and perhaps others). When invoked, this keyword forces equal coordinate increments on the two axes to be displayed as equal distances on the axes on the screen. Thus, for example, the statements

```idl
IDL> s = findgen(101)/10.0
IDL> x = cos(s) & y = sin(s)
IDL> plot, x, y, /isotropic
```

will produce a graph in which the circle appears as a circle.

Unfortunately, while a graph produced in this way and written to a PostScript file will preserve the desired aspect ratio, the graph itself will not be centered in its bounding box and hence, for example, in a center environment in \LaTeX{} will not directly produce a centered graph on the page. One way to address this complication is to wrest control of some further defaults from the program so as to adjust the aspect ratio and control the portion of the window in which the plot appears. We illustrate the necessary adjustments first for on-screen displays. The keywords xsize and ysize to the window command and the system variables \texttt{!x.margin} and \texttt{!y.margin} are the important features of IDL in this context. Even then, however, we must be careful, since the size parameters for the window command specify the size of the entire window, including the title bar. We must begin by making sure that the region of the window available for plotting is square, which means that the vertical extent of the window must in fact be (about\footnote{The value 30 was determined empirically by measuring on the screen.}) 30 pixels larger than the horizontal extent. A statement like

```idl
IDL> window, 0, xsize=400, ysize=430
```

will achieve the desired end. Then, however, we must also override the default sizes that the plot command assumes for the margins around the axes, setting them instead so that the axes will embrace a square region. Those margins are controlled by the system variables \texttt{!x.margin} and \texttt{!y.margin}, each of which is a two-element vector. We might, for example set them to the values
IDL> !x.margin=[ 5.0, 2.0 ]
IDL> !y.margin=[ 3.0, 4.0 ]

(The sum of the two values in the vector assigned to !x.margin must be the same as the sum of
the two values in the vector assigned to !y.margin.) With these stipulations, the statements
with which this subsection began will now produce a graph that actually looks like a circle on the screen.
Similarly, the statements

IDL> window, 0, xsize=400, ysize=430
IDL> !x.margin=[ 5.0, 2.0 ]
IDL> !y.margin=[ 6.0, 1.0 ]
IDL> dt = 2.0*!pi/100.0
IDL> theta = dt*findgen(101)
IDL> a = 3.0
IDL> r = a*(1.0 - cos(theta))
IDL> plot, r, theta, /polar, 
IDL> subtitle='The Cardioid', 
IDL> thick=3.0, charsize=1.25
IDL> axis, 0, 0, xaxis=0
IDL> axis, 0, 0, yaxis=0

will produce a graph of the cardioid in which the two axes have identical scale increments on the
two axes so that, for example, one scale unit along the horizontal axis represents the same physical
length as one scale unit along the vertical axis.

Communicating this circle to the printer through a PostScript file, however, requires additional
specifications. Again, one uses xsize and ysize keywords, this time to the device command, so
as to set the size of the plotting area on the paper away from its default size. One also needs the
flag /inches to establish the units in which dimensions are specified. Further (though not here
illustrated), one may need the keywords xoffset and yoffset to make sure that the entire graph
actually appears on the printed page. The statements

IDL> set_plot,'ps'
IDL> device, filename='test.ps', 
IDL> xsize=4.0, ysize=4.0*(float(430)/float(400)), /inches
IDL> plot, x, y
IDL> device,/close
IDL> set_plot, 'x'

will set the plotting area to a size that has the same aspect ratio as the display produced on the
screen, thus assuring that the printed plot will also appear as a circle. Compare Fig. 2.29, which
was produced by the procedures here described, with Fig. 2.7, in the production of which we ignored
these issues of scaling.

2.16.6 Working with Color

By default, IDL produces its displays in a grey-scale, using color table 0, which is officially named
‘B-W linear’. The command loadct allows that default to be changed to any of three dozen or more
built-in alternative color tables, each of which makes 255 distinct colors available. For example, the
statements

device, decomposed=0
loadct, 3
plot, x, y, color=145
will load color table 3 as the current color table and then produce a graph of $y$ versus $x$ using color 145 from that table.\textsuperscript{68} Color table 3 will remain the current table until it is explicitly changed. A list of the available built-in color tables will be displayed if the command \texttt{loadct} is issued without an argument. Alternatively, the command \texttt{xloadct} will bring up a browser in which we can select any of the available color tables and see the resulting spectrum before making the selection final.

Even more generally, IDL allows us to construct our own correspondence between colors and color numbers. Each color in the spectrum is specified most simply by giving three values that determine the relative proportion of the primary colors red, green, and blue.\textsuperscript{69} For example, red would be represented by the RGB triplet (255,0,0), green by the triplet (0,255,0), and blue by the triplet (0,0,255). The table

<table>
<thead>
<tr>
<th>Color Name:</th>
<th>Black</th>
<th>Blue</th>
<th>Green</th>
<th>Cyan</th>
<th>Red</th>
<th>Magenta</th>
<th>Yellow</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color Index:</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Red value</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>255</td>
<td>255</td>
<td>255</td>
<td>255</td>
</tr>
<tr>
<td>Green value</td>
<td>0</td>
<td>0</td>
<td>255</td>
<td>255</td>
<td>0</td>
<td>0</td>
<td>255</td>
<td>255</td>
</tr>
<tr>
<td>Blue value</td>
<td>0</td>
<td>255</td>
<td>0</td>
<td>255</td>
<td>0</td>
<td>255</td>
<td>0</td>
<td>255</td>
</tr>
</tbody>
</table>

specifies eight colors. As described in the section on color tables in the IDL manuals, these eight colors can be loaded into the first eight elements of the IDL default color table with the statements

\begin{verbatim}
IDL> red = [ 0, 0, 0, 0, 255, 255, 255, 255 ]
IDL> green = [ 0, 0, 255, 255, 0, 0, 255, 255 ]
IDL> blue = [ 0, 255, 0, 255, 0, 255, 0, 255 ]
IDL> tvlct, red, green, blue
\end{verbatim}

\textsuperscript{68}Setting the keyword \texttt{decomposed} to zero (false) is essential to guarantee that the \texttt{color} keyword in the \texttt{plot} command will refer to the selected color table.

\textsuperscript{69}RGB designations are the most straightforward. IDL also allows specification by HSV (hue-saturation-value) and HLS (Ostwald system), which are described in the IDL manuals.
where the command tvlct loads the specified color table into the proper internal storage. Then, in subsequent plot commands, the color red, for example, would be identified as color 4. Once these color values have been loaded into the current color table, the statements\(^{70}\)

```
IDL> x = findgen(101)/10.0
IDL> y = sin(x)
IDL> plot, x, y, yrange=[-2.0,10.0], /nodata ; Draw axes
IDL> for i = 0, 5 do oplot, x, y+i, thick=3.0, color=i+1
```

will draw a succession of six sine curves, one above the other and each with a different color.

If we want to write a color PostScript file so that the colors seen on the screen are reflected in the printed output (on a color printer), we must modify the process so far described for creating a PostScript file to invoke the keyword `color` to the `device` command (so that IDL’s current color table is incorporated into the file and, hence, downloaded to the printer). The statements

```
set_plot, 'ps'
device, filename='test.ps', /color
... execute plot commands
device, /close
set_plot, 'x'
```

will write the requisite file.

*Once we are finished with our own private color table, we can use the statement `loadct,0` to return to the original default color table.*

### 2.17 Miscellaneous Occasionally Useful Tidbits

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

#### 2.17.1 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 in an IDL path will be properly interpreted even in Windows.

#### 2.17.2 IDL’s Default Directory

If IDL’s CLI is launched from a shell window of the operating system, the default directory (current working directory) will be the directory from which IDL was launched. In the GUI, the default directory is displayed in the box labeled ‘Current Directory’ just above the panel in which commands are entered. In either interface, the default directory can be displayed with the pair of statements

```
IDL> cd, current=direct & print, direct
```

\(^{70}\)In the initial plot command, we use the default color rather than color 7 for white so that, when we later produce hard copy (and the black background on the screen is turned to white), the axes (in white on the screen) will be turned to black in the hard copy. If we used color 7 (white), it would stay white in the hard copy and the axes would be invisible against the (now) white background.
In either interface, the default directory can be changed with statements like

```
IDL> cd, 'FullPath'
```

Provide full path of new directory

```
IDL> cd, 'DirectoryName'
```

Shift focus to subdirectory of current directory

```
IDL> cd, '../DirectoryName'
```

.. backs up one directory

Alternatively, but only in the GUI, the default directory can be changed by editing the text in the box labeled ‘Current Directory’, which action will construct and execute an appropriate command that will be automatically displayed at the IDL> prompt.

When the default directory is unknown or ill-defined, including a full specification of a path in a quoted file name is necessary.

### 2.17.3 Customizing IDL

The configuration of IDL can be customized in two ways, specifically by

1. A startup command file, the path to which is provided in the environment variable `IDL_STARTUP` for the CLI and in the text entry box labeled ‘Startup file:’ in the Window/Preferences/IDL tab in the GUI and/or

2. A (separate) command file, which can be invoked at any time by a specific command submitted at the IDL prompt.

While individual users can invoke the first method, that route is more often reserved to the system manager, who can supply a standard startup file and thereby create a standard starting IDL environment for all users. The Local Guide will explain whether the end user needs to establish any environment variables for the CLI and/or enter a specific path in the Window/Preferences/IDL tab for the GUI). Most often, the proper parameters will be established without any action on the part of the end user.

The second method of customization is recommended for individual end users, who can use it by creating a supplementary startup command file whose execution after IDL is started will provide additional customization (and perhaps override some of the features established by the system-supplied startup file). Then, in the CLI, the statement

```
idl startup
```

to the operating system when IDL is started or, in either the CLI or the GUI, the IDL statement

```
@startup
```

issued as the first statement to IDL after IDL starts will execute the file `startup.pro` in the default directory. Whichever command is used, the system-wide startup file is executed first; then, the individual file will be executed. (The name of the individual startup file is, of course, arbitrary.)

### 2.17.4 IDL’s Search Path

Whenever an explicit path is not specified for a file that IDL must read (either data file or pro-file), the program looks first in the current default directory (see Section 2.17.2) and then in a sequence of directories defined by the search path, which is stored in the system variable `!path` and can at any time be displayed with the statement `print, !path`. The initial search path is defined by default or, when IDL is launched, by an optional statement in the file pointed to as described in item 1.
in Section 2.17.3. The specific initial search path may be site-specific and is identified in the Local Guide. A statement like

```idl
!path = '/usr/people/abcdef:' + !path
```

will add the directory `/usr/people/abcdef` at the beginning of the current search path.

### 2.17.5 Restoring IDL’s Initial State

Sometimes, we would like to restore IDL to its initial state. The simplest way to achieve this objective is to exit from IDL and then restart it. IDL also provides the executive command `.reset_session`, which will remove all defined variables and functions (including common blocks) but will not reset system variables to their defaults. If you anticipate the need to restore those variables without exiting from IDL, you can save them immediately upon starting IDL with a statement like

```idl
save, /system_variables, filename='defaultvars.dat'
```

(Choose any convenient file name.) Then, at any time later in the session (or in another session), you can restore the default values with the statement

```idl
restore, 'defaultvars.dat'
```

(The file name in the `restore` statement must, of course, match the one you used in the `save` statement, and a full path must be included in the filename if the current default directory does not contain the referenced file.)

### 2.17.6 Common Storage in IDL

Common storage was introduced in Section 2.8. There is, however, no way short of executing the executive command `.reset_session` or exiting from and then restarting IDL to redefine a common area of a given name once it has been created at IDL’s main command level. Since these routes both entail removing much more than the common area from IDL’s workspace, neither may provide in all circumstances an acceptable way to remove a common area. Note, however, that a common area defined at IDL’s main command level (i.e., at the IDL prompt `IDL>`) can contain more variables than are used in various procedures or functions subsequently called in the session. Thus, for example, if we know that we will need at most six variables in a common area of a particular name, we can define that area at the main command level with six variables. Subsequent routines referring to that named common can then use no more than six variables but can, in fact, use fewer, should fewer be sufficient for the purposes of that routine.

### 2.17.7 Reading Data from the Keyboard

Occasionally, and perhaps especially in various pro-files, we may wish to prompt for one or more values by displaying a message on the screen and then read those values from characters typed at the keyboard. The keyboard is identified to IDL as input channel 0, so that, when executed, the statement

```idl
readf, 0, x
```

Be aware that full paths to different directories are separated by a character that depends on the operating system (`:` in UNIX and `;` in Windows; see also the Local Guide). We use `:` in this example.
will prompt with a colon ‘:’ and wait for the value of \( x \) to be entered from the keyboard; entry is terminated by pressing \( \langle \text{RETURN} \rangle \). Alternatively, the statement

```plaintext
read, x
```

will accomplish the same end without the need to specify an input channel. Should a prompting message be desired, we need simply provide it in quotation marks (single or double) as a first argument, as for example in the statement

```plaintext
read, 'Please enter a value for x: ', x
```

or, alternatively, by using the keyword `prompt`, as in

```plaintext
read, prompt = 'Please enter a value for x: ', x
```

If two or more values are to be entered, then the statement requesting the entry would be

```plaintext
read, 'Enter x and y coordinates: ', x, y
```

and, separated by space(s) or a comma, the values will be entered in a single line.

### 2.17.8 NaN and Inf

Division by zero will produce a warning but also assign to the variable being evaluated the value `Inf`. Some other arithmetic infelicities will result in the assigning of the value `NaN`—not a number—to the variable being evaluated. Rather than terminating execution of a loop when one or the other of these events occurs, IDL simply assigns the appropriate value to the current variable and goes on to the next value of the loop index. Other IDL routines know how to ignore these special values. While ultimately we should develop procedures that do not encounter these errors, in quick operations and in the early development of a procedure, we will not be forced by IDL to straighten out these problems before meaningful results can be obtained in subsequent operations.

## 2.18 References

Those seeking additional information about IDL should

- Consult the on-line help messages that can be brought up by typing ‘?’ at the IDL prompt or by accessing the Help menu in the menubar of the GUI interface to IDL.
- Google 'IDL manuals' and examine some of the links that emerge, including in particular
  - the link to IDL PDF guides,
    ```plaintext
    www.harrisgeospatial.com/docs/PDF-GUIDES.html
    ```
    which provides access, among others, to a downloadable document titled *Using IDL*,
  - the link to the Exelis Documentation Center
    ```plaintext
    www.harrisgeospatial.com/docs/home.html
    ```

and
any of several links to documents produced by third-party entities.

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

- *Learning IDL*, which provides a pedagogically oriented introduction to a large fraction of IDL’s capabilities and is an appropriate source for the first time user for whom the full reference manual may well seem daunting.
- *Using IDL*, which focuses on the interactive use of IDL and briefly describes the interface and many of the more commonly used commands for plotting functions, displaying images, processing signals, and performing various mathematical operations.
- *Building IDL Applications*, which focuses on the writing of programs in the IDL language.
- *The IDL Reference Guide*, which is a three-volume set detailing all capabilities of IDL and, in particular, containing a systematic, alphabetically arranged directory of all commands, option variables, keywords, and . . .
- *The IDL HandiGuide*, which—in ten or a dozen pages—summarizes all IDL commands, environment variables, system variables, graphics keywords, and other useful information.

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.

### 2.19 Exercises

#### 2.19.1 Writing IDL Statements

2.1. Write and test IDL 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.

2.2. Look up the procedure `sort` both in the on-line help and in the printed IDL 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.

2.3. Look up the procedures `lusol` and `ludc` both in the on-line help and in the printed IDL manuals, then use the process described therein to solve the linear equations solved by other means in Section 2.3.4.

2.4. Describe and test a sequence of IDL statements that uses a `for/do` 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.

2.5. (a) Describe and test a sequence of IDL statements that uses a `for/do` loop to evaluate $\sum_i a_i$ 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 IDL statements that uses the built-in function `total` to achieve the same end.

2.6. Write and test an IDL function that accepts two three-component vectors as input and returns a three-component vector containing the cross product of the two input vectors.
2.7. 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.

2.19.2 Finding Eigenvalues and Eigenvectors

2.8. Patterning your code after that in lueigen, write and test a function that accepts a symmetric matrix as input and returns a matrix, each column of which is structured with its first element one of the eigenvalues of the input matrix and its remaining elements the components of the eigenvector belonging to that eigenvalue.

2.9. 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 diagonals just above and just below the main diagonal 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.\(^{72}\)

(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, i.e., a graph of vec[i] versus i. 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

IDL> plot, [ [0.0], [ evecs[0,*] ], [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

IDL> x = findgen(12)/11.0
IDL> plot, x, [ [0.0], [ evecs[0,*] ], [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 $50 \times 50$.

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

2.10. 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 energy of the states with principal quantum number

\(^{72}\)Using the statement 'print, matrix' to print a $10 \times 10$ floating matrix matrix will result in a printout that is difficult to read because each row of the matrix will occupy more than one line on the screen. If the matrix happens to contain only integers, you might use instead the statement 'print, fix(matrix)'. With this statement, each row of the matrix will take only one line on the screen. The statement 'print, fix(100*matrix)' will show more decimal places in a compact display, though you will need mentally to adjust all printed values by the extraneous factor of 100.
CHAPTER 2. INTRODUCTION TO IDL

2.19.3 Graphing Scalar Functions of a Single Variable

2.11. Create and test a sequence of IDL statements that will produce a graph that is similar to Fig. 2.6 except that the axes are drawn along the lines $y = 0$ and $x = 0$. Note the strategy used in Section 2.10.3.

2.12. (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.

2.13. 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 $V(z)/(E_0a)$ versus $z/a$. Hint: Remember the IDL function $\text{abs}$.

2.14. 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.
2.15. 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 $\delta$ over the interval $-3\pi \leq \delta \leq 3\pi$ for various values of $r$, and write a paragraph describing these graphs.

2.16. The thin lens equation

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}$$

relates object distance $s$, image distance $s'$, and focal length $f$. Here, $s > 0$ for an object to the left of the lens, $s' > 0$ for an image to the right of the lens, and $f > 0$ for a converging lens. With $s$ ranging from large positive to large negative values, obtain graphs of $s'$ versus $s$ for various values of $f$ (both positive and negative), and write a paragraph describing these graphs.

2.17. According to the special theory of relativity, the mass $m$, the momentum $p$, and the kinetic energy $K$ of a particle moving with speed $v$ are given in terms of the rest mass $m_0$ and the speed of light $c$ by the equations

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

where $\beta = \frac{v}{c}$. Obtain graphs of $m/m_0$, $p/m_0c$, and $K/m_0c^2$ versus $\beta$, superimposing on each a graph of the corresponding non-relativistic expression, and write a paragraph describing these graphs.

2.18. 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

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

(2.22)

where $\kappa = \frac{2\pi n}{\lambda}$ 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.

2.19. 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

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

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. **Hint:** Remember the IDL function $\text{abs}$.

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

2.21. Journal articles often contain graphs showing data taken by several different scientists regarding the same relationship, with each experimentalist’s contribution marked with a different symbol. Describe an IDL procedure to produce such a graph and, if you have suitable data available (or if you can invent such data), demonstrate that your procedure works.
2.22. When a photon of initial energy $E_0$ undergoes Compton scattering from 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.

2.23. 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

$$E(\theta) = \frac{q}{4\pi \epsilon_0 r^2} \left( \frac{1 - \beta^2}{1 + \beta^2 \sin^2 \theta} \right)^{3/2}$$

where $\theta$ is the polar angle of the observation point, $r$ is the radial coordinate at that point, and $\beta = v/c$ ($c$ is the speed of light). Obtain graphs of $E(\theta)/(q/4\pi \epsilon_0 r^2)$ versus $\theta$ and, to reveal some of the behavior more visibly, of $E(\theta)/E(0)$ versus $\theta$ for various values of $\beta$ on the interval $-\pi \leq \theta \leq \pi$, and write a paragraph describing these graphs.

2.24. 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

$$I(\delta) = \cos^3 \delta (1 + \cos \delta)$$

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.

2.19.4 Graphing Scalar Functions of Two Variables

2.25. The Planck radiation law gives the expression

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

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 $\overline{\lambda} = \lambda/\lambda_0$. 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 $\overline{T} = T/T_0$. (You might find it informative to evaluate $T_0$ for $\lambda_0 = 550$ nm.) With these changes, the expression to be plotted can be recast in the form

$$u(\overline{\lambda},\overline{T}) = \frac{8\pi ch}{\lambda_0^5} \left( \frac{1}{e^{\frac{ch}{\lambda_0 kT}} - 1} \right)$$

and the question now becomes one of plotting this quantity using the dimensionless variables $\overline{\lambda}$ and $\overline{T}$.

2.26. 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.
2.27. Consider two circular current loops, each of radius \( a \) and lying with its center on and its plane perpendicular to the z axis. The first loop is centered at the point \((0, 0, b)\) and the second loop is centered at the point \((0, 0, -b)\). The axial component of the magnetic field at the point \((0, 0, z)\) is given by the equation

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

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.

2.28. 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.

2.29. Refine the technique leading to Fig. 2.15 showing the displacement of a square drumhead so that, both on-screen and in hard copy, the display will actually be square. The approach described in Section 2.16.5 will be useful. Hint: Remember the keyword isotropic.

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

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

where \( \kappa \) is the thermal conductivity of the potato, \( c \) is its heat capacity per unit volume, \( j_0(x) \) and \( j_1(x) \) are the zeroth- and first-order spherical Bessel functions, and \( \beta_n \) is the \( n \)-th root of \( j_0(x) \), i.e., \( j_0(\beta_n) = 0 \). Obtain graphs of \( u(r, t)/u_0 \) as a function of \( r/a \) for various values of \( t \). Obtain also a graph of the temperature \( u(0, t) \) at the center of the potato as a function of \( t \) and determine how long it takes the potato to bake if, by being baked, one means that the temperature at the center has risen to 175 °C, i.e., to a value such that \( u(0, t)/u_0 = 0.875 \). Hints: (1) Note that

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

Thus, the \( n \)-th root of \( j_0(x) \) is \( \beta_n = n \pi \). (2) Express times in units of \( ca^2/\kappa \) but then, taking the radius of the potato to be \( a = 0.05 \) m and taking \( \kappa \) and \( c \) for the potato to be those of water \( [\kappa = 0.63 \text{ J/(m s K)}, \ c = 4.2 \times 10^6 \text{ J/(K m}^3\text{)}] \), 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.

2.31. 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 IDL, 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 \( GmM/R \).
2.19.5 Graphing Scalar Functions of Three Variables

2.32. Following the pattern illustrated in Section 2.13, explore at least one of the three-dimensional scalar fields

\[ p_{3,1,0}(x, y, z) = \frac{8}{(27)^{3/2}} \rho^2 \left( 1 - \frac{\rho}{6} \right)^2 e^{-2\rho/3} \cos^2 \theta \]

\[ p_{3,1,1}(x, y, z) = \frac{4}{(27)^{3/2}} \rho^2 \left( 1 - \frac{\rho}{6} \right)^2 e^{-2\rho/3} (1 - \cos^2 \theta) \]

\[ p_{3,2,1}(x, y, z) = \frac{3}{(27)^{3/2}} \rho^4 e^{-2\rho/3} \cos^2 \theta (1 - \cos^2 \theta) \]

\[ p_{3,2,2}(x, y, z) = \frac{3}{4(27)^{3/2}} \rho^4 e^{-2\rho/3} (1 - \cos^2 \theta)^2 \]

giving the probability density for the hydrogen states \((n, l, m) = (3, 1, 0), (n, l, m) = (3, 1, 1), (n, l, m) = (3, 2, 1), (n, l, m) = (3, 2, 2)\). 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.

2.33. 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

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

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

2.19.6 Graphing Vector Fields

2.34. 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

```plaintext
xold ←− x0
yold ←− y0
ds ←− chosen step size
LOOP
    FX ←− \(F_x(xold, yold)\)
    FY ←− \(F_y(xold, yold)\)
    FM ←− \(\sqrt{FX^2 + FY^2}\)
    xnew ←− xold + ds*FX/FM
    ynew ←− yold + ds*FY/FM
    Draw line from \((xold, yold)\) to \((xnew, ynew)\)
    xold ←− xnew
    yold ←− ynew
EXIT_LOOP WHEN DONE
END_LOOP
```

Set starting point.

Choose step.

Start loop to generate steps.

Calculate components of force.

Calculate magnitude of force.

Step \(dx\) along force line.

Draw segment of force line.

Shift focus to new point.

Repeat.

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

```plaintext
trace_field, xold, yold, ds, n
```

where \texttt{xold} and \texttt{yold} convey the desired starting point for a field line, \texttt{ds} conveys the desired step size, and \texttt{n} stipulates the number of steps to be taken before completing the line and returning to the calling program. Though we might ultimately want to provide a means to define the field
outside of the tracing procedure, for purposes of this exercise suppose that the field of interest is given by

\[ F_x(x, y) = \frac{x - 1.5}{[(x - 1.5)^2 + y^2]^{3/2}} - \frac{x + 1.5}{[(x + 1.5)^2 + y^2]^{3/2}} \]

\[ F_y(x, y) = \frac{y}{[(x - 1.5)^2 + y^2]^{3/2}} - \frac{y}{[(x + 1.5)^2 + y^2]^{3/2}} \]

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, \( \text{trace\_field} \) 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. \( \text{Hint:} \) You may want to use the command \( \text{plot} \) with the keyword \( \text{nodata} \) set to \( \text{true} \) to establish the axes before beginning to trace any field lines. \( \text{Optional:} \) Refine the procedure so that

(a) the loop is terminated not only after execution of a fixed number of steps but also when the field line has gone out of bounds, approaches too closely to a singular point, or returns to a point too close to its starting point (closed field line),

(b) it obtains its field definitions from two functions whose names are supplied as arguments when it is called, and/or

(c) it uses a more refined predictor-corrector algorithm in which the values \( x_{\text{new}} \) and \( y_{\text{new}} \) determined above are instead regarded as \( x_{\text{pred}} \) and \( y_{\text{pred}} \), the field is calculated at that predicted point, and then a final step to \( x_{\text{new}} \) and \( y_{\text{new}} \) is made from \( x_{\text{old}} \) and \( y_{\text{old}} \) by using the average of the fields at the points \((x_{\text{old}}, y_{\text{old}})\) and \((x_{\text{pred}}, y_{\text{pred}})\).

2.35. Since equipotential curves are perpendicular to the field lines representing the associated (conservative) force field, an algorithm for tracing equipotential curves in two dimensions differs from the algorithm described in the previous exercise only by arranging for the stepping to occur at right angles to the calculated field rather than in the direction of the field. That objective is accomplished by replacing the two statements evaluating \( x_{\text{new}} \) and \( y_{\text{new}} \) in the algorithm in the previous exercise with the statements

\[ x_{\text{new}} \leftarrow x_{\text{old}} - ds*F_{\text{Y}/F_{\text{M}}} \]

\[ y_{\text{new}} \leftarrow y_{\text{old}} + ds*F_{\text{X}/F_{\text{M}}} \]

Following a pattern similar to that described in the previous exercise, cast the loop in this algorithm as an IDL procedure \( \text{trace\_equipot} \) that would be called with a statement like

\[ \text{trace\_equipot, xold, yold, ds, n} \]

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.

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

\[ \text{MAG} = \sqrt{(H_x)^2 + (H_y)^2} \]

for example, will generate an array \( \text{MAG} \) 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 2.12. Explore the magnetic field given in Eq. (2.19) in this way.
2.37. Read the IDL manuals about the procedures \texttt{flow3} and \texttt{scale3}. Then use \texttt{flow3} to explore the velocity field

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

(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. Hints: (1) Note that \texttt{flow3} will not function properly unless its invocation has been preceded by an appropriate invocation of \texttt{scale3}. (2) In establishing a scale, you may find it useful to assume a length scale \( a \), recast the field in the form

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

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

and try to draw graphs of the field \( \mathbf{v}/a\omega \) as functions of the parameter \( \alpha/a\omega \) in the space whose axes are labeled \( x/a, y/a, \) and \( z/a \).

2.19.7 Animation

2.38. 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

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

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 \( 1/n \).

2.39. 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 \texttt{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 pro-file to accomplish this task. Try your procedure with the two functions

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

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

2.40. 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 procedure pro-file in IDL to animate the motion of this drumhead 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 2.15.

2.19.8 Supplementary Exercises

2.41. Write a program to (a) accept as input an arbitrary two- to four-digit integer with at least two different digits, (b) rearrange the four digits to create (i) the largest and (ii) the smallest integers from the four digits, (c) subtract the smallest number from the largest numbers, and (d) repeat steps (b)–(c) until the same number occurs twice in succession. As a guard against an infinite loop, limit the number of iterations to 10. You should find, probably with some surprise, that the end result is always 6174, known as a Kaprekar Constant, and that no more than seven iterations will be required. **Note:** If the input number or any result after you complete step (c) has only three digits, add a leading zero before moving to the next iteration. **Note:** If you code appropriately, the adding of zeroes may happen automatically. **Optional:** Explore whether numbers with three or five digits exhibit a similar property and, if they do, identify the corresponding Kaprekar Constant. Google “Kaprekar Constant” to locate several articles on this and similar interesting numbers.
2.A Listing of lugen_grid

pro lugen_grid, x, xrange=xrange, nx=nx, y, yrange=yrange, ny=ny, z, zrange=zrange, nz=nz

; The procedure lugen_grid generates 2D and 3D arrays of evenly spaced x, y, and z points. Information about the third coordinate is optional and can be omitted, in which case two-dimensional arrays will be produced.

; In the call to this procedure, x, y, z are arrays in which the coordinates will be stored, xrange, yrange, zrange are two-component vectors giving the range of values on each axis, and nx, ny, nz are the number of subdivisions in each range.

; Set nz appropriately if not supplied
if not keyword_set(nz) then nz = 0

; Guard against entry of floating values
nx=fix(nx) & ny=fix(ny) & nz=fix(nz)

; Set spacing between grid points; create arrays for x and y
dx = ( xrange[1] - xrange[0] )/float(nx)
dy = ( yrange[1] - yrange[0] )/float(ny)
x = fltarr( nx+1, ny+1, nz+1 )
y = fltarr( nx+1, ny+1, nz+1 )

; Set spacing and create array for z (if necessary)
if nz ne 0 then begin
dz = ( zrange[1] - zrange[0] )/float(nz)
z = fltarr( nx+1, ny+1, nz+1 )
endif

; Generate lists of values of x, y, and (if necessary) z
xtmp = xrange[0] + dx*findgen(nx+1)
ytmp = yrange[0] + dy*findgen(ny+1)
if nz ne 0 then ztmp = zrange[0] + dz*findgen(nz+1)

; Write values into arrays for x, y, and (if necessary) z
for i = 0, nx do begin
    for j = 0, ny do begin
        for k = 0, nz do begin
            x[i,j,k] = xtmp[i]
y[i,j,k] = ytmp[j]
if nz ne 0 then z[i,j,k] = ztmp[k]
        endfor
    endfor
endfor
endfor
end
Chapter 8

Introduction toMATHEMATICA

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

Conceived and developed by Stephen Wolfram and first released to the public in June of 1988,
Mathematica® is a large program for manipulating expressions symbolically. Mathematica can
perform symbolic algebra; evaluate derivatives, integrals, and Taylor series; solve algebraic and
differential equations; manipulate complex numbers; find eigenvalues and eigenvectors; produce two-
and three-dimensional graphs; and accomplish numerous other tasks. In this chapter, we describe
some ways by which expressions can be defined for Mathematica and then illustrate some of the
statements by which Mathematica can be instructed to manipulate these expressions. Users should
also work through the on-line help messages within Mathematica itself and, in particular, look at
the examples with which most on-line help messages conclude (Section 8.2). Further details can
be found in printed and on-line Mathematica documentation (Section 8.19). We shall refer to this
documentation collectively as the Mathematica manuals.

Two user interfaces to Mathematica are provided, a text-based interface (TBI), which gives
direct access to the Mathematica kernel, and a notebook interface (NBI), which displays equations
in a more elegant form, offers assistance for those who have trouble remembering the commands,
and has a notebook capability to facilitate documentation. In essence, the TBI gives the user direct
access to the Mathematica kernel while the NBI is simply a front-end that facilitates construction of
statements that the NBI then formulates properly and passes on to the kernel, which may actually
be running on a different (and perhaps faster) computer than is used for the front-end. In this
book, we assume that the user is working in the NBI. We will, however, describe the structure of
input statements without addressing the ways in which the NBI facilitates assembling them, so the
statements we describe can equally well be submitted to the TBI. We shall present Mathematica
output with the approximate appearance it will have in the NBI.

Except where occasionally noted otherwise, for the rest of this book, information provided de-
scribes the NBI explicitly and may not accurately reflect the behavior of the TBI.

1 Mathematica is a registered trademark of and also a commercially available program produced and marketed
by Wolfram Research, Inc., of Champaign, Illinois. (See Appendix Z for full contact information.) Its use at any
particular site is subject to the provisions of whatever license that site has negotiated with Wolfram Research, Inc.
The terms of that license are explained in the Local Guide.

2 See the Mathematica manuals for information about how to use templates, palettes, and other aids available in
the NBI.
8.1 Beginning a Mathematica Session

Detailed instructions for initiating a session with Mathematica will be found in the Local Guide. In general, Mathematica will be started either (1) by typing the command `mathematica` for the NBI (or `math` for the TBI) at the prompt from the operating system,\(^3\) (2) by double-clicking ML on an appropriate icon on the desktop, or (3) by selecting ‘Mathematica’ (or an equivalent phrase) from a menu.\(^4\) Presently, an introductory screen offering a number of options will appear. Opening the New Document drop-down menu and selecting the option ‘Notebook’ produces a blank notebook titled ‘Untitled-1’ and a cursor—but no prompting character(s). We must from the outset be aware that

1. Labels, e.g., ‘\texttt{In[12]:=}’ for input and ‘\texttt{Out[12]=}’ for output, will appear after a statement has been submitted to Mathematica.

2. Internally, Mathematica is case sensitive, i.e., Mathematica treats upper and lower case letters as distinct. By convention, built-in Mathematica commands use names whose initial letter is capitalized.

3. By default, Mathematica formats input statements using Wolfram Language input. Other formats can be selected by opening the menu accessed by the ‘+’ sign that appears at the top of the notebook.

4. Lines of Mathematica code can be terminated in any of three ways, depending on whether we want to

   (a) submit the statement for execution and instruct Mathematica to display its response, in which case we terminate the statement by typing (SHIFT/ENTER) on the main keyboard or by typing ⟨ENTER⟩ on the numeric keypad.\(^5\)

   (b) submit the statement for execution and instruct Mathematica to suppress explicit display of its response, in which case we terminate the statement with a semicolon followed by (SHIFT/ENTER) on the main keyboard or ⟨ENTER⟩ on the numeric keypad.

   (c) continue entering the statement on a new line, in which case we terminate the line by typing simply ⟨ENTER⟩, but we must be clever in exploiting this feature. Mathematica will expect additional input only if the part entered before typing ⟨ENTER⟩ is incomplete. To enter the statement \texttt{a + b + c} on two lines, for example, we could not break it after the \texttt{b} but we could break it after either of the + signs.

5. Two or more statements can be placed in a single physical line by separating the statements with semicolons, though output will be suppressed for all but the last statement in the line (unless it is followed by a semicolon).

6. To abort execution of a statement by the Mathematica kernel and return Mathematica to a state from which new statements can be entered, either type ⟨ALT/.⟩ or, alternatively,

   • Select the block containing the statement whose execution is to be terminated by clicking ML on the corresponding bracket at the right end of the screen,

   • Open the Evaluation menu from the menu bar across the top of the notebook, and

   • Select ‘Abort Evaluation’ from that menu.

   There may be a short delay during which Mathematica will clean up its workspace before actually “prompting” for fresh input.

---

\(^3\)In this command to the operating system, case may be important.

\(^4\)Access to Mathematica is controlled by a license manager. When Mathematica is started, the license manager checks to make sure that a license is available. If no license is available (or the license manager happens not to be running), the desired session will not start. Restarting a stopped license manager requires action by the system administrator.

\(^5\)On some keyboards the ⟨ENTER⟩ key will be labeled ⟨RETURN⟩.
7. Typing the statement Quit[] whenever Mathematica is ready for input cleans Mathematica’s workspace and reinitializes its internal statement counter but does not close the notebook or return control to the operating system. To exit altogether from Mathematica, select ‘Exit’ from the File menu or click ML on the ‘×’ button in the upper right corner above the Mathematica menu bar. If the notebook contains unsaved content, a pop-up box will give you the option of saving that content before Mathematica actually returns control to the computer’s operating system.

8. To close the current notebook without exiting from Mathematica, select ‘Close’ from the File menu. If, after closing a notebook, you open a new notebook by selecting ‘New→Notebook’ from the File menu or ‘Notebook’ from the Mathematica opening screen, statement numbers in that new notebook will continue from where they left off in the closed notebook. To reinitialize the internal statement counter, execute the statement Quit[], preferably before closing the first notebook but alternatively right after opening the new notebook.

9. The Mathematica command Run[ "OSCommand" ] effects temporary escape from Mathematica to the underlying operating system (OS) to effect execution of OSCommand by that operating system. Control is then returned to Mathematica. The resulting output (if any) is displayed in the command window from which Mathematica was launched.6

10. Although the feature can be disabled, by default some types of error detected by Mathematica’s front end will produce an audible beep. For further information about that beep when it occurs, select ‘Why the beep?’ from the HELP menu.

### 8.2 On-Line Help

Mathematica provides extensive on-line help. Users can enter a hypertext browser by selecting ‘Wolfram Documentation’ (or some equivalent phrasing) from the HELP menu. The browser can be used to search for an item by entering the name of the item in the text entry box at the top of the browser and then clicking ML on the magnifying glass next to the text entry box; presently, a description of the item will appear. Alternatively, we can navigate the hierarchy of help messages by opening various submenus represented by colored boxes in the window that pops up when ‘Wolfram Documentation’ is selected.

Alternatively, information about a particular item can be obtained simply by typing a question mark ?—which must be the first character on the line—followed by the name of the item about which information is sought (e.g., ?Sin) whenever Mathematica is prepared to accept input. This action brings up an inline brief description of the specified item, including a link—the symbol >>—whose selection will bring up the full documentation for the item sought. Beginning a line with two question marks, e.g. ??Sin, appends to the message produced by one question mark a statement about the attributes assigned to the specified symbol.

Finally, when a specific Mathematica symbol is highlighted in a notebook, selecting ‘Find Selected Function’ from the HELP menu will bring up the full documentation for that function.

Items in the on-line help library that describe specific Mathematica commands typically include not only a description of the function and syntax of the command but also examples of the proper use of the command and links to further details and related topics at other points in the on-line messages. Many of the examples contain active statements that can be edited and reexecuted from within the help message. Your attention is drawn particularly to the on-line help messages for the commands D, Integrate, Simplify, Expand, Factor, Solve, Sum, and DSolve.

---

6In the NBI on any platform of course, this feature is unnecessary, since we can easily use resources of the desktop to execute other programs without exiting from Mathematica. Indeed, any output produced by Run[ "OSCommand" ] executed from the NBI will be displayed in a pop-up Shell window which, however, will promptly disappear when the command ends.
8.3 Basic Entities in Mathematica

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

- **expressions**, lists, and character strings, which are the primary quantities with which *Mathematica* works;
- **commands** and **functions**, which accept zero, one, or more arguments as input and return various outputs, thereby providing the means by which the objects enumerated in the previous item are manipulated;
- **system variables**, e.g., `$HomeDirectory`, `$IgnoreEOF`, `$Packages`, `$Path`, `$RecursionLimit`, `$TextStyle`, and `$TopDirectory`, whose values—default or user-specified—control the behavior of various system functions or record various pieces of information about the session. Some of these variables are maintained by *Mathematica* and are beyond the reach of individual users; others can be customized to the user’s specific ends.
- **constants**, including `Pi` for $\pi$, `I` for $\sqrt{-1}$ (which, except in exponents, will be printed in *Mathematica* output with a somewhat thicker vertical line but which we will display in output more simply as the lower case Roman character `i`), `Infinity` for $+\infty$, `E` for the base of the natural logarithm (which, except in exponents, will be printed in *Mathematica* output with an extra vertical line but which we will display in output more simply with the lower case Roman character `e`), and `True` and `False` for the two logical values.
- **special symbols**, including `+` (addition), `-` (subtraction), `*` (multiplication), `/` (division), `:=` (function definition), `/.` (perform replacement), `==` (equality), and `->` (transformation rule).
- **options**, e.g., `Assumptions`, `PlotRange`, `TextStyle`, and `TargetFunctions`, which may appear among the arguments to functions, may be assigned values, and—when present—modify the output produced by the functions. A full list of the options available, including the default value of each, for a particular function will be displayed when the statement `Options[ Function ]`, e.g., `Options[ Plot ]` is executed.

Consistent with *Mathematica* structure, many of these entities are represented internally in the standard form consisting of a head followed by arguments in the form of a list. The command `FullForm`, which will be introduced in Section 8.8.2, provides a convenient means to display the internal form of these—and many other—*Mathematica* objects.

8.4 Variable Names in Mathematica

Legal variable names begin with an alphabetic character and cannot coincide with reserved words like `True`, `Sin`, and `Factor` but are otherwise unrestricted. They can contain alphabetic and numeric characters but not the underscore character. Within reason, they can be of any length. Lower-case and upper-case letters are treated as distinct. Typing `{ESCAPE} followed by the spelled-out name of a Greek letter followed by `{ESCAPE}` in an input line will result in the display of the lower-case or upper-case Greek letter both in the input line and in the associated output line, e.g., `{ESCAPE}gamma{ESCAPE}` in an input line will produce $\gamma$ in the input and output lines while `{ESCAPE}Gamma{ESCAPE}` in an input line will produce $\Gamma$.

Note particularly that, in almost all situations, *Mathematica* will assume by default that any variable to which no explicit value has been assigned represents a complex quantity.

---

7Since the names for built-in *Mathematica* objects (commands, special variables, options, ...) begin with an upper-case letter, avoiding upper-case letters in user-defined variable names will avoid naming conflicts.

8This character is reserved in *Mathematica* for another purpose, to be outlined in Section 8.6.

9The actual maximum length depends on the operating system, but is quite large even in the most restrictive case.
8.5 Expressions in Mathematica

Expressions are written in Mathematica using a syntax very much like that used in ordinary algebra and are supplied merely by typing the expression at Mathematica’s prompt for input, e.g.,

\[
\text{In[12]:=} \quad x + a \ast \cos[b \ast x]
\]

to which, when the expression is entered by typing (SHIFT/ENTER), Mathematica will respond

\[
\text{Out[12]} = x + a \cos[bx]
\]

The symbols +, −, ∗, /, and ^ are used for addition, subtraction, multiplication, division, and exponentiation. Except for those in quoted strings, spaces are ignored and may be freely inserted for legibility. Functions available for use in expressions include

- General functions, e.g., Abs[], N[], Trunc[], Round[], Sqrt[], Sign[], Random[], and \[].
- Trigonometric, hyperbolic and exponential functions and their inverses, e.g., Sin[], Cos[], Tan[], ArcSin[], ArcCos[], ArcTan[], ArcTan[], Sinh[], Cosh[], Tanh[], ArcSinh[], ArcCosh[], ArcTanh[], Exp[], and Log[]. Note that each function name is capitalized and the argument(s) of the function is (are) enclosed in square brackets.
- Special functions, e.g., Bessel and modified Bessel functions, Airy functions, elliptic integrals, gamma and beta functions, the error function, Legendre polynomials, Hermite polynomials, and Laguerre polynomials.

Full details on these and other functions, including information about the form of their arguments, will be found in the Mathematica manuals. A few will be introduced in the remainder of this chapter.

Internally, Mathematica regards a simple expression to be composed of a head, which conveys the operator involved, followed by a list of the operands known individually as part one, part two, ... (The head can also be referred to as part zero.) Thus, for example, the two expressions

\[
a + b + c + d \quad \text{and} \quad abc
\]

are stored internally as the expression

\[
\text{Plus[ a, b, c, d ] \quad and \quad Times[ a, b, c ]}
\]

---

10In this section, we use non-sequential statement numbers to emphasize that many statements may intervene between those illustrated. These statements will all be executable as written, but your statement numbers will almost certainly differ from those presented here.

11In the TBI, the characters In[12]:= will be displayed as a prompt before the statement is typed. In the NBI, no prompt is displayed; the characters—best called a label in this case—will be added after the statement has been submitted for execution.

12See item 4 in Section 8.1 for alternatives. We shall make no further mention of the need for an appropriate termination to each statement.

13Note that we will occasionally rearrange the display of Mathematica output lines to improve legibility. No substance is lost in that rearrangement.

14Mathematica actually allows omission of the symbol ∗ for conveying multiplication. Thus, the statement \(x + a \cos[b \ast x]\) is fully equivalent to the one illustrated in line In[12]. Spaces separating variables to be multiplied are then critical, since \(xy\) would be interpreted as the two-character name of a (single) variable, not as the product of two variables \(x\) and \(y\). While recognizing the simpler appearance of statements in which the asterisk is omitted, we will throughout this book nonetheless note the multiplication explicitly.

15and those separating entities to be multiplied; see previous footnote.
respectively. More complicated expressions are stored as nested simple expressions. Thus, for example, the expressions

\[ a + b - c + d \quad \text{and} \quad a + b(c + d) + e \]

have the internal representations

\[
\text{Plus}[a, b, \text{Times}[-1, c], d] \quad \text{and} \quad \text{Plus}[a, \text{Times}[b, \text{Plus}[c, d]], e]
\]

respectively.

As a final example, which anticipates a later need, suppose the expression of interest is the product of three binomial factors, e.g.,

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

For this expression, we would generate Mathematica’s internal representation in several steps finding

\[
\text{Times}[(x + 3)^2, x - a, x + a] \\
\text{Times}[\text{Power}[x + 3, 2], \text{Plus}[-a, x], \text{Plus}[a, x]] \\
\text{Times}[\text{Power}[\text{Plus}[3, x], 2], \text{Plus}[-1, a], x], \text{Plus}[a, x]]
\]

Here, the first line shows the head and parts 1, 2, and 3 of the original expression while the second line reveals that part 1 of the original expression is itself a structure having a head (\text{Power}; the head of part 1 of the original expression), part 1 \((x+3);\) part 1 of part 1), and part 2 \((2;\) part 2 of part 1). Further, the third line reveals that some of the parts of the parts themselves have parts. Thus, for example, part 2 of part 1 of part 1 of the original expression is \(x\) while part 1 of part 2 of part 2 of the original expression is \(-1\). These relationships can be involved and confusing, but some manipulations depend critically on being able to identify the parts, subparts, subsubparts, and \ldots\ of a complicated expression.\footnote{The third line conveys the representation actually used in Mathematica. Unfortunately, Mathematica does not always store the various parts in the same order in which they were entered. After all, the expressions \text{Plus}[a, b, c] and \text{Plus}[b, a, c] are absolutely equivalent in value. However an expression is entered, Mathematica chooses a standard order to facilitate subsequent manipulations. Presently, we shall encounter the Mathematica command \text{FullForm}, which permits us to instruct Mathematica to display its internal representation.}

\section{8.6 Assigning Values to Variables; Defining Functions}

Especially when elaborate expressions are involved, we may want to refer to these expressions after they have been entered without having to retype them. Mathematica provides for this shorthand in two ways. First, as a session proceeds, Mathematica maintains an internal record of the contents of each \texttt{Out} line, identifying each with the number \(n\) that appears in square brackets after the word \texttt{Out}. The item at line \texttt{Out}[\textit{n}] can then be referred to in subsequent statements either by the label \texttt{Out}[\textit{n}] or by the label \%\textit{n}. For example, a later request that Mathematica square the expression on line \texttt{Out}[12] in the previous section could be communicated to Mathematica either with the statement

\[
\text{In}[18]:= \%12^2 \\
\text{Out}[18]= (x + a \cos[b x])^2
\]

or with the statement

\[
\text{In}[18]:= \%12^2 \\
\text{Out}[18]= (x + a \cos[b x])^2
\]
Alternatively, if we anticipate frequent reference to a particular expression, we may wish to use the assignment operator ‘\=' to assign or *bind* the expression to a variable. The statement,

\begin{verbatim}
In[20]:= bigexp = x + a * Cos[b*x]
Out[20]= x + a Cos[b x]
\end{verbatim}

for example, associates the variable *bigexp* with the expression following the assignment operator ‘\='. Subsequently, whenever that variable is included in an expression, it will automatically be replaced by the expression to which it has been bound, e.g., had we bound the expression to the variable *bigexp*, we might later have written

\begin{verbatim}
In[30]:= bigexp^2
Out[30]= (x + a Cos[b x])^2
\end{verbatim}

The wisdom of choosing variable names with mnemonic significance should be obvious.

Once a name has been assigned, we can at any time remind ourselves of the value bound to that variable simply by asserting its name in a statement like

\begin{verbatim}
In[47]:= bigexp
Out[47]= x + a Cos[b x]
\end{verbatim}

or, alternatively, by interrogating the system with a statement like

\begin{verbatim}
In[48]:= ?bigexp
Global 'bigexp
bigexp = x + a Cos[b x]
\end{verbatim}

Further, if we can’t remember whether we have already assigned a value to a particular variable, we can invoke the statement

\begin{verbatim}
In[75]:= Names[ "bigexp" ]
Out[75]= { bigexp }
\end{verbatim}

to find out.\(^{17}\) Finally, when we have no further need for a particular variable and wish to tell *Mathematica* to forget that we had once bound a value to it, we could enter either of the statements

\begin{verbatim}
In[97]:= Clear[ bigexp ]
In[98]:= Remove[ bigexp ]
\end{verbatim}

the first of which removes any value or definition assigned to the named variable but leaves the variable (and its attributes) known to *Mathematica*\(^{18}\) while the second of which completely deletes not only any assigned value or definition but also the name itself so the variable is no longer known to *Mathematica* in any way.

\(^{17}\)The statement will return a null list \(\{\}\) if no value has been assigned to the specified variable. Note that the wildcard \(\ast\) can be used in the string. Further, the statement *Names[]* will return a (long) list of *all* the symbols known, including those that *Mathematica* defines for its own purposes.

\(^{18}\)Alternatively, the shorter statement *bigexp\ast*. will achieve the same end.
When special functions of significance to a particular session would be useful, we might exploit the Mathematica operator := and the underscore suffix to a variable name\(^{19}\) to append one or more user-defined functions to the built-in functions. For example, the statements

\[
\text{In}[123]:= \text{spectrig}[x_] := 1 + \sin[x]^2 \\
\text{In}[124]:= \text{mode}[x_,y_] := \sin[\pi x] \cdot \sin[2\pi y]
\]

define the functions \text{spectrig}(x) and \text{mode}(x,y) for subsequent use just as we would use any of Mathematica’s built-in functions.\(^{20}\) Further, if we need to be reminded of the function definition given to a variable, we need simply enter statements like

\[
\text{In}[170]:= \text{?spectrig} \\
\text{In}[171]:= \text{?mode}
\]

Finally, when we have no further need for a particular function and wish to tell Mathematica to forget that we had once associated a function with a particular variable, we would enter an appropriate one or two of the statements

\[
\text{In}[201]:= \text{Clear}[\text{spectrig}]
\text{In}[202]:= \text{Remove}[\text{spectrig}]
\text{In}[203]:= \text{Clear}[\text{mode}]
\text{In}[204]:= \text{Remove}[\text{mode}]
\]

to remove the function definitions or the definitions and their names from Mathematica’s “memory”.

Within Mathematica, some potential variable names are protected. The names \text{Pi} (for π), \text{EulerGamma} (for Euler’s constant, 0.577216), and \text{Sin} (for the sine function), for example, are protected. An attempt to assign a value to a protected variable will fail, as illustrated in the statement

\[
\text{In}[227]:= \text{Pi} = 4.9 \\
\text{... Set: Symbol }\pi\text{ is Protected.}
\text{Out}[227]= 4.9
\]

though the content of the message may vary with the version of Mathematica. Further, in the illustrated example, both the symbol ‘...’ and the word ‘Set’ are hot keys that give access to additional information.

In rare circumstances, we may wish to use the (protected) symbol anyway. If so, we must unprotect it before using it in the new context, as illustrated in the statements

\[
\text{In}[240]:= \text{Unprotect}[\text{Pi}]; \text{Pi} = 4.9
\]

Once we are finished with the non-standard use, we can restore the original status of the variable with the statements

\[
\text{In}[256]:= \text{Pi} = .; \text{Protect}[\text{Pi}];
\]

---

\(^{19}\)This suffix, which is used only on the left-hand side of a function definition, flags the variable so that Mathematica will understand that whatever expression, however complicated, appears as the argument of the function is to be substituted for the variable without the suffix on the right-hand side in the function definition.

\(^{20}\)Note the use entirely of lower-case letters in naming the user-defined functions. By adhering to this convention, we avoid conflict with Mathematica’s reserved words.
where the first statement removes the non-standard definition, restoring the default value, and the second statement reestablishes the protection of the variable. Tampering with specially defined variables and other reserved words, however, is risky, and should be undertaken only occasionally and only for good reason.

8.7 Creating and Examining Lists

We have already mentioned in Section 8.3 that Mathematica works with several different types of object, and we have in Sections 8.5 and 8.6 described how to create expressions and assign values to variables. To simplify the quick tour of selected capabilities to be presented in Section 8.8, we include here a description of how lists and lists of lists can be created and examined.

A list is an ordered, comma-separated sequence of elements that is enclosed in braces \{\ldots\}. Lists are created with statements like

\begin{verbatim}
In[301]:= lst1 = {a, b, a, d, e}
Out[301]= {a, b, a, d, e}
In[302]:= lst2 = {"David", 5.00, a+b}
Out[302]= {"David", 5.00, a+b}
\end{verbatim}

An entire list can be referred to by its name (e.g., lst2), while a particular element in the list can be referred to by placing its index in double square brackets after the name of the list (e.g., lst2[[2]]).

Mathematica contains a few special commands to facilitate the creation of commonly needed lists, including the following statements:

\begin{verbatim}
In[303]:= Range[ 5 ]
Out[303]= {1, 2, 3, 4, 5}
In[304]:= Range[ 5, 10 ]
Out[304]= {5, 6, 7, 8, 9, 10}
In[305]:= Range[ 1, 2, 1/4 ]
Out[305]= {1, 5/4, 7/4, 9/4}
\end{verbatim}

Create a list containing values ranging from 1 to 5 in steps of 1.
Create a list containing values ranging from 5 to 10 in steps of 1.
Create a list containing values ranging from 1 to 2 in steps of \(\frac{1}{4}\).

We draw your attention also to the command Table, which can be used to create a list containing values that can be calculated from a formula. For example, the statement

\begin{verbatim}
In[306]:= Table[ n^2, {n, 4} ]
Out[306]= {1, 4, 9, 16}
\end{verbatim}

creates a list containing the squares of the integers from 1 to 4 while the statement

\begin{verbatim}
In[307]:= Table[ Sin[Pi*n/2], {n, 6} ]
Out[307]= {1, 0, -1, 0, 1, 0}
\end{verbatim}

creates a list of the values of sin(\(n\pi/2\)) for \(n\) ranging in integer steps from 1 to 6.

Lists of lists are also easily created. We simply place several lists, each of which is enclosed in braces, as the elements of a larger list. The statement

\begin{verbatim}
In[308]:= lst1 = {a, b, a, d, e}
Out[308]= {a, b, a, d, e}
In[309]:= lst2 = {"David", 5.00, a+b}
Out[309]= {"David", 5.00, a+b}
\end{verbatim}

Alternatively, a particular element of a list can be extracted with the function Part to be introduced presently. The element lst2[[2]] can also be identified as Part[lst2,2].
for example, produces the output

\[
\text{Out}[308]= \{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}\}
\]

In essence, the statement at In[308] creates a list of three elements, each of which is itself a list of three elements. The elements of this list of lists can be extracted with statements like

\[
\text{Out}[309]= \text{lstlst}[2]
\]

\[
\text{Out}[310]= 6
\]

Further, since a list of lists can be viewed as a two-dimensional matrix, Mathematica provides the statement

\[
\text{Out}[311]= \text{lstlst} // \text{MatrixForm}
\]

\[
\begin{pmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9
\end{pmatrix}
\]

for creating and displaying a matrix whose \( m \times n \) element is \( \sin(mx) \sin(ny) \).

8.8 Using Mathematica . . .

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

Note that the manner of editing statements submitted to Mathematica depends on whether the statement is entered in the TBI or the NBI. In the TBIs on some—but not all—platforms, the arrow keys can be used to retrieve previous statements and move the cursor within a statement, the backspace and delete keys can be used to remove characters, and other keys can be used to insert characters at the position of the cursor. Further—again in some TBI’s, input lines already existing on the screen can be highlighted and then copied to the current input point.

In the NBI, in-line editing can be accomplished in the same way using the left and right arrow keys to move the cursor, the backspace or delete key to remove characters, and other keys to insert new characters. The up and down arrow keys, however, move the cursor not only through the statements on input lines but also through components of the displayed output. The cursor can be located in previous statements by typing the up arrow key a sufficient number of times (or by repositioning the cursor with a click of ML), the statement can then be then be edited, and the edited form reexecuted, but the new statement will take the place of the old; it will not be copied to the new input focus in the window. Alternatively, a previous statement (or portion thereof) can be highlighted, copied into the clipboard by selecting ‘Copy’ from the Edit menu, and then—after clicking ML at the desired point of insertion—selecting ‘Paste’ from the Edit menu.

Note that, whichever interface is in use, the cursor need not be positioned at the end of a line when the line is to be submitted for execution by pressing (SHIFT/ENTER); the full line will be submitted regardless of the position of the cursor within the line.

To keep the several segments in this section separated from one another, we shall periodically submit the statement Quit[]. In the TBI, this statement will terminate the Mathematica session, so we will then have to restart Mathematica; in the NBI, this statement leaves the session running in the same Notebook window. In both cases, however, the action will clear Mathematica’s workspace and, in particular, reinitialize the statement counter. The action is, of course, not necessary, and the entire “conversation” could have been “held” perfectly well without these occasional breaks or—for that matter—with breaks in different places, so long as the breaks do not occur between the point at which a variable is defined and the point at which it is used. The “conversation” in this section should be started with a fresh invocation of Mathematica.

Finally, note that, throughout this section, various components of Mathematica (functions, commands, operators, . . . ) are introduced simply by presenting statements in which they are used, frequently with brief comment. Note also that, in the NBI, various entered components will be grouped together into cells, a grouping conveyed by square brackets along the right edge of the Notebook window; the function of these groupings will be addressed briefly in Section 8.16. Details on all mentioned components and on many others can be found in the Mathematica manuals.22

8.8.1 . . . for Arithmetic

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

\[
\text{In}[1]:= 2 + 3 \\
\text{Out}[1]= 5
\]

Enter a simple numeric expression, terminating with (SHIFT/ENTER). Mathematica returns a value, doing the arithmetic.

---

22Input statements will look much the same regardless of the platform on which we are operating. Output is here presented approximately as it would appear in the NBI. In the TBI, the content of the output will be the same but, because that interface is limited to character output, its appearance will be less “pretty”.

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

Compute \( \sqrt{9} \). Note the capitalization of `Sqrt` and the *square* brackets enclosing the argument.

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

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

Find 9 factorial with postfix operator `!`.

Ask for floating value with default number of digits.

Ask for floating value with default number of digits with postfix operator `//`.

Ask for floating value with 20 digits.\(^\text{23}\)

Ask for floating value of \( \pi \) with 32 digits.

(or `Exit[]`)

We have, of course, illustrated only some (+, /, `Sqrt`, `Factorial`, !, N, `//N`) of the operators and functions that might be used to persuade *Mathematica* to manipulate with numbers. You should make it a point to read about these and other commands, functions, options, constants, and operators in the *Mathematica* manuals.

8.8.2 ... for Algebra

*Mathematica* is, of course, capable of much more than simple arithmetic. We next illustrate how algebraic manipulations can be requested, beginning with the statements

\[ \text{In}[1]:= (x+3)^2*(x^2-a^2) \]

\[ \text{Out}[1]= (3 + x)^2(-a^2 + x^2) \]

Enter an expression. Note that the order of the various pieces in the output may differ from that in the input. Compare the response produced on your system by this statement to the response produced by the statement \((x^2-a^2)*(x+3)^2\).

\(^{23}\)While the statement `N[4/6, 20]` yields an evaluation with 20 digits, the statement `N[4/6, 10]` will on some platforms yield the same value as is displayed by the statement `N[4/6]`. Indeed, on those platforms, `N[4/6, n]` yields the same value as `N[4/6]` for any \( n \) less than 17. Only for \( n \geq 17 \) does *Mathematica* display a value with the specified number of digits. This behavior stems from the nature of the communication between the kernel (which does the computing) and the front end (which displays the results).
8.8. USING MATHEMATICA...

In[2]:= Expand[ % ]
Out[2]= $-9a^2 - 6a^2x + 9x^2 - a^2x^2 + 6x^3 + x^4$

Expand the expression. Again, the order of terms may be platform and interface dependent. Here, the symbol % refers to the most immediate past output. (%% and %%% refer to the penultimate and antepenultimate outputs, respectively. The symbol %n refers to the output on line Out[n].)

In[3]:= fct = Factor[ % ]
Out[3]= $-(a - x)(3 + x)^2(a + x)$

Factor the expression in the previous output, assigning the result to the variable fct.

We note, of course, that this factoring hasn’t quite returned us to our starting point. To achieve that end, we need to multiply the first and third factors without expanding the second. To do so, we will have to carve the expression apart, manipulate with its parts, and then reconstruct the expression. As discussed in Section 8.5, Mathematica sees this expression as a structure with nested heads and lists of operands. Indeed, we can persuade Mathematica to reveal the first level in that structure by exploiting the Mathematica functions Head and Part in the statements.

In[4]:= {Head[fct], Part[fct,1], Part[fct,2], Part[fct,3], Part[fct,4]}
Out[4]= {Times, $-1$, $a - x$, $(3 + x)^2$, $a + x$}

Alternatively, we could exploit not only the function Part but also Mathematica’s ability to construct a loop in the statement

In[5]:= For[ i=0, i<=4, i++, Print[ Part[fct,i] ] ]
Times
$-1$
$a - x$
$(3 + x)^2$
$a + x$

to show all operands of the expression at once. (Note that we have also here illustrated that the head of an expression can be referred to as part 0 of the expression.)

Yet another way to reveal how Mathematica sees the expression fct internally exploits the command FullForm, as in the statement

In[6]:= FullForm[ fct ]

Once again, we see that fct is a product of four factors $[-1, a - x, (3 + x)^2, a + x]$, though here we have a fuller breakdown of the factors into their own parts.

Now, using the capabilities of the command Part, we might achieve the entire objective of recasting the expression fct with the single nested statement.

Out[7]= $-(3 + x)^2(a^2 - x^2)$

24While the number of operands is invariant, their order may be platform, interface, and perhaps context dependent. Compare the parts identified here by the integers 0, 1, 2, and 3 with the parts identified by those integers in your environment.

25We repeat the previous footnote: The order of the factors in fct may be platform, interface, and perhaps context dependent. Thus, the numbers identifying various operands in your environment may differ from those used here. Beware!
in which we extract parts 1, 2, 3, and 4 from \texttt{fct}, expand the product of parts 2 and 4, and then reconstruct the expression by multiplying that result by parts 1 and 3. This time, we have succeeded in reconstructing the original expression. For comparison, note that this expression has the internal representation

\begin{verbatim}
In[8]:= FullForm[ fct ]
\end{verbatim}

We continue with some additional manipulations. First, using the replacement operator \texttt{/}. and the operator \texttt{->} (which ultimately prints as \texttt{\rightarrow} in the NBI) to specify a transformation rule, we set the constant \texttt{a} to the specific value 2 with the statement

\begin{verbatim}
In[9]:= fct /. a -> 2
Out[9]= - (3 + x)\(^2\)(4 - x\(^2\))
\end{verbatim}

Alternatively, we might have embedded the transformation rule in the command \texttt{Replace} by using the statement\footnote{Some versions of \textit{Mathematica} prior to Version 4.1 did not admit the third argument in the command \texttt{Replace}.}

\begin{verbatim}
In[10]:= Replace[ fct, a->2, 3 ]
Out[10]= -(3 + x)\(^2\)(4 - x\(^2\))
\end{verbatim}

Here the third argument specifies how deeply into the structure of \texttt{fct} the replacement is to be carried. Since the \texttt{a} we wish to replace appears at level 3 in \texttt{fct} (see the output of \texttt{FullForm[fct]}), we must extend the replacement to level 3 to achieve the desired substitution. However we make the replacement, we could then find a partial fraction expansion of the reciprocal of this polynomial with the statement

\begin{verbatim}
In[11]:= eq1 = Apart[ 1/%, x ]
Out[11]= 1/100(-2 + x) + 1/4(2 + x) + 1/5(3 + x)\(^2\) + 6/25(3 + x)
\end{verbatim}

Numerous rewrites of this expression are now possible. For example, replacing \texttt{x} with 3\texttt{z} would be accomplished by using the \texttt{Replace} command in the statement

\begin{verbatim}
In[12]:= Replace[ eq1, x -> 3*z, 4 ]
Out[12]= 1/100(-2 + 3z) + 1/4(2 + 3z) + 1/5(3 + 3z)\(^2\) + 6/25(3 + 3z)
\end{verbatim}

Alternatively, this same replacement could be accomplished with the statement

\begin{verbatim}
In[13]:= eq1 /. x -> 3*z
\end{verbatim}

though we suppress the output, since it is the same as that at line Out[12].

To illustrate a more complicated recasting, suppose that we wanted to square the denominator of the third term in \texttt{eq1} without affecting any other term. The statement \texttt{Expand[eq1]} does nothing. (Try it.) We need a more sophisticated operation that rearranges only the third term (part 3 of the original expression). To achieve the desired end, we extract that term, carve it up, manipulate with the proper piece, and substitute that piece into the proper point in the original expression. The task takes several steps, beginning with the identification of the proper part with the statements
Extract third part of `eq1`.

```math
\text{Out}[14]= \frac{1}{5(x+3)^2}
```

Extract second part of third part of `eq1`.

```math
\text{Out}[15]= \frac{1}{(x+3)^2}
```

Alternatively, extract the second part of the third part of `eq1` in one step.

```math
\text{Out}[16]= \frac{1}{(x+3)^2}
```

Then we expand the denominator of this piece of `eq1` and substitute it into `eq1` as the second part of the third part with the statements

```math
\text{Out}[17]= \frac{1}{9+6x+x^2}
```

(Note that the particular part to be replaced (part \{3,2\}) is conveyed to \texttt{ReplacePart} with a \texttt{list}.

```math
\text{Out}[18]= \frac{1}{100(-2+x)} - \frac{1}{4(2+x)} + \frac{6}{25(3+x)} + \frac{1}{5(9+6x+x^2)}
```

We have, of course, only mentioned a few (\texttt{Expand}, \texttt{Factor}, \texttt{Part}, \texttt{Replace}, \texttt{ReplacePart}) of the numerous functions \textit{Mathematica} makes available for manipulating expressions algebraically. We have also introduced the control structure \texttt{For}, the postfix replacement operator \texttt{//}, and the special symbol \texttt{\rightarrow} (which ultimately is displayed as \texttt{\rightarrow} in the NBI). The differences among several commands, including \texttt{Expand}, \texttt{ExpandAll}, \texttt{Factor}, \texttt{Simplify}, \texttt{FullSimplify}, \texttt{Together}, \texttt{Collect}, \texttt{Apart}, \texttt{Cancel}, \texttt{Numerator}, \texttt{Denominator}, and \texttt{Factor} are explored in one of the exercises. You should make it a point to read about these functions and any associated options in the \textit{Mathematica} manuals.

### 8.8.3 ... for Complex Variables

\textit{Mathematica} is also capable of manipulating complex variables. Indeed, in most contexts, \textit{Mathematica} assumes that any variable to which no specific value has been assigned represents a complex value. We illustrate some of \textit{Mathematica}'s main capabilities in this category with the statements

```math
\text{Out}[1]= \sin(x+y\cdot i)
```

Define complex expression. The constant \texttt{i} stands for $\sqrt{-1}$ and is displayed as \texttt{i} in output.

```math
\text{Out}[2]= \cosh(y)\sin(x) + \text{i}\cos(x)\sinh(y)
```

Recast \texttt{fct} in form $a + ib$. In contrast to the default behavior, the function \texttt{ComplexExpand} assumes that variables to which no value has been assigned represent \textit{real} values.
CHAPTER 8. INTRODUCTION TO MATHEMATICA

In[3]:= ComplexExpand[ Re[fct] ]

In[4]:= ComplexExpand[ Im[fct] ]

In[5]:= ComplexExpand[ Conjugate[fct] ]

Extract real part of fct.

Extract imaginary part of fct.

Find complex conjugate of fct.

In[6]:= fct1 = (x + i*y)/(x - i*y)
Out[6]= x + i y

In[7]:= ComplexExpand[ fct1, TargetFunctions -> {Re, Im} ]
Out[7]= \frac{x^2 + 2 i x y}{x^2 + y^2} + \frac{y^2}{x^2 + y^2}

Recast fct1 in form a + ib. ComplexExpand usually tries to express its output in terms of the real and imaginary parts of the input variables. Sometimes, however, the option TargetFunctions must be used to specify which of Re, Im, Abs, Arg, Conjugate and Sign are to be used. We include that option here simply to illustrate it; the option is not actually necessary in this context.

Find complex absolute value of previous result.

And simplify it.

Find complex absolute value of fct1, noting that this route includes the simplification that we had to request explicitly in the previous route.

Find complex argument of fct1. In Mathematica, ArcTan[a, b] = tan^{-1}(b/a).

In[8]:= ComplexExpand[ Abs[fct1] ]
Out[8]= \sqrt{4 x^2 y^2 + \left(x^2 + y^2\right)^2}

In[9]:= Simplify[ % ]
Out[9]= 1

In[10]:= ComplexExpand[ Abs[ fct1 ] ]
Out[10]= 1

In[11]:= ComplexExpand[ Arg[ fct1 ], TargetFunctions -> {Re, Im} ]
Out[11]= ArcTan[\frac{x^2 + 2 i x y}{x^2 + y^2}, \frac{y^2}{x^2 + y^2}]

Simplify it, though Mathematica doesn’t seem to recognize that ArcTan[x^2 - y^2, 2yx] would be even simpler.

To conclude this section, we record quickly how to convert complex numbers from Cartesian form \(z = x + iy\) to polar form \(|z|e^{i\text{Arg}(z)}\) and from polar form to Cartesian form. Starting with a complex number in Cartesian form, we would use statements like

In[13]:= fct3 = x + I*y
Out[13]= x + i y

In[14]:= ComplexExpand[ Abs[ fct3 ], TargetFunctions -> {Re, Im} ]
Out[14]= \sqrt{x^2 + y^2}

In[15]:= ComplexExpand[ Arg[ fct3 ], TargetFunctions -> {Re, Im} ]

to find the absolute value and argument of the number. Conversely, we would use statements like
8.8. USING MATHEMATICA ...

\[\text{In[16]:= } fct4 = r \ast \text{Exp}[\text{I}\ast\text{Theta}]\]
\[\text{Out[16]= } e^{i\theta}r\]
\[\text{In[17]:= } \text{ComplexExpand}\left[\text{ExpToTrig}[fct4]\right]\]
\[\text{Out[17]= } r \cos[\theta] + ir \sin[\theta]\]
\[\text{In[18]:= } \text{Quit}[\]}

to find the Cartesian expression of a complex number originally given in polar form.

We have, of course, only mentioned some (\text{ComplexExpand}, \text{Re}, \text{Im}, \text{Conjugate}, \text{Abs}, \text{Arg}, \text{ExpToTrig}) of the functions \text{Mathematica} makes available for manipulating complex numbers. Further, we have also introduced the option \text{TargetFunctions} for the command \text{ComplexExpand}, and we have seen additional contexts in which the command \text{Simplify} can be valuable. You should make it a point to read about these functions and any associated options in the \text{Mathematica} manuals.

8.8.4 ... for Trigonometry

\text{Mathematica} also makes available a number of commands and functions for manipulating with trigonometric, hyperbolic, and exponential functions. Note, for example, the capabilities illustrated in the statements

\[\text{In[1]:= } \text{trig1 }= \sin[x] \ast \cos[x]\]
\[\text{Out[1]= } \cos[x] \sin[x]\]
\[\text{In[2]:= } \text{TrigToExp}[\text{trig1}]\]
\[\text{Out[2]= } \frac{1}{4}i e^{-2ix} - \frac{1}{4}i e^{2ix}\]
\[\text{In[3]:= } \text{ExpToTrig}[\text{trig1}]\]
\[\text{Out[3]= } \frac{1}{2} \sin[2x]\]
\[\text{In[4]:= } \text{TrigReduce}[\text{trig1}]\]
\[\text{Out[4]= } \frac{1}{2} \sin[2x]\]
\[\text{In[5]:= } \text{TrigExpand}[\text{trig1}]\]
\[\text{Out[5]= } \cos[x] \sin[x]\]
\[\text{In[6]:= } \text{trig2 }= \sin[2x+y]\]
\[\text{Out[6]= } \sin[2x+y]\]
\[\text{In[7]:= } \text{TrigExpand}[\text{trig2}]\]
\[\text{Out[7]= } 2 \cos[x] \cos[y] \sin[x] + \cos[x]^2 \sin[y]
- \sin[x]^2 \sin[y]\]
\[\text{In[8]:= } \text{TrigReduce}[\text{trig2}]\]
\[\text{Out[8]= } \sin[2x+y]\]
\[\text{In[9]:= } \text{TrigExpand}[\text{trig2}\text{/}.x\to z/2\text{/}.z\to 2\ast x]\]
\[\text{Out[9]= } \cos[y] \sin[2x] + \cos[2x] \sin[y]\]
\[\text{In[10]:= } \text{TrigExpand}[\text{trig2}]\]
\[\text{Out[10]= } 2 \cos[x] \cos[y] \sin[x] + 2 \cos[x]^2 \sin[y]
- \sin[x]^2 \sin[y]\]

Define a trig expression.

Convert to exponential form.

Convert to standard trig functions.

Recast with multiple angle identities.

Expand functions of multiple angles.

Define another trig function.

Expand fully using trig addition formulae.

Return to original form.

Expand, but don't expand 2x. Here, we replace 2x with z for the expansion but then replace z with 2x after the expansion has been evaluated.

Expand again to complete expansion.
We have, of course, only mentioned a few (TrigToExp, ExpToTrig, TrigReduce, TrigExpand, and Expand) of the functions Mathematica makes available for doing trigonometry. You should make it a point to read about these and other commands and any applicable special arguments or options in the Mathematica manuals.

8.8.5 ... for Algebraic Equations

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

\[
\text{In[1]:=} \quad \text{poly = Expand[ (x^2-a^2)*(x+3)^2 ];} \quad \text{Define a polynomial.}
\]

\[
\text{In[2]:=} \quad \text{soln = Solve[ \{\text{poly == 0}, \{x\}\];} \quad \text{Solve the equation poly = 0 for } x. \text{ There are four roots, though two of them are identical.}
\]

Note that the arguments to the command Solve are lists, and that the roots have been returned as a list of lists of transformation rules—a behavior that would seem more reasonable had we been solving a system of equations for more than one unknown. In the simple case of a single equation and a single unknown, the braces can be omitted, as in

\[
\text{In[3]:=} \quad \text{soln1 = Solve[ poly == 0, x ];} \quad \text{Note also the use of the double equal sign in setting up a logical equality between two quantities as opposed to the assignment of a value to a variable (for which Mathematica uses a single equal sign).}
\]

Of course, once we have obtained roots, we should verify their correctness by backsubstitution into the original equation. To do so, we would execute statements like

\[
\text{In[4]:=} \quad \text{poly /. x -> -a;} \quad \text{Verify first root.}
\]

\[
\text{In[5]:=} \quad \text{poly /. soln;} \quad \text{Verify all roots in a single statement.}
\]

\[
\text{In[6]:=} \quad \text{For[i=1, i<=4, i++,} \quad \text{Print[ poly /. soln[[i]] ]} \quad \text{Verify all roots, using the For structure to code a loop.}
\]

Perhaps the most famous of all solutions is that of a quadratic equation. The command Solve clearly knows the quadratic formula, as illustrated by the statements
8.8. USING MATHEMATICA . . .

In[7]:= a\*x^2 + b\*x + c;
In[8]:= Solve[ %==0, x ]

Out[8]=

\[
\begin{cases}
  x \rightarrow -\frac{-b - \sqrt{b^2 - 4ac}}{2a}, \\
  x \rightarrow -\frac{-b + \sqrt{b^2 - 4ac}}{2a}
\end{cases}
\]

In[9]:= % /. {a->3, b->4, c->2}

Out[9]=

\[
\begin{cases}
  x \rightarrow -\frac{1}{6} \left(-4 - 2i\sqrt{2}\right), \\
  x \rightarrow -\frac{1}{6} \left(-4 + 2i\sqrt{2}\right)
\end{cases}
\]

We can, of course, always ask for floating point evaluations with the command \texttt{N}, finding here that

In[10]:= N[ % ]

Out[10]= \{
  \{x \rightarrow -0.666667 + 0.471405i\},
  \{x \rightarrow -0.666667 - 0.471405i\}\}

Finally, let us illustrate \textit{Mathematica}'s capability to solve simultaneous linear equations. To define and solve a representative system, we invoke the statements

In[11]:= eqn1 = x - y == 4
In[12]:= eqn2 = x + y == 0
Out[12]= x + y == 0

In[13]:= Solve[ {eqn1, eqn2}, {x,y} ]

Out[13]= \{
  \{x \rightarrow 2, y \rightarrow -2\}\}

Note that the equations and the variables are presented to \texttt{Solve} as lists and that each solution—here there is only one—is presented as a list. Were there several solutions, \textit{Mathematica} would have returned a list of lists.

We need not, of course, have specific numerical values for the coefficients in the equations to be solved. Solution of a more symbolic system of linear equations is illustrated in the statements

In[14]:= eqn1 = a\*x + b\*y == e;
In[15]:= eqn2 = c\*x + d\*y == f;
In[16]:= Solve[ {eqn1, eqn2}, {x,y} ]

Out[16]= \{
  \{x \rightarrow \frac{-de - bf}{bc - ad}, y \rightarrow \frac{-ce + af}{bc - ad}\}\}

In[17]:= Remove[ eqn1, eqn2 ]

In[18]:= Quit[]

We have, of course, only mentioned one (\texttt{Solve}) of the functions \textit{Mathematica} makes available for solving algebraic equations exactly. Your attention is drawn also to the function \texttt{Nsolve} for finding solutions numerically and to the function \texttt{LinearSolve}, which is built to solve matrix equations of the form $Ax = b$. You should make it a point to read about these functions in the \textit{Mathematica} manuals. Rootfinding is discussed in greater depth in Chapter 14.
8.8.6 ... for Manipulating Vectors and Matrices

Vectors and matrices represent numerous physical quantities. We have already seen in Section 8.7 how to create lists. Vectors are simply lists and (two-dimensional) matrices are lists of lists. Once they have been created, Mathematica can manipulate these objects in numerous ways. To illustrate some of these capabilities, we begin by defining a matrix and evaluating its transpose with the statements

\begin{verbatim}
In[1]:= mat1 = {{1,2}, {3,4}};
In[2]:= MatrixForm[ mat1 ]
Out[2]//MatrixForm=
\begin{array}{c}
1 \\
3
\end{array}
\begin{array}{c}
2 \\
4
\end{array}

In[3]:= mat2 = Transpose[ mat1 ];
In[4]:= MatrixForm[ mat2 ]
Out[4]//MatrixForm=
\begin{array}{c}
1 \\
2
\end{array}
\begin{array}{c}
3 \\
4
\end{array}
\end{verbatim}

Then, we evaluate the inverse, the trace, and the determinant of the matrix with the statements

\begin{verbatim}
In[5]:= MatrixForm[ Inverse[ mat1 ] ]
Out[5]//MatrixForm=
\begin{array}{cc}
-2 & 1 \\
3 & -1 \\
2 & 2
\end{array}

In[6]:= val1 = Tr[ mat1 ]

In[7]:= val2 = Det[ mat1 ]
\end{verbatim}

Other common manipulations include evaluating the standard matrix product and an element-by-element product. The standard (non-commutative) matrix product is specified by the infix operator ‘.’ in a statement like

\begin{verbatim}
In[8]:= MatrixForm[ mat1.mat2 ]
Out[8]//MatrixForm= 
\begin{array}{c}
5 \\
11
\end{array}
\begin{array}{c}
11 \\
25
\end{array}
\end{verbatim}

Element-by-element (commutative) multiplication is accomplished with the infix operator ‘*’ in a statement like

\begin{verbatim}
In[9]:= MatrixForm[ mat1*mat2 ]
Out[9]= 
\begin{array}{c}
1 \\
6
\end{array}
\begin{array}{c}
6 \\
16
\end{array}
\end{verbatim}

Finally (for this section anyway), Mathematica includes functions for finding eigenvalues and eigenvectors, which we illustrate with the statements
8.8. USING MATHEMATICA ...

In[10]:= spinx = { {0, 1/Sqrt[2], 0},
{1/Sqrt[2], 0, 1/Sqrt[2]},
{0, 1/Sqrt[2], 0} };

Enter a matrix bound to the variable spinx.

In[11]:= MatrixForm[ % ]

Display the matrix in matrix form.

Out[11]=
\[
\begin{pmatrix}
0 & \frac{1}{\sqrt{2}} & 0 \\
\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\
0 & \frac{1}{\sqrt{2}} & 0
\end{pmatrix}
\]

In[12]:= Eigenvalues[ spinx ]

Find eigenvalues.

Out[12]= \{-1, 1, 0\}

In[13]:= Eigenvectors[ spinx ]

Find eigenvectors.

Out[13]= \{\{1, -\sqrt{2}, 1\}, \{1, \sqrt{2}, 1\}, \{-1, 0, 1\}\}

In[14]:= Eigensystem[ spinx ]

Find eigenvalues and eigenvectors with a single statement.

Out[14]= \{\{-1, 0, 1\}, \{1, -\sqrt{2}, 1\}, \{-1, 0, 1\}, \{1, \sqrt{2}, 1\}\}\}

In[15]:= MatrixForm[%]

Display result in matrix form, in which the first row contains the eigenvalues and the second row contains the associated eigenvectors.

Out[15]=
\[
\begin{pmatrix}
-1 & 1 & 0 \\
1, -\sqrt{2}, 1 & 1, \sqrt{2}, 1 & \{-1, 0, 1\}
\end{pmatrix}
\]

In[16]:= Quit[]

We have, of course, only mentioned a few (MatrixForm, Transpose, Inverse, Tr, Det, Eigenvalues, Eigenvectors, Eigensystem) of the functions available for manipulating matrices. Important additional functions include AppendColumns, AppendRows, BlockMatrix, TakeRows, TakeColumns, TakeMatrix, SubMatrix, ZeroMatrix, and LinearEquationsToMatrix, which are contained in the add-on package\textsuperscript{27} LinearAlgebra'MatrixManipulation' and the function Normalize, which is contained in the add-on package LinearAlgebra'Orthogonalization'. You should make it a point to read about these functions and any associated arguments in the Mathematica manuals.

8.8.7 ... for Basic Calculus

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

\textsuperscript{27} Packages are discussed in Section 8.9.
In[1]:= x^2 * Exp[-x^2]  
Out[1]= e^{-x^2}x^2  

In[2]:= D[ %, x ]  
Out[2]= 2e^{-x^2}x - 2e^{-x^2}x^3  

In[3]:= Simplify[ D[ %, {x,2} ] ]  
Out[3]= 2e^{-x^2}(1 - 5x^2 + 2x^4)  

To evaluate integrals, we would instead use statements like  
In[4]:= Integrate[ %2, x ]  
Out[4]= e^{-x^2}x^2  

In[5]:= int1 = x/(a^2+x^2)^(3/2)  
Out[5]= x/(a^2+x^2)^(3/2)  

In[6]:= Integrate[ int1, {x, 0, Infinity} ]  
Out[6]= ConditionalExpression[1/\sqrt{a^2}, \text{Re}[a] \neq 0]  

Interestingly, Mathematica in this case returns a result revealing that the integral is not well defined. In the returned conditional expression the first argument is the evaluation of the integral, but only when the condition in the second argument is true. If the specified condition is false (\text{Re}(a)=0), then the denominator in the integrand passes through zero somewhere in the interval of integration and the integrand is indeed ill defined. If, however, the specified condition is true, then the integrand is well defined, having the value \frac{1}{\sqrt{a^2}}, a result not written as \frac{1}{a} because a is not known to be positive or even real.  

We can instruct Mathematica to be more specific by exploiting an optional argument to \textbf{Integrate}, as for example in the statements  
In[7]:= Integrate[ int1, {x, 0, Infinity}, Assumptions -> \{\text{Im}[a] == 0\} ]  
Out[7]= ConditionalExpression[1/\text{Abs}[a], a \neq 0]  

In[8]:= Integrate[ int1, {x, 0, Infinity}, Assumptions -> \{a > 0\} ]  
Out[8]= \frac{1}{a}  

Finally (for this section), we point out that Mathematica can generate series expansions of prescribed functions:  
In[9]:= 1/(1+Exp[-x])  
Out[9]= \frac{1}{1+e^{-x}}  

In[10]:= ser = Series[ %, \{x, 0, 7\} ]  
Out[10]= \frac{1}{2} + \frac{x}{4} - \frac{x^3}{48} + \frac{x^5}{480} - \frac{17x^7}{80640} + O[x^8]  

While a series displays on the screen as an algebraic expression, internally its form differs from that of many other (i.e., of ordinary) expressions. The statement
8.8. USING MATHEMATICA

In[11]:= FullForm[ ser ]
Out[11]//FullForm= SeriesData[
  x, 0, List[
    Rational[1,2], Rational[1,4], 0, Rational[-1,48],
    0, Rational[1,480], 0, Rational[-17,80640]], 0, 8, 1]

reveals that the head of an expression that conveys a series is SeriesData. The remaining parts of
the expression are, in order, the variable \( x \) in which the series is expressed, the value \( x_0 \) (here 0) of
\( x \) about which the expansion is made, a list of the coefficients of the successive powers of \( x - x_0 \) in
the expansion, and two arguments (here 0 and 8) which convey the starting and ending powers as
multiples of the final argument (here 1). The powers appearing thus are 0/1, 2/1, 3/1, ... 8/1.

Unfortunately, for some operations, the expression of a series in the form dictated by SeriesData
cannot be used. In those cases, we may need to convert the series into an ordinary expression by
invoking the command Normal with the statement

In[12]:= pol = Normal[ ser ]
Out[12]= \( \frac{1}{2} + \frac{x}{4} - \frac{x^3}{48} + \frac{x^5}{480} - \frac{17 x^7}{80640} \)

In[13]:= FullForm[ pol ]
Out[13]= Plus[ Rational[1,2], Times[Rational[1,4],x], Times[Rational[-1,48],Power[x,3]],
  Times[Rational[1,480],Power[x,5]], Times[Rational[-17,80640],Power[x,7]] ]

In[14]:= Quit[]

Note that the omitted terms in the series are not indicated in \( pol \) and that \( pol \) has expression type
Plus. Further, compare the storage of a series as revealed by the statement FullForm[ser] with
the storage of a polynomial as revealed by the statement FullForm[pol].

By default, Mathematica commands automatically carry out particular operations that, in
Mathematica’s mind, simplify expressions submitted to them. As we have seen above, the com-
mand \( D \) for differentiation, for example, will result in immediate evaluation of any derivatives that
can be evaluated. Sometimes, however, we may wish to enter an expression and suppress imme-
diate evaluation. Mathematica provides the command HoldForm for that purpose and the command
ReleaseHold to request explicit evaluation at a later time. To illustrate, suppose we define the
function

\[ f[y_] := a \sin(\Omega y) \]

Then, the two statements

\[ D[ f[y], y ] \]
Out[2]= \( a\omega \cos[y\omega] \)

In[3]:= d1 = HoldForm[ D[ f[y], y ] ]
Out[3]= \( \partial_y f[y] \)

reveal the difference. In the first, the derivative is evaluated; in the second, it is left unevaluated.
At some later time, we could evaluate the derivative in this last expression with the statement

In[47]:= ReleaseHold[ d1 ]
Out[47]= \( a\omega \cos[y\omega] \)
This feature might, for example, be useful if we had several functions whose derivative we wished to evaluate. We would simply define the derivative with \texttt{HoldForm}. Then, we could evaluate the derivative of each new function by changing the definition of $f[y]$ before executing the simple statement illustrated on Line In[47].

\texttt{In[48]:= Quit[]}

We have, of course, only mentioned a few (\texttt{D, Integrate, Series, Normal}) of the numerous functions \texttt{Mathematica} makes available for doing calculus. One important additional function is \texttt{Limit}. You should make it a point to read about these functions and any associated options in the \texttt{Mathematica} manuals.

\section{8.8.8 \ldots for Laplace Transforms}

In Section 1.5.2, we defined the Laplace transform $\tilde{f}(s)$ of a function $f(t)$ by the integral

$$
\tilde{f}(s) = \int_0^\infty e^{-st} f(t) \, dt \quad (8.1)
$$

and showed a number of the properties exhibited by this integral transform. Among other properties, the Laplace transform has the capacity to convert a linear, ordinary \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 (time-space) of the original problem. The initial conditions are automatically incorporated in this approach. Note, unfortunately, that inverting the Laplace transform of the solution to find the solution itself is rarely easy.

\texttt{Mathematica} provides for quick evaluation of an assortment of integral transforms. In particular the function \texttt{LaplaceTransform} will calculate a Laplace transform and the function \texttt{InverseLaplaceTransform} will invert a transform to return to the original space. We illustrate the use of those functions with the statements

\begin{verbatim}
In[1]:= Sin[\[Omega]*t];

In[2]:= LaplaceTransform[ %, t, s ]
Out[2]= \(\frac{\omega}{s^2 + \omega^2}\)

In[3]:= InverseLaplaceTransform[ %, s, t ]
Out[3]= Sin \[t \omega]\n
In[4]:= LaplaceTransform[ Exp[-a*t], t, s ]
Out[4]= \(\frac{1}{a + s}\)

In[5]:= Simplify[ InverseLaplaceTransform[ 1/( (s+a)*(s+b) ), s, t ] ]
Out[5]= \(-\frac{\text{e}^{-a t} - \text{e}^{-b t}}{a - b}\)

In[6]:= Quit[]
\end{verbatim}

Define a function. Note that the expression \texttt{\[Omega]} will ultimately print as $\omega$.

Calculate the Laplace transform of the function, where $t$ is the primary variable and $s$ is the variable in terms of which the transform is to be expressed.

Then, invert the transform to return to the original function.

Calculate the Laplace transform of the function $e^{-at}$.

Calculate and simplify the inverse Laplace transform of the function $(s+a)^{-1}(s+b)^{-1}$. 


8.8.9 ... for Ordinary Differential Equations

Among Mathematica’s strongest suits is its ability to solve ordinary differential equations (ODEs). We will return in Chapter 11 to a more extended discussion of those capabilities. For this chapter, we limit ourselves to a quick illustration, choosing the motion of an object of mass \( m \) attached to a spring having constant \( k \) and to a dashpot (shock absorber) having damping constant \( b \). The appropriate equation of motion is communicated to Mathematica with the statement

\[
\text{In[1]:= } \text{deq1 = } m x''[t] + b x'[t] + k x[t] == 0
\]

To solve this equation for \( x \) as a function of \( t \), we can proceed in either of two ways. If we submit the statement

\[
\text{In[2]:= soln1 = DSolve[ deq1, x[t], t ]}
\]

Mathematica returns the solution in the form

\[
\text{Out[2]= } \left\{ x[t] \to e^{(-b - \sqrt{b^2 - 4km}) t/2m} C[1] + e^{(-b + \sqrt{b^2 - 4km}) t/2m} C[2] \right\}
\]

Here, \text{DSolve} has used the symbols \( C[1] \) and \( C[2] \) for the initially undetermined integration constants in the solution to a second-order differential equation. Alternatively, we might phrase the invocation of \text{DSolve} in the form

\[
\text{In[3]:= soln2 = DSolve[ deq1, x, t ]}
\]

in which case Mathematica returns the solution in the form

\[
\text{Out[3]= } \left\{ x \to \text{Function}\left[t, e^{(-b - \sqrt{b^2 - 4km}) t/2m} C[1] + e^{(-b + \sqrt{b^2 - 4km}) t/2m} C[2] \right]\right\}
\]

where again the symbols \( C[1] \) and \( C[2] \) have been used for the integration constants.

The two forms of the solution look similar. Each is expressed as a transformation rule. The solution \( \text{soln1} \), however, is not a function. The rule expressing \( \text{soln1} \) allows substitution of the solution for \( x[t] \) in any expression. For example, the statement

\[
\text{In[4]:= } x[t] + \text{D}[ x[t], t ] /\text{. soln1}
\]

yields a result in which \( x[t] \) has been replaced in the first term but the derivative has been left unevaluated. At times, this first form of the solution will be adequate. If, however, we expect to manipulate with the solution, the second form is preferable, since it expresses the solution as a true function of the independent variable \( t \). Had we used \( \text{soln2} \) in the example at \text{In[4]}, i.e., had we submitted the statement

\[
\text{In[5]:= } x[t] + \text{D}[ x[t], t ] /\text{. soln2}
\]

\footnote{When the variable with respect to which a derivative is to be evaluated is the only quantity on which a function depends, we can use a prime \( ' \) to convey differentiation rather than the more explicit form \( \text{D}[ x[t], t ] \).}

\footnote{As is often the case, we here present the output in a form that is more transparent than—i.e., different from—the form presented directly in the \text{Mathematica} output. In what follows, if we extract explicit parts of the expressions, we will do so in ways that do not depend on these differences.}
Mathematica would have returned a more complicated result in which the solution would have been substituted both in the first and in the second term and then the derivative would have been evaluated.

Whichever form of the solution we have requested, Mathematica’s response to the simple statements we have used has been blindly to write a solution that is mathematically correct but has not recognized that the most convenient form of this solution will depend on the algebraic sign of the quantity under the square root in the two exponents. There are three cases:

1. \( b^2 > 4km \) (overdamped), in which case the square roots are real and both terms have real—though in both cases decaying—exponentials.

2. \( b^2 = 4km \) (critically damped), in which case the two terms are the same and Mathematica has returned only one solution when the general solution should be a linear superposition of two linearly independent solutions. The returned solution is mathematically correct, but we must carefully evaluate the limit as \( b^2 \) approaches \( 4km \) to extract a more transparent form.

3. \( b^2 < 4km \) (underdamped), in which case the square roots are imaginary and the complex exponentials that then appear would more appropriately be written in terms of sines and cosines of real arguments, each multiplied by the decaying exponential \( e^{-bt/2m} \).

To achieve a more appropriate form in the case of underdamped motion, we need to tell Mathematica that \( b^2 < 4km \), an end that is (unfortunately) not easily accomplished, partly because Mathematica assumes that variables to which no value has been assigned represent complex quantities and does not have a facility for overriding that assumption except within the context of a few specific commands. The following sequence of statements will achieve the desired objective.\(^{30}\) We begin by using the statement

\[ \text{In}[6] := \text{soln3 = soln1[[1]][[1]][[2]]} \]

\[ \text{Out}[6] = e^{(-b - \sqrt{b^2 - 4km})t/2m} C[1] + e^{(-b + \sqrt{b^2 - 4km})t/2m} C[2] \]

to extract the second part of the first part of the first part of \( \text{soln1} \), i.e., to extract an expression giving the solution itself. Then, we simplify the appearance of this expression by introducing (temporarily) the variables

\[ \gamma = \frac{b}{2m} \quad \text{and} \quad \omega = \sqrt{\frac{k}{m}} \] (8.2)

with the statement

\[ \text{In}[7] := \text{soln4 = soln3} /. \{ b -> 2*\text{\[Gamma]}*m, k -> \text{\[Omega]}^2*m \} \]

\[ \text{Out}[7] = e^{(-2m\gamma - \sqrt{4m^2\gamma^2 - 4m^2\omega^2})t/2m} C[1] + e^{(-2m\gamma + \sqrt{4m^2\gamma^2 - 4m^2\omega^2})t/2m} C[2] \]

Next, we invoke the statement\(^{31}\)

\[ \text{In}[8] := \text{soln5 = Simplify[ soln4, m > 0 ]} \]

\[ \text{Out}[8] = e^{(-t(\gamma + \sqrt{\gamma^2 - \omega^2}))} \left( C[1] + e^{(2t\sqrt{\gamma^2 - \omega^2})} C[2] \right) \]

to clear the \( m \)'s, the statement

\[ \text{In}[9] := \text{soln6 = Expand[ soln5 ]} \]

\[ \text{Out}[9] = e^{(-t(\gamma + \sqrt{\gamma^2 - \omega^2}))} C[1] + e^{(2t\sqrt{\gamma^2 - \omega^2}) - t(\gamma + \sqrt{\gamma^2 - \omega^2})} C[2] \]

\(^{30}\)The author continues to search for less cumbersome routes but, so far anyway, has failed to find any.

\(^{31}\)Some versions of Mathematica prior to Version 4.1 did not admit the second argument in the command Simplify.
8.8. USING MATHEMATICA ...

... to distribute the initial factor over both terms, and the command Map in the statement

\[ \text{In[10]:= soln7 = Map[Expand, soln6, 3]} \]

\[ \text{Out[10]= } e^{-t\gamma - t\sqrt{\gamma^2 - \omega^2}} C[1] + e^{-t\gamma + t\sqrt{\gamma^2 - \omega^2}} C[2] \]

... to apply Expand to some of the internal pieces of the expression in soln6. Finally, we simplify the exponents further by recognizing that the square roots are, in fact imaginary (when \(4km > b^2\), i.e., \(\omega > \gamma\)), a recognition achieved by making the substitution

\[ \text{In[11]:= soln8 = soln7 /. Sqrt[\[Gamma]^2-\[Omega]^2] -> I*a} \]


... At last, we are closing in on a more convenient presentation of the solution for the underdamped motion of a harmonic oscillation. In essence, we wish to replace the exponentials \(e^{\pm \text{i} at}\) with \(\cos at \pm \text{i} \sin at\). We take the next step with the statement

\[ \text{In[12]:= soln9 = Factor[Map[ComplexExpand, soln8 }} \]

\[ \text{Out[12]= } e^{-\text{i}\alpha t - \text{i} \gamma t} C[1] + e^{\text{i} \alpha t - \text{i} \gamma t} C[2] \]

From this solution, however, we note that the expression would have a yet simpler form if we substituted \(A\) for \(C[1] + C[2]\) and \(B\) for \(-i(C[1] - C[2])\). We achieve that objective with the statements


\[ \text{Out[13]= } \left\{ \left\{ C[1] \rightarrow \frac{1}{2}(A + iB), \ C[2] \rightarrow \frac{1}{2}(A - iB) \right\} \right\} \]

\[ \text{In[14]:= soln10 = Simplify[ soln9 /. %[[1]]} \]

\[ \text{Out[14]= } e^{-\text{i}\alpha t}(A \text{Cos}[at] + B \text{Sin}[at]) \]

... We, of course, are not finished until we have imposed suitable initial conditions on the solution. To do so, we determine \(A\) and \(B\) to reflect the simple assumption that the motion is started by drawing the oscillator aside and releasing it from rest, i.e., that

\[ x(0) = x_0 \quad \text{ and } \quad \frac{dx}{dt}(0) = 0 \]

(8.3)

Then, we find the constraints imposed on the integration constants with the statements

\[ \text{In[15]:= eq1 = x0 == soln10 /. t -> 0} \]

\[ \text{Out[15]= } x_0 == A \]

\[ \text{In[16]:= eq2 = 0 == D[ soln10, t ] /. t -> 0} \]

\[ \text{Out[16]= } 0 = aB - A\gamma \]

\[ \text{In[17]:= Solve[\{eq1,eq2\}, \{A,B\}} \]

\[ \text{Out[17]= } \left\{ \left\{ B \rightarrow \frac{x_0\gamma}{a}, A \rightarrow x_0 \right\} \right\} \]

\[ \text{In[18]:= soln11 = soln10 /. %[[1]]} \]

\[ \text{Out[18]= } e^{-\text{i}\alpha t}\left(x_0 \text{Cos}[at] + \frac{x_0\gamma \text{Sin}[at]}{a}\right) \]

\[ \text{In[19]:= soln12 = Collect[ soln11, x0 }} \]

\[ \text{Out[19]= } e^{-\text{i}\alpha t}x_0\left(\text{Cos}[at] + \frac{\gamma \text{Sin}[at]}{a}\right) \]
To interpret this result in terms of the original dimensional parameters, we need merely remember that:

\[ a = \sqrt{\omega^2 - \gamma^2} \quad ; \quad \gamma = \frac{b}{2m} \quad ; \quad \omega^2 = \frac{k}{m} \]  \hspace{1cm} (8.4)

These values could, of course, be reinserted into the solution obtained at line Out[19]. At long last, we have reached a form of the solution that reveals its character as an oscillation with an exponential decay of its amplitude. We leave it to the reader to use Mathematica to verify that this solution in fact satisfies the original equation and the imposed initial conditions. (Similar simpler expressions for the critically and overdamped cases are deduced in the exercises.)

In the above example, we chose to leave the initial conditions unspecified until after the solution had been obtained and recast in a more transparent form. Alternatively, we could have incorporated the initial conditions in the original invocation of DSolve with an expanded statement of the form

\[
\text{In[20]:= } \text{soln13 = DSolve}\{\text{deq1, x[0]==x0, x'[0]==0 }, x[t], t \}\]

The resulting solution would still have been presented in an exponential form in which, for the underdamped case, the square roots appearing would have been imaginary. We would still have to work our way through a number of transformations before having the solution in the standard form for the underdamped case. Note specifically that the equations and initial conditions have here been presented to DSolve as a list.

Much of our manipulation in the above illustration would have been simplified if we had expressed the original equation in dimensionless form before seeking its solution. To that end, we submit the statements

\[
\text{In[21]:= } \text{Quit[]}
\]

\[
\text{In[1]:= } \text{deq1 = m\times''[t] + b\times'[t]+k\times[t]==0}
\]

\[
\text{Out[1]= } kx[t] + bx'[t] + mx''[t] == 0
\]

to clear the workspace and redefine the equation. Then, with the statements

\[
\text{In[2]:= } \text{deq2 = deq1 /. \{k -> m*\[Omega]^2, b -> 2*m*\[Gamma]\}}
\]

\[
\text{Out[2]= } \omega^2 x[t] + 2\gamma x'[t] + \omega x''[t] == 0
\]

\[
\text{In[3]:= } \text{deq3 = Cancel[ deq2[[1]]/m ] == 0}
\]

\[
\text{Out[3]= } \omega^2 x[t] + 2\gamma x'[t] + x''[t] == 0
\]

we introduce the quantities \( \omega \) and \( \gamma \) defined by \( \omega^2 = k/m \) and \( \gamma = b/(2m) \), where \( \omega \) is the natural frequency of the oscillator, and remove the parameter \( m \) from the picture altogether. Next, we change the independent variable to \( \tau = \omega t \). One successful process for achieving that end involves (1) introducing a new function with the statement

\[
\text{In[4]:= } x[t_] := y[\[Omega]*t]
\]

(2) reevaluating \( \text{deq3} \) with the statement

\[
\text{In[5]:= } \text{deq4 = deq3}
\]

\[
\text{Out[5]= } \omega^2 y[t \omega] + 2\gamma y'[t \omega] + \omega^2 y''[t \omega] == 0
\]

(3) replacing the quantity \( t \omega \) (which is now the variable with respect to which differentiation is to be evaluated in \( \text{deq4} \)) with a single variable \( \tau \) by executing the statement
8.8. USING MATHEMATICA...

In[6]:= deq5 = deq4 /. t*\[Omega] -> \[Tau]

(4) introducing a new dimensionless variable β defined by γ = ωβ with the statement

In[7]:= deq6 = deq5 /. \[Gamma] -> \[Omega]*\[Beta]
Out[7]= ω^2 y[\tau] + 2\beta ω^2 y'[\tau] + ω^2 y''[\tau] == 0

and, finally, (5) removing the common factor of ω^2 with the statement

In[8]:= deq7 = Expand[ deq6[[1]]/\[Omega]^2 ] == 0
Out[8]= y[\tau] + 2\beta y'[\tau] + y''[\tau] == 0

At this point, we have swallowed all of the dimensional constants either into a rescaling of the time (τ = ωt) or into a single dimensionless parameter (β = b/mω). We could, of course, rescale what was originally x(t) and is now y(τ) to some chosen dimensionless quantity, but the equation is linear in this quantity and we can simply interpret the equation as it stands in those terms. We would then solve the equation as before with the statement

In[9]:= soln = DSolve[ {deq7, y[0]==x0, y'[0]==0}, y[\[Tau]], \[Tau] ]
Out[9]= {{y[\tau] -> 1/2 √-1 + β^2 + e(-\beta - \sqrt{-1 + \beta^2})τ + e(-\beta + \sqrt{-1 + \beta^2})τ + e(-\beta - \sqrt{-1 + \beta^2})τ \sqrt{-1 + \beta^2} + e(-\beta + \sqrt{-1 + \beta^2})τ \sqrt{-1 + \beta^2})}}

We leave it to the reader to use *Mathematica* to show that this solution is equivalent to the result shown at line Out[19] when β < 1.32

In[10]:= Quit[]

We have, of course, mentioned only the command *DSolve* as a function for solving ordinary differential equations. We have also seen additional contexts in which replacement of variables with the operator ‘/.’ and the commands *Expand, Factor, Solve, ComplexExpand,* and *Simplify* are useful. Finally, we have met the commands *Map, Collect,* and *Cancel* for the first time. You should make it a point to read about these commands and any associated options in the *Mathematica* manuals.

8.8.10 ... for Vector Calculus

*Mathematica* includes several commands to facilitate vector calculus.33 The main commands are *Grad, Div, Curl,* and *Laplacian* for the gradient, divergence, curl, and Laplacian, respectively. Each accepts a first argument—a scalar function for *Grad* and *Laplacian*; a vector function for *Div* and *Curl*—and a second argument, a list that identifies the variables in the coordinate system in which the quantity is expressed. An optional third argument—a string—specifies the coordinate system to be used, with "Cylindrical" and "Spherical" as the main alternatives to the default ("Cartesian").

32Here, the assumption β < 1 is equivalent to the assumption γ/ω < 1 or γ^2 < ω^2 or b^2/4km^2 < k/m or b^2 < 4mk. Thus, the dimensionless inequality β < 1 here is exactly equivalent to the dimensional inequality b^2 < 4km that appeared earlier in this section.

33Prior to Version 9, vector calculus was provided in *Mathematica* by the add-on package *VectorAnalysis,* which needed to be added with the statement << *VectorAnalysis*'. In subsequent versions, this capacity was made part of the main kernel, though with rather different commands and syntax. The previous package may remain available for a while, but it is officially obsolete and is likely to disappear altogether at some point.
We begin this illustration by defining scalar and vector quantities in each of the three common coordinate systems with the statements

\begin{align*}
\text{In}[1] &: \quad \text{scar} = x^2 + y^2 + z^2 \\
\text{Out}[1] &:= x^2 + y^2 + z^2 \\
\text{In}[2] &: \quad \text{scyl} = r^2 + z^2 \\
\text{Out}[2] &:= r^2 + z^2 \\
\text{In}[3] &: \quad \text{ssph} = r^2 \\
\text{Out}[3] &:= r^2 \\
\text{In}[4] &: \quad \text{vcar} = \{x, y, z\} \\
\text{Out}[4] &:= \{x, y, z\} \\
\text{In}[5] &: \quad \text{vcyl} = \{r, 0, z\} \\
\text{Out}[5] &:= \{r, 0, z\} \\
\text{In}[6] &: \quad \text{vsph} = \{r, 0, 0\} \\
\text{Out}[6] &:= \{r, 0, 0\}
\end{align*}

Define a scalar in Cartesian coordinates.

Define a vector in Cartesian coordinates.

Define a scalar in cylindrical coordinates.

Define a vector in cylindrical coordinates.

Define a scalar in spherical coordinates.

Define a vector in spherical coordinates.

Then, we evaluate the gradients with the statements,

\begin{align*}
\text{In}[7] &: \quad \text{Grad}[\text{scar}, \{x, y, z\}] \\
\text{Out}[7] &:= \{2x, 2y, 2z\} \\
\text{In}[8] &: \quad \text{Grad}[\text{scyl}, \{r, \phi, z\}, \text{"Cylindrical"}] \\
\text{Out}[8] &:= \{2r, 0, 2z\} \\
\text{In}[9] &: \quad \text{Grad}[\text{ssph}, \{r, \theta, \phi\}, \text{"Spherical"}] \\
\text{Out}[9] &:= \{2r, 0, 0\}
\end{align*}

Evaluate $\nabla$ in Cartesian coordinates.

Evaluate $\nabla$ in cylindrical coordinates.

Evaluate $\nabla$ in spherical coordinates.

the Laplacian with the statements

\begin{align*}
\text{In}[10] &: \quad \text{Laplacian}[\text{scar}, \{x, y, z\}] \\
\text{Out}[10] &:= 6 \\
\text{In}[11] &: \quad \text{Laplacian}[\text{scyl}, \{r, \phi, z\}, \text{"Cylindrical"}] \\
\text{Out}[11] &:= 6 \\
\text{In}[12] &: \quad \text{Laplacian}[\text{ssph}, \{r, \theta, \phi\}, \text{"Spherical"}] \\
\text{Out}[12] &:= 6
\end{align*}

Evaluate $\nabla^2$ in Cartesian coordinates.

Evaluate $\nabla^2$ in cylindrical coordinates.

Evaluate $\nabla^2$ in spherical coordinates.

the divergence with the statements

\begin{align*}
\text{In}[13] &: \quad \text{Div}[\text{vcar}, \{x, y, z\}] \\
\text{Out}[13] &:= 3 \\
\text{In}[14] &: \quad \text{Div}[\text{vcyl}, \{r, \phi, z\}, \text{"Cylindrical"}] \\
\text{Out}[14] &:= 3 \\
\text{In}[15] &: \quad \text{Div}[\text{vsph}, \{r, \theta, \phi\}, \text{"Spherical"}] \\
\text{Out}[15] &:= 3
\end{align*}

Evaluate $\nabla \cdot$ in Cartesian coordinates.

Evaluate $\nabla \cdot$ in cylindrical coordinates.

Evaluate $\nabla \cdot$ in spherical coordinates.

and, finally, the curl with the statements

\begin{align*}
\text{In}[16] &: \quad \text{Curl}[\text{vcar}, \{x, y, z\}] \\
\text{Out}[16] &:= \{0, 0, 0\} \\
\text{In}[17] &: \quad \text{Curl}[\text{vcyl}, \{r, \phi, z\}, \text{"Cylindrical"}] \\
\text{Out}[17] &:= \{0, 0, 0\} \\
\text{In}[18] &: \quad \text{Curl}[\text{vsph}, \{r, \theta, \phi\}, \text{"Spherical"}] \\
\text{Out}[18] &:= \{0, 0, 0\}
\end{align*}

Evaluate $\nabla \times$ in Cartesian coordinates.

Evaluate $\nabla \times$ in cylindrical coordinates.

Evaluate $\nabla \times$ in spherical coordinates.

\textit{Mathematica} can also evaluate these vector derivatives when the dependence in the first argument on the coordinates is not explicit. For example

\begin{align*}
\text{In}[19] &: \quad \text{Laplacian}[f[r, \phi, z], \{r, \phi, z\}, \text{"Cylindrical"}] \\
\text{Out}[19] &:= f^{(0, 0, 2)}[r, \phi, z] + \frac{f^{(0, 2, 0)}[r, \phi, z]}{r} + \frac{f^{(1, 0, 0)}[r, \phi, z]}{r} + f^{(2, 0, 0)}[r, \phi, z]
\end{align*}
Here, the superscripts on \( f \) indicate how many derivatives with respect to each of the three variables are to be evaluated, e.g.,

\[
f^{(2,0,0)}[r, \phi, z] = \frac{\partial^2 f}{\partial r^2} \quad ; \quad f^{(1,1,0)}[r, \phi, z] = \frac{\partial^2 f}{\partial r \partial \phi}
\]

where the independence of the mixed derivatives on the order in which the derivatives are evaluated is critical to use of this notation.

We have, of course, mentioned only the commands \texttt{Grad}, \texttt{Div}, \texttt{Laplacian}, and \texttt{Curl} for working with coordinate dependent vectors and scalars. \texttt{Mathematica} also includes a variety of commands to transform the coordinates of a point from one system to another, to calculate the area or volume of appropriately defined regions, to create two and three dimensional graphical displays of vector fields, and many more. You should make it a point to bring up Wolfram Documentation from the NBI \texttt{Help} menu, search for ‘Vector Analysis’ in that documentation, and study the multitude of available commands. Be aware, also, that \texttt{Mathematica} can apply these commands to functions and vectors in more than three dimensions.

### 8.9 Packages

To this point, we have supposed that the \texttt{Mathematica} we are using has been configured in the way set up by the default installation, so that all of the functions so far invoked will be available to us once \texttt{Mathematica} has been launched. At some sites, however, \texttt{Mathematica} may have been configured to conserve memory, and some of the features normally loaded automatically may have been left out. At other sites, additional features giving further commands may have been included when \texttt{Mathematica} is started. This flexibility is provided through the use of \texttt{Mathematica} packages, segments of \texttt{Mathematica} code which may or may not be automatically loaded when \texttt{Mathematica} is started up and which define features that will be present only when the package defining those features has been loaded. Any package not loaded automatically can, of course, be explicitly loaded to make its features available. A list of the packages included with the standard \texttt{Mathematica} distribution can be found by accessing the Wolfram Documentation either from the \texttt{Help} menu in the NBI or on line, searching for ‘Packages’, and opening the link labeled something like ‘Wolfram System Standard Extra Packages’. The resulting page has links to descriptions of the packages. A list of the packages that have been loaded into the current \texttt{Mathematica} configuration is contained in the special variable $\$\texttt{Packages}$, whose value can be displayed by the simple statement\(^{34}\)

\[
\text{In}[9] := \$\texttt{Packages}
\]

\[
\text{Out}[9] = \text{ListOfLoadedPackages}
\]

If a needed package is not loaded automatically when \texttt{Mathematica} is launched, it can be explicitly loaded with one of the commands\(^ {35}\)

\[
\text{In}[23] := <<\texttt{PackageName} \\
\text{In}[38] := \texttt{Get[ "PackageName"] } \\
\text{In}[79] := \texttt{Needs[ "PackageName"]}
\]

Here << and \texttt{Get} are synonymous and will load the package directly, even if it has previously been loaded. In contrast, \texttt{Needs} will load the package only if it has not already been loaded.

Beyond the standard packages, individual users can create additional packages for special purposes. Thus, if we are lucky, we might be able to find a user-contributed package with special features

\(^{34}\)The list output by this statement may vary from site to site, depending on how \texttt{Mathematica} is configured. It will also reflect the inclusion of any packages explicitly loaded during the session.

\(^{35}\)Note that the tic following \texttt{PackageName} in the first statement is a \textit{forward} tic.
of value to a particular exploration. In the absence of such luck, we could also create our own pack-
package to serve those purposes. The details of writing such packages are laid out in the *Mathematica*
manuals but are beyond the scope of this book.

With each new version of *Mathematica*, some features that had been provided with add-on
packages may be moved to the main kernel, in which case those features will be available without
explicitly loading the package. Indeed, prior to Version 9, the capacity for evaluating several vector
derivatives described in Section 8.8.10 was available only if a package named *VectorAnalysis* was
loaded explicitly. Starting with Version 9, those features were built into the main kernel, though the
names of the commands and their syntax suffered revision. For backward compatibility, the original
package is retained in subsequent versions, at least for a while, but it may in time be completely
removed. Further, loading the package will also generate a message alerting the user that the package
has become obsolete.

### 8.10 Loops, Logical Expressions, and Conditionals

Among the most ubiquitous programming structures is the loop, which provides a means by which a
statement or block of statements can be executed some number of times, typically with small changes
controlled by a loop index. Several such structures are available in *Mathematica*. The simplest is
the *For* loop, which we have already met in Sections 8.8.2 and 8.8.5. Briefly, the statements

```mathematica
In[1]:= x = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0}
Out[1]= {0, 0, 0, 0, 0, 0, 0, 0, 0, 0}
In[2]:= For[i = 1, i <= 10, i = i + 2, x[[i]] = i^2]
In[3]:= x
Out[3]= {1, 9, 25, 49, 81, 0}
```

create a 10-element list of zeros and then assign the squares of the odd integers to the odd elements
of that list. Note that the C-like structure `i++` can be used instead of the expression `i = i + 1` if
the index is to be incremented by 1.

In essence, the *For* loop requires *a priori* knowledge of how many times the loop will be executed.
For cases when that knowledge is not in hand, *Mathematica* makes available the *While* loop, whose
general syntax has the form

```mathematica
While[(condition), (block of statements)]
```

In this loop, the block of instructions must assume responsibility for changing the condition to a
value that will stop the loop. More specifically, these loops are illustrated in the statements

```mathematica
In[4]:= x = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0};
In[5]:= i = 1;
In[6]:= While[i < 11, x[[i]] = i^3; i = i + 2];
In[7]:= x
Out[7]= {1, 9, 27, 125, 343, 0, 729, 0}
```

Note that,

- The block of statements in this structure starts after the first comma and continues until the
  closing bracket.
- There is *no* punctuation after the last statement in the block of statements.
- Individual statements in that block are separated by semicolons.
• The statement incrementing \( i \) is essential so that the initially true condition \( i < 11 \) is ultimately toggled to false and the loop terminates.

Finally, a word about composite conditions: The illustrations above made use of (single) conditions like \( i < 11 \), which will evaluate to true or false depending on the value of \( i \). Indeed, the Mathematica function \( \text{Boole} \), for example,

\[
\text{In[8]} := i := 5:
\text{In[9]} := \{ \text{Boole}[i > 10], \text{Boole}[i < 10] \}
\text{Out[9]} = [0, 1]
\]

allows the testing of a logical condition. In this Boolean context, 0 equates to false and 1 to true. The six logical operators \( ==, <, >, <=, >=, \) and \( != \) can be used for test for equal, less than, greater than, less than or equal, greater than, and unequal. In some situations, one may need to (1) combine two or more conditions into a composite condition or (2) negate a condition. Mathematica makes available the logical operators \( &&, | |, \) and \( ! \), and the logical functions \( \text{Xor} \) (exclusive or), \( \text{Nand} \), and \( \text{Nor} \) to facilitate combining conditions and negating conditions. More specifically, with \( (C_1) \) and \( (C_2) \) individual conditions

• \( ((C_1) \&\& (C_2)) \) will be true if \( (C_1) \) and \( (C_2) \) are both true and false otherwise,

• \( ((C_1) \mid\mid (C_2)) \) will be true if either \( (C_1) \) or \( (C_2) \) is true or both \( (C_1) \) and \( (C_2) \) are true and false otherwise, and

• \( \text{Xor}(C_1), (C_2) \) will be true if either \( (C_1) \) or \( (C_2) \) is true and false if neither condition is or both are true, and

• \( ! (C_1) \) will be true if \( (C_1) \) is false and false if \( (C_1) \) is true.

For example, the coding

\[
\text{In(10)} := x = \{ 0, 0, 0, 0, 0, 0, 0, 0, 0 \};
\text{In(11)} := i = 1;
\text{In[12]} := \text{While}[ (i > 0 \&\& i < 11), x[[i]] = i^3; i = i+1 ]
\text{In[13]} := x
\text{Out[13]} = \{1, 8, 27, 64, 125, 216, 343, 512, 729, 1000\}
\]

illustrates a composite condition controlling the While statement. Note that parentheses around the composite condition can in this case be omitted, but they—and perhaps parentheses around some internal pieces—may be necessary in some cases.

Logical conditions appear not only to control loops but also to structure branches in a sequence of statements. As with most programming languages, Mathematica also provides the function \( \text{If} \) for this purpose. Its fundamental syntax is

\[
\text{If}[ (\text{condition}), (\text{statements if true}), (\text{statements if false}) ]
\]

though the second block of statements can be omitted if nothing is to be done when the condition is false. Thus, for example, the statement

\[
\text{For[ } i=1, i<=10, i++, \text{If[ } x[[i]] < 0, x[[i]] = -x[[i]] \text{ ] ]}
\]

will replace each negative element in a list \( x \) with the corresponding positive value, the statement

\[
\text{For[ } i = 1, i<=10, i++, \text{If[ } x[[i]] > 10, x[[i]] = 10 \text{ ] ]}
\]
will replace all values greater than 10.0 with the value 10.0, and the statement

\[\text{If[ } a >= 0, b = a, b = -a \] \]

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 true clause or the false clause will be separated by commas and enclosed within braces, as in

\[\text{If[ } \langle \text{condition} \rangle, \{ \langle \text{statement} \rangle, \langle \text{statement} \rangle, \ldots \} \{ \langle \text{statement} \rangle, \langle \text{statement} \rangle, \ldots \} \] \]

For selection among more than two options, the statement would have to be nested with care. All conditions may be single or composite.

As is the case in most languages, loops and conditional structures can be nested, though constructing the syntax correctly can sometimes be challenging.

### 8.11 Command Files

In addition to responding to statements supplied interactively, \textit{Mathematica} can execute statements read from a command file. To begin, we create a text file containing the desired statements just as they would be typed interactively. If that file is then stored in the default directory\(^{36}\) or somewhere in \textit{Mathematica}'s search path,\(^{37}\) execution of the command(s) contained within it would then be specified by typing either of the statements

\[\text{\textless\textless FileName} \quad \text{or} \quad \text{Get["FileName"]}\]

when \textit{Mathematica} is ready to accept input. There is no default file type.\(^{38}\) As the command file is read and executed, the statements themselves will \textit{not} be displayed on the screen. Whether the statements are terminated with semicolons or not, the output will also not be displayed on the screen (except for the last statement in the file, whose output will be displayed if it ends with no semicolon and suppressed if it ends with a semicolon). Further, no line labels will be attached to any of the intermediate results, and the output from the last statement in the file (if any) will be given the label Out\([n]\), where \( n \) is the identifying number of the statement requesting the reading of the file. Only if intermediate results are explicitly printed (with the command \texttt{Print[]} will these results be displayed; only if intermediate results are explicitly bound to a variable will those results be available once execution of the command file has been completed. Except for the absence of the In and Out labels, the \textit{Mathematica} environment after the command file has been executed is identical to the environment that would have existed had the statements been entered interactively.

Sometimes, creation of a command file facilitates repeated use of quantities not defined in standard \textit{Mathematica}. Use of such a file can also facilitate the debugging of a sequence of statements, since—when an error is reported—we need only edit the file appropriately and resubmit it to \textit{Mathematica}; repeated typing of the correct parts of the sequence is not then necessary. To illustrate matrix multiplication, for example, we could create the file \texttt{testmultiply.m} containing the statements listed in Table 8.1.\(^{39}\)

Then, assuming that this file has been stored in the default directory,\(^{40}\) we would execute the statements it contains by submitting to \textit{Mathematica} one or the other of the statements

---

36 See Section 8.18.3.
37 See Section 8.18.4.
38 Conventionally, text files containing \textit{Mathematica} statements are given names with file type \texttt{.m}. Further, the types \texttt{.nb} and \texttt{.mx} are conventionally used for files containing \textit{Mathematica} notebooks and \textit{Mathematica} definitions in a special format, respectively.
39 Comments in \textit{Mathematica} code are introduced with the symbol \(\ast\) and terminated with the symbol \(\ast\).
40 See again Section 8.18.3.
Table 8.1: The Mathematica command file testmultiply.m.

(* Command file testmultiply.m *)

(* Enter and display matrix *)
mat1 = { {1,2}, {3,4} }
Print[ MatrixForm[ mat1 ] ]
(* Evaluate and display transpose *)
mat2 = Transpose[ mat1 ]
Print[ MatrixForm[ mat2 ] ]
(* Evaluate and display matrix product *)
Print[ MatrixForm[ mat1.mat2 ] ]
(* Evaluate and display element-by-element product *)
Print[ MatrixForm[ mat1*mat2 ] ]

Table 8.2: The Mathematica command file crossdot.m.

(* crossdot.m *)

(* Command file to define functions to evaluate the cross and dot products of two three-component vectors. *)
lucross[a_,b_] :=
{ a[[2]]*b[[3]] - a[[3]]*b[[2]],
  a[[3]]*b[[1]] - a[[1]]*b[[3]],
  a[[1]]*b[[2]] - a[[2]]*b[[1]] };
ludot[a_,b_] := a[[1]]*b[[1]] + a[[2]]*b[[2]] + a[[3]]*b[[3]];

<< testmultiply.m or Get[ "testmultiply.m" ]

All output produced by the Print statements will be displayed on the screen but only the expressions explicitly bound to mat1 and mat2 will be available by those names once execution of the file is completed.

As a second example, suppose that we frequently had two three-component lists defining two three-dimensional vectors and that we wanted to be able easily to evaluate their cross and dot products. We might create a command file listed in Table 8.2. Here, beyond the comments, the first four lines—actually a single statement—define the function lucross, which takes two three-component lists as arguments and returns a three-component list whose components are those of the cross product of the two vectors. The last line defines the function ldot, which takes the same arguments but returns a scalar value equal to the dot product of those two lists treated as three-dimensional vectors. This file is named crossdot.m and can be accessed in the directory $HEAD/mathematica. Thus, for example, the statement

41See the note at the beginning of this chapter.
Get[ "$HEAD/mathematica/crossdot.m" ];

will load this file from the directory associated with this book, defining the two functions \texttt{ludot} and \texttt{lucross} for subsequent use.\footnote{Be aware, as well, that the dot product of two vectors can also be evaluated with the non-commutative infix operator \texttt{.}, which will actually provide the sum of the products of corresponding components regardless of the dimensionality of the vectors. Further, in addition to the several vector derivatives already introduced in Section 8.8.10, \texttt{Mathematica} includes the commands \texttt{Dot} and \texttt{Cross}, which are more carefully structured than \texttt{ludot} and \texttt{lucross} to evaluate dot and cross products.}

Finally in this section, we note that \texttt{Mathematica} also provides means by which we can display the contents of a command file conveniently and easily from within \texttt{Mathematica}. To display the file \texttt{crossdot.m}, we submit to \texttt{Mathematica} the statement

\begin{verbatim}
FilePrint[ "crossdot.m" ] or FilePrint[ "$HEAD/mathematica/crossdot.m" ]
\end{verbatim}

Here, the first form applies if the file is in a directory in the search path and the second form refers to the file in the directory associated with this book. The contents of the file will be displayed in the NBI or, if you are using the TBI, in the \textit{Shell} window from which \texttt{Mathematica} was launched.

\section*{8.12 High-Resolution Graphing}

We illustrate the production of high resolution graphics by plotting the solution for the damped harmonic oscillator obtained in Section 8.8.9. First, we reclaim the solution with the statement

\begin{verbatim}
In[1]:= soln3 = x0 * Exp[-\[Gamma]*t] * (\[Gamma]*Sin[a*t]/a + Cos[a*t])
\end{verbatim}

\begin{verbatim}
Out[1]= e^{-t\gamma}x_0\left(\cos(at) + \frac{\gamma\sin(at)}{a}\right)
\end{verbatim}

In addition, we calculate and display the velocity with the statement

\begin{verbatim}
In[2]:= soln3dot = Simplify[ D[ soln3, t ] ]
\end{verbatim}

\begin{verbatim}
Out[2]= -e^{-t\gamma}x_0(a^2 + \gamma^2)\sin(at)/a
\end{verbatim}

Then, recognizing that $\gamma$ and $a$ are not independent ($a^2 = k/m - \gamma^2$) and hence that assigning values to $\gamma$ and $a$ in effect fixes $k/m$, we use the statements

\begin{verbatim}
In[3] := \[Gamma] = 0.1; a = 1.0; x0 = 1.0;
\end{verbatim}

to assign specific values to $\gamma$, $a$, and $x_0$. Finally, we create an inline graph showing position and velocity with the statement

\begin{verbatim}
In[4]:= Plot[ {soln3,soln3dot}, {t,0.0,20.0} ]
\end{verbatim}

which plots the identified expressions over the intervals $0.0 \leq t \leq 20.0$ in $t$ and $-1.0 \leq y \leq 1.0$ in $y$ (the vertical coordinate). The graph is produced as a coarse character plot if we are working in the \texttt{Mathematica} TBI and as a smooth plot in the NBI.\footnote{The number of points distributed over the range of the independent variable is determined in \texttt{Mathematica} to keep the amount of turning from segment to segment small. Numerous options—see the \texttt{Mathematica} manuals—give the user the ability to override the defaults.}

Alternatively, we could modify the resulting graph by exploiting a few of the many options supported by the command \texttt{Plot}. The statement
Figure 8.1: Graph of position and velocity for a damped oscillator with $\gamma = 0.1$, $a = 1.0$, $x(0) = x_0 = 1.0$, and $v(0) = 0.0$. Here, the solid and dashed lines show $x(t)$ and $v(t)$, respectively.

Damped Oscillator

For example, will produce the same graph but with

- the vertical axis running from $-1.0$ to $+1.0$,  
- at least 100 points spaced over the range of the independent variable (the default is 25),  
- the indicated labels on the horizontal and vertical axes,  
- the indicated title in 16-point, bold type, Times font on the graph as a whole,  
- the first expression (but not the axes) displayed black with a line of thickness heavier than the default (i.e., with a thickness set to the indicated fraction of the height of the graph as a whole) and the default (solid) linestyle and the second expression displayed black with the same line thickness but with a dashed line style with alternate dashes and gaps, each of size 0.02 units, and  
- the text $\gamma = 0.1$, $a = 0.1$ displayed on the face of the graph and centered at the point [15.0, 0.75].

These (and other) options to the command Plot are fully described in the Mathematica manuals. The graph resulting from the statement on line In[5] is shown in Fig. 8.1.

We could explore other features of this motion by, for example, changing the value of $\gamma$ and reploting the graph with the statements

$\text{In[5]} := \text{Plot} \{ \{\text{soln3, soln3dot}, \{t,0.0,20.0\}, \text{PlotRange}\to\{-1.0,1.0\}, \text{PlotPoints}\to100, \text{AxesLabel}\to\{"t","x,v"\}, \text{PlotLabel}\to\text{StyleForm["Damped Oscillator",}  
\text{FontSize}\to16, \text{FontFamily}\to\"Times\", \text{FontWeight}\to\"Bold\"\}, \text{PlotStyle}\to\{\{\text{Thickness}[0.01],\text{Black}\},  
\{\text{Thickness}[0.01],\text{Black},\text{Dashing}[\{0.02,0.02\}]\}\}, \text{Epilog}\to\{\text{Text[\"\[Gamma]=0.1, a=0.1", \{15.0, 0.75\}]}\}\]$

44See the Mathematica manuals at Epilog and Text for details on the way to specify text displayed on a graph.
CHAPTER 8. INTRODUCTION TO MATHEMATICA

Figure 8.2: Graph of velocity versus position for a damped oscillator with $\gamma = 0.2$, $a = 1.0$, $x(0) = 1.0$, and $v(0) = 0.0$.

In[6]:= \[Gamma\] = 0.2;
In[7]:= Plot[ {soln3, soln3dot}, {t,0.0,20.0}, PlotRange->{-1.0,1.0},
PlotPoints->100, AxesLabel->"t","x,v"],
PlotLabel->StyleForm[ "Damped Oscillator",
FontSize->16, FontFamily->"Times", FontWeight->"Bold" ],
PlotStyle->{ {Thickness[0.01],Black},
{Thickness[0.01],Black,Dashing[{0.02,0.02}]} } ]

or, for these new values of the parameters, we could produce a graph of the trajectory in the phase plane ($soln3dot$ versus $soln3$) with the statements

In[8]:= ParametricPlot[ {soln3, soln3dot}, {t,0.0,20.0},
PlotRange->{{-1.0, 1.0}, {-1.0, 1.0}},
PlotStyle->{Thickness[0.01],Black},AspectRatio->1.0 ]

Here, the new option AspectRatio permits us to tell Mathematica to use a plotting area that is square (aspect ratio 1.0). The resulting graph is shown in Fig. 8.2.

Mathematica can, of course, also produce surface plots and contour plots in several different ways. The simplest and quickest such plots are produced with statements like

In[9]:= mode23 = Sin[ 2*Pi*x ] * Sin[ 3*Pi*y ]
Out[9]= Sin[2\pi x] Sin[3\pi y]
In[10]:= Plot3D[ mode23, {x,0.0,1.0}, {y,0.0,1.0},
PlotPoints->25, ColorFunction->GrayLevel ]
In[11]:= lvls = Range[-9,9,1]/10.0;
In[12]:= ContourPlot[ mode23, {x, 0.0, 1.0}, {y, 0.0, 1.0},
AspectRatio->1.0, ContourShading->False,
Contours->lvls, PlotPoints->50 ]
In[13]:= Quit[]
Line In[9] here defines the expression to be displayed. Line In[10] produces a surface plot with 25 points (default 15) along each axis and coloring set to a gray scale (since we anticipate black and white printing in this book). Finally, line In[11] defines the values at which contour lines are too be drawn (from $-0.9$ to $0.9$ in steps of $0.1$) and line In[12] produces a contour map at those levels (option Contours) with shading turned off, the aspect ratio adjusted so that the display will be square, and the function evaluated at 50 points along each axis. The resulting graphs are shown in Figs. 8.3 and 8.4.

We have, of course, only mentioned a few (Plot, ParametricPlot, Plot3D, and ContourPlot) of the functions and only some (PlotRange, PlotPoints, AxesLabel, PlotLabel, PlotStyle, FontSize, FontWeight, ColorFunction, AspectRatio, ContourShading, and Contours) of the options that modify the behavior of these functions. Several other functions for producing graphical displays and numerous additional options are available. You should make it a point to read about these functions and the associated options in the Mathematica manuals. In particular, note that the statement Options[Function]\(^{45}\) will return a list of all of the options that affect Function, attaching to each its current default value.

8.13 Making Hard Copy ...

The method used to make hard copy of all or part of a Mathematica dialog depends on whether the TBI or the NBI is in use; whether the material to be output is textual or graphical; whether the graphical material is inline or in a separate window; whether the output is to go directly to a printer or to a PostScript, ASCII, HTML, or ... file; and whether we are working on a UNIX, Windows, or Macintosh platform. In the following subsections we address each of several possibilities in turn.

\(^{45}\)See Section 8.18.5 for a fuller discussion.
Figure 8.4: Contour map of the function $z(x, y) = \sin(2\pi x) \sin(3\pi y)$.

8.13.1 ... from the NBI

The menu structure and procedures to print text and graphics from the NBI may differ from version to version and platform to platform. For obtaining hard copy in the NBI, several items in the FILE menu are relevant:

- The option ‘Print...’ opens a pop-up window in which to specify a printer and, if the OPTIONS or PREFERENCES menu is opened, select various parameters (page range to be printed, number of copies to be printed, treatment of color in the output, ...). In Windows, a capacity to print a selection from the active notebook is provided as an option in the ‘Page Range’ panel of the ‘Print...’ option in the FILE menu; in UNIX, a capacity to print a selection from the active notebook is provided in the separate option ‘Print Selection...’ in the FILE menu. In either case, the option will be available only if a portion of the notebook has, in fact, been selected.

- The option ‘Printing Settings→’ from the FILE menu will open a menu offering several options, including:
  - ‘Page Setup’, which opens a pop-up window in which to specify paper size, orientation (portrait or landscape), and margins.
  - ‘Printing Options’, which opens a pop-up window in which to specify whether to print cell brackets, registration marks, and a few other features.
  - ‘Headers and Footers’, which opens a pop-up window in which to specify the content and position of desired page headers and footers.
  - ‘Show Page Breaks’, a check box, to select whether page breaks should be displayed or hidden in the NBI.
  - ‘Printing Environment’, which opens a pop-up window in which to select among the options ‘Working’, ‘Slide Show’, or ‘Printout’.

---

46 This description has been worked out in LINUX and Windows environments. The same features will be available in other environments, but the details of the behavior may be slightly different. Further, the details here described apply specifically to *Mathematica* 11. Other versions may make these capabilities available on other ways.

47 ‘Print to file’ is among the options but offers fewer formats than ‘Save’, ‘Save As’, and ‘Save Selection As’.
8.13. MAKING HARD COPY...

- The option ‘Print Preview...’ displays on screen what will ultimately be printed. That on-screen display will be produced in an Adobe Reader window from which a PDF file can, of course, be saved.

- The options ‘Save...’, ‘Save As...’, and ‘Save Selection As...’, each of which brings up a browser in which you can identify the directory in which all or the selected part of a notebook should be saved, the name to give the resulting file, and the file type to be used. Available file types include tex, ps, eps, pdf, and html, though not all types are available with all options. In particular, the filetype ps is available with ‘Save...’ and ‘Save As...’ and the file type eps is available with ‘Save Selection As’. The resulting file can then be printed directly or incorporated in a more extensive document.

These routes to output information include the Mathematica-generated labels for all included elements of the notebook.

Mathematica also provides a facility for writing a description of a specified graphic object (not the entire notebook) into a file. We simply need to inform Mathematica of the form of the driver it is to use in creating the file, the name it is to give the file, and any controlling options. For example, with graph the label of the graphics display to be output to a file, either of the statements

Export[ "trial.eps", graph, "EPS" ]

which determines the format of the output file from the third argument, or

Export[ "trial.eps", graph ]

which determines the format of the output file from the extension in the file name,\(^48\) will tell Mathematica to convert the graphic object identified by the variable graph into encapsulated PostScript and write that description to the file trial.eps in the default directory.\(^49\) When displayed on a computer screen with ghostview, an EPS file creates a display that includes a number of controlling buttons surrounding the graph itself. These surrounding buttons are not included in PDF or JPEG files, nor do they appear, for example, when the EPS file is incorporated in a L\TeX\, document. This route to exporting a graph does not include Mathematica-generated output labels in the file.

Note, however, that in some contexts Mathematica may not always transcribe all characters in the text in the on-screen graphic (e.g., in labels on the axes or in titles) into the proper PostScript characters. Unless some Mathematica-specific fonts have been made available to the program that will interpret the PostScript file, opening and closing parentheses, the forward slash, and probably other characters may be rendered in unexpected forms when the resulting PostScript file is printed or viewed on the screen. In exporting graphics, we must either be sure that the appropriate Mathematica fonts are known by the program(s) that will display the file, or we must constrain Mathematica as it produces the graphic by limiting its choice of fonts for use in the graphic. The former route is complicated (and necessary if, for example, Greek letters are to be included in the exported text).\(^50\) While not always satisfactory, the latter route is far simpler; it involves the stipulation of an additional option in the command producing the graphic. For the sake of a quick illustration, we simply note that the statements

\begin{verbatim}
graph = Plot[ Sin[t], {t, 0, 10}, AxesLabel->"sin(t)", "x/a"],
    FormatType->OutputForm ]
Export[ "trial.eps", graph, "EPS" ]
\end{verbatim}

will produce an encapsulated PostScript file in which the parentheses and the slash in the labels on the axes will be properly rendered, even when the specific Mathematica fonts are not available to the program that will display trial.eps.

---

\(^48\) Other available formats include "PDF" (file type pdf and "JPEG" (file type jpeg). A full list is stored in the system variable $ExportFormats.

\(^49\) See Section 8.18.3.

\(^50\) For details, see the Mathematica manuals.
8.13.2 ... from the TBI

Printing of text displayed in the TBI is most simply accomplished by using the cutting and pasting capabilities of the operating system to transfer that character information into a text file, which can then be edited in a standard text editor and sent to an available printer with the standard commands to the operating system (or perhaps by selecting appropriate options from a menu available with the editor).

Graphics commands like Plot may not produce very satisfactory on-screen results when executed in the TBI. Even so, if it is given a variable name, a graphic object can still be written into a file as high-resolution graphic for subsequent printing. As in the previous section, we simply need to inform Mathematica of the form of the driver it is to use in creating the file, the name it is to give the file, and any controlling options. For example, either of the statements

\[
\text{Export["trial.eps", graph, "EPS"]}
\]

or

\[
\text{Export["trial.eps", graph ]}
\]

will tell Mathematica to convert the graphic object identified by the variable graph into encapsulated PostScript and write that description to the file trial.eps in the default directory.

8.14 Output in \LaTeX Format

A particularly useful feature of Mathematica is its ability to write \LaTeX versions of expressions to the screen or into a file. Full details are laid out in the Mathematica manuals. In brief, the statement\footnote{Note the upper-case X in the third character.}

\[
\text{TeXForm[ expr1 ]}
\]

for example, converts the specified expression into \LaTeX format and displays it on the screen (from where standard cut and paste operations—when they are available—can move the description to wherever we might want it). Alternatively, either of the statements

\[
\text{TeXForm[ expr1 ] > FileN}\]
\[
\text{Put[ TeXForm[ expr1 ], "FileName" ]}
\]

will create a new (ASCII) file and write the \LaTeX description of the indicated expression into that file, overwriting any existing file by the same name. To append the \LaTeX description of the expression to an existing file, one or the other of the statements

\[
\text{TeXForm[ expr1 ] >>> FileName}
\]
\[
\text{PutAppend[ TeXForm[ expr1 ], "FileName" ]}
\]

is necessary. Once created, this file can, of course, be edited or incorporated into other \LaTeX documents.

Automatic translation of arbitrary expressions into \LaTeX is a daunting task. While fairly sophisticated, the Mathematica translator is not perfect. Edits will often be necessary to “correct” the translator’s glitches. In particular, the translator writes only the \LaTeX description of the
8.15 Animation

The Mathematica command \texttt{Animate} facilitates the creation of animated displays in the NBI. The simple statement

\begin{verbatim}
Animate[ Plot[Sin[Pi*x]*Cos[2*Pi*t], {x,0.0,1.0}], 
 PlotRange->{ {0.0,1.0},{-1.0,1.0} }, 
 {t, 0.0, 1.0, 0.05} ]
\end{verbatim}

will, for example, display the fundamental mode of a vibrating string, keeping the axes the same size throughout all frames. A more complicated animation, display of the 23 mode of a square membrane, is produced by the statement

\begin{verbatim}
Animate[ Plot3D[Sin[2*Pi*x]*Sin[3*Pi*y]*Cos[2*Pi*t], {x,0.0,1.0}, {y,0.0,1.0}, 
 PlotRange-> { {0.0,1.0}, {0.0,1.0},{-1.0,1.0} }], 
 {t,0.0,1.0,0.05}, 
 AnimationRate-> 0.1]
\end{verbatim}

In both cases, the button labeled ‘Pause’ in the window displaying the animation will stop the animation and allow subsequent statements to be submitted to Mathematica.

Unfortunately, the command \texttt{Animate} has too many possibilities and options for full discussion here. Details can be found in the Mathematica manuals.

\footnote{For example, output of the \texttt{Mathematica} expression \texttt{Sin[x+y]/Cos[x-y]} with \texttt{TeXForm} yielded the correct but unintended result \texttt{\sin (x+y) \sec (x-y)}. Note also the unnecessary spaces after \texttt{sin} and \texttt{cos}.}

\footnote{\LaTeX\ files written by some earlier versions of Mathematica make use of several Mathematica-specific \LaTeX\ commands which are defined in the Mathematica-supplied \LaTeX\ package named \texttt{notebook2e} and the \LaTeX\-supplied package named \texttt{latexsym}. The files defining these packages must be accessible on your computer system before a \LaTeX\-produced \LaTeX\ source file can be processed and printed. The file \texttt{latexsym.sty} is probably available without effort on your part. If the file \texttt{notebook2e.sty} has not been placed where \LaTeX\ can find it automatically, you may have to copy it from the directory \texttt{\$MATHHEAD/SystemFiles/IncludeFiles/TeX/texmf/tex/latex/wolfram}, in which it was placed when \texttt{Mathematica} was installed on your system.}
8.16 Using the Notebook

Within the NBI, the window in which Mathematica commands are entered has full capabilities as a notebook, which means in particular that it can contain textual sections, e.g., documentation and motivation, as well as executable Mathematica commands. Notebooks are constructed out of cells, which may contain executable Mathematica statements (and the associated output)—the only cells we have so far seen—or may contain text. For purposes of organizing larger documents, these cells can themselves be assembled into sections and subsections, much in the way a conventional book would be assembled.

Suppose we already have on screen a Mathematica notebook containing a few Mathematica statements and their output. Note, in particular, that each statement and its output embraces two cells, one containing the input statement and the other containing the associated Mathematica response. The input cell is marked at the right side of the Notebook window with a closing square bracket whose upper corner contains also a diagonal line, and the output cell is similarly marked except that the bracket also has a short horizontal line below the diagonal line. Further, the two cells are themselves grouped together at the right side of the Notebook window with a closing square bracket that has no adornments at its corners. Finally, note that double clicking ML on the outer square bracket will “close” the pair, hiding the output and leaving only the input statement. In this arrangement, the outer bracket now has half an arrowhead at its lower end to indicate that the cell has been closed. Double clicking ML on the outer bracket when it is in this state will reopen the cell.

Now, suppose that we want to precede each statement with an explanatory paragraph. We could achieve that end for one of the statements by

1. Moving the cursor to the position at which a new cell is desired, placing it so that it becomes a horizontal I-beam.
2. Clicking ML to create a horizontal line (cell insertion bar) across the screen.
4. Typing the desired text, using the various options in the Format menu to specify font, type size, type style, alignment, ... As you start typing, a new cell will be created for the typed text. This cell will be bracketed on the right with a square bracket that has a short horizontal line below its top bar—the identifier of a text cell.

This process can, of course, be repeated to provide a textual description for each of the remaining statements.

In essence, the previous paragraph provides all we need to know to add documentation as an intrinsic part of a notebook. As our notebooks grow in size, however, we may wish to organize their contents into appropriate structural units and provide titles. To those ends, we need simply notice that the Style submenu of the Format menu contains numerous items beyond the one introduced in the previous paragraph. Positioning the cursor at the point of a desired textual insertion, clicking ML to create a horizontal line (cell insertion bar) across the screen, and then selecting some other item (Title, Subtitle, Section, Subsection, ...) from the Style submenu of the Format menu will then set Mathematica to create a cell of the selected style when we begin typing. Further, Mathematica will select a type size and weight appropriate to the selected cell style and, for sections and subsections, will place a marker at the beginning of the line to draw attention to its status in clarifying the structure of the document. Note also that, as a document is divided into sections and subsections, the bars along the right edge become more numerous, reflecting the divisions created by the inserted sections and subsections.

For example, to add a title to the entire document, we create an insertion bar at the very beginning of the document, select ‘Title’ from the Style submenu of the Format menu, and type
8.17. CUSTOMIZATIONS

the text of the title. Then, selecting this cell by highlighting its bracket at the right edge of the screen, we can adjust the font, type style, type size, alignment, and several other properties by appropriate selections from submenus accessible in the FORMAT menu.

Mathematica output in this notebook can, of course, always be quickly regenerated by executing the statements. Frequently, we may wish to suppress the presence of that output in the actual notebook. We can

- remove selected items of output by selecting the items to be removed and then selecting ‘Cut’ from the EDIT menu.
- remove all output from the notebook by selecting ‘Delete All Output’ from the CELL menu.

Finally, we note that, after all the work of creating a notebook, the notebook itself can be saved by selecting ‘Save’ from the FILE menu and, in the resulting pop-up window, specifying the file name (conventional file type .nb) and the directory, and clicking ‘Save’ on the ‘Save’ button. In a later session, the worksheet can be loaded into Mathematica by selecting ‘Open’ from the FILE menu and identifying the file to be opened in the pop-up browser.

This short section can, of course, not do full justice to the wide spectrum of features of Mathematica's notebook interface. You must work with it, explore the menus in the NBI, and read the Mathematica manuals.

8.17 Customizations

When Mathematica is started up, the configuration both of the kernel and of the front-end is defined by a number of default specifications built into the program, by statements in an assortment of system-wide initialization files and by statements in several user-created initialization files. Specifically,\(^{54}\)

- The file $BaseDirectory/Kernel/init.m configures the Mathematica kernel for all users and the file $BaseDirectory/FrontEnd/init.m configures the Mathematica front-end for all users. These files are maintained by the system manager and can be viewed but not edited by individual users.
- The file $UserBaseDirectory/Kernel/init.m modifies the configuration of the kernel established by the system-wide file and must be explicitly created by the individual user.
- The file $UserBaseDirectory/FrontEnd/init.m modifies the configuration of the front-end established by the system-wide file but is not created directly by the user. Instead, Mathematica maintains this file to reflect changes made via statements executed by Mathematica.

These files are simply command files—see Section 8.11—that can contain any statement that might otherwise be typed interactively. Be aware, however, that none of the statements or their responses will be displayed as the initialization file(s) is (are) read. Even so, for example, variables to which values are assigned and the features defined by packages read in will be available when reading of the file has been completed. As Mathematica is started, the system-wide files are executed and then the user-created file is executed. Thus, the user-created file can be used to override statements in the system-wide file, should that be desired. Together, these files—if they exist at all—define the default configuration set up by Mathematica when finally it is ready to accept input.

\(^{54}\)Here, the system variable $BaseDirectory identifies the top directory in the hierarchy in which the Mathematica structure is stored and the system variable $UserBaseDirectory identifies the top directory in which user-created configuration files are stored. Both of these directories may be hidden.
8.18 Miscellaneous Occasionally Useful Tidbits

In this section, we comment on a number of other features that are occasionally useful. Full details, of course, are available in the *Mathematica* manuals.

8.18.1 Turning off Messages

Many of *Mathematica*'s messages can be suppressed by executing the command `Off[... ]` and turned back on by executing the command `On[... ]`. Here, the ellipses stand for the name of the message. For example, to turn off display of the message that alerts you that a symbol you have just typed is nearly the same as some other symbol that you may have intended to type, you can execute the statement

```
Off[General::Spell1]
```

though full turning off of these messages may also require execution of the statement

```
Off[General::Spell]
```

These messages can be turned on again with the statements

```
On[General::Spell1]
On[General::Spell]
```

The system variable `$MessageGroups` stores a list of the existing message groups (the piece of the argument before the double colon).

8.18.2 Specifying Directories

Different operating systems use different forms to specify the path that identifies a particular directory, for example, to set the default directory or to identify the location of a file to be read or written. Within *Mathematica*, commands for performing these operations should separate directories in a path with the structure that is native to the operating system in use (e.g., forward slash in UNIX, backslash in Windows). For convenience, and to simplify command files to be used with several operating systems, *Mathematica* will translate forward slashes into the appropriate separator for the operating system in use.

8.18.3 Default Directory

The default directory or current working directory in *Mathematica* is set when *Mathematica* is launched. Unless an initialization file has been created with a command that changes that directory, it will probably be the user’s home directory or the directory from which *Mathematica* was launched by a command at a Shell prompt. The command `Directory[]` will cause the display of the current default directory. As illustrated by the statements

- `SetDirectory[]` Set directory to user’s home directory.
- `SetDirectory["..."]` Set directory to the parent of current directory.
- `SetDirectory["progs"]` Set directory to progs subdirectory of current directory.
- `SetDirectory["/apps/MATHEM/userprogs"]` Set directory to indicated directory.
the command `SetDirectory` allows dynamic specification of the current working directory in a variety of ways. Note that the separator between successive directories in a path varies from operating system to operating system. The illustrated separator is the one used in UNIX. Further information about these commands will be found in the Mathematica manuals. Further, the statement `FileNames[]` will display a list of the files in the current directory and the statement `FileNames["*.m"]` will display a list of all files in that directory with filetype `.m`. If the file is not in the default directory, then its full path must be specified as part of the filename.

### 8.18.4 Search Path

Whenever an explicit path is not specified for a file that Mathematica is to read, the program looks in a sequence of directories defined by the search path. The initial search path is defined in a Mathematica initialization file or, in the absence of such a file, by Mathematica itself. The initial search path is site-specific and is identified in the Local Guide. If not defined by an initialization file, the path is likely to contain, in order, a few specific Mathematica directories in the user’s home directory, the default directory, the user’s home directory, and finally a succession of directories descending from the directory `$MATHEMHEAD`. The system variable `$Path` contains a list of the directories in the current search path. Thus, the statement `$Path` will display that list, and statements like

```mathematica
Prepend[ $Path, "/usr/share/CPSUP/mathematica" ]
Append[ $Path, "/usr/local/packages" ]
Insert[ $Path, "/usr/local/temp", 5 ]
```

will create a new value for the list `$Path`, that new value containing all of the elements of the original list in the original order but having the indicated directory at the beginning, at the end, or in the fifth position, respectively. Similarly, statements like

```mathematica
Delete[ $Path, 5 ]
ReplacePart[ $Path, "/usr/local/myfiles", 7 ]
```

will create a new value for the list `$Path` by removing altogether the element in the fifth position or by replacing the element in the seventh position with the indicated directory, respectively.

### 8.18.5 Displaying and Changing Default Options

In the course of introducing several commands, we have learned that the precise effect of many commands depends on the values of associated options. Initially, we accepted Mathematica’s defaults. Then we learned that arguments like

```
OptionName -> NewValue
```

(e.g., `PlotRange -> {-1.0,1.0}`) can be used within a command to override the default value for the indicated option, but only for that execution of the command. If the default value is to be overridden for all subsequent executions of the associated command, however, changing the default with a statement like

```
SetOptions[ CommandName, OptionName -> NewValue ]
```

---

55 See Section 8.18.2.

56 For definiteness, we use the UNIX form for indicating paths. Remember that the precise form in which paths are specified depends on the operating system. The form appropriate at your site is described in the Local Guide.
(e.g., `SetOptions[ Plot, PlotRange -> {-1.0, 1.0} ]`) may be more convenient—though this change will, of course, be in effect only until it is changed again (or until a new Mathematica session is launched). Finally, the statement

`Options[ CommandName ]`

(e.g., `Options[ Plot ]`) will display a list of all options admitted by the indicated command, each with its current default value.

### 8.18.6 Placing Text in Graphs

With suitable commands, we can generate a graphical display to which specified text has been added in the body of the display. Suppose, for example, we wanted to label a graph of the sine function with a string of characters positioned not as a title on the graph as a whole or as a label on an axis but rather as a notation at specified coordinates within the graph. We might use statements like

```mathematica
In[48]:= Plot[ Sin[x], {x, 0.0, 10.0} ];
In[49]:= Graphics[ Text[ "New Graph", {5.0,0.5} ] ];
In[50]:= Show[ { %48, %49 } ]
```

The first of these statements creates a graphics object containing the desired sine curve, the second statement uses the graphics primitive `Text` to create a second graphics object containing the specified text centered at the specified location in the coordinates of the graph and oriented horizontally. And the third statement creates and displays a composite graphics object assembled from

- the graphics object created by line `In[48]` and known by the name `%48`, and
- the graphics object created by line `In[49]` and known by the name `%49`.

We could, of course, have nested the first two statements within the third, but the result would have been syntactically difficult to fathom.

Of course, this quick example only hints at the potential of these commands. The Mathematica manuals describe many more graphics primitives that can be used as arguments to the `Graphics` command and also describe numerous options associated both with the `Graphics` command and with the `Show` command. Careful study of those items in the Mathematica manuals is prerequisite to full exploitation of Mathematica’s abilities to create graphics objects and combine them in elaborate displays.

### 8.18.7 Space Curves

Sometimes it is desirable to view the trajectory of a particle in three-dimensional space. MATHEMATICA has the ability to plot space curves from knowledge of the trajectory defined parametrically by the functions \( x(t) \), \( y(t) \), and \( z(t) \) giving the coordinates of points on a three-dimensional path. With that information, the command `ParametricPlot3D` will calculate the coordinates of the points along the trajectory, project these points onto the two-dimensional screen, and connect consecutive points with lines. To illustrate this feature, consider the equations

\[
x(t) = 5 \cos t \quad ; \quad y(t) = 5 \sin t \quad ; \quad z(t) = \alpha t
\]

(8.5)

---

57 The `Graphics` command accepts one or two additional arguments, each of which is a list. For example, the alternative command `Graphics[ Text[ "New Graph", {5.0,0.5}, {-1,0}, {0,1} ] ]` will place the left end of the text at the specified point (third argument) and orient the text vertically and reading from the right side of the graph (fourth argument). Other options for the third and fourth arguments are described in the Mathematica manuals.
Figure 8.5: Space curve of a charged particle moving in a constant magnetic field along the z axis.

where $\alpha$ is a constant, describing the trajectory of a charged particle moving in a constant magnetic field directed along the z axis. The starting point is, of course, to evaluate $x$, $y$, and $z$. Then, we invoke the command `plot3d` to plot the graph. The statements

```
In[1]:= alpha = 0.5;  
Set alpha.

In[2]:= ParametricPlot3D[{5*Cos[t], 5*Sin[t], alpha*t}, {t, 0, 30},  
PlotStyle -> {Thickness[0.02], Black}]
```

will produce the graph in Fig. 8.5.

### 8.18.8 Directing Plot to Separate Window

The several simple graphics commands we have introduced all direct their output to the window in which the commands are issued. Should you wish to have a new graphics window created for a particular plot, you need only add the Option

```
DisplayFunction -> CreateDialog
```

to whatever else in incorporated in the command. For example, the statement

```
Plot[ Tanh[x], {x, -5.0, 5.0} ]
```

produces the specified graph in line while the statement

```
Plot[ Tanh[x], {x, -5.0, 5.0}, DisplayFunction->CreateDialog ]
```

directs the graph to a newly created (small) separate window. By the second route, the first graph remains on the screen even if the in line graph scrolls off as further statements are executed.
8.19 References

In this chapter, we have introduced only the most important features of Mathematica. A full description of all commands and features is contained in the Mathematica manuals, including

- several documents and resources as described in Section 8.2.
- a number of documents, links to which can be found by entering “Mathematica help” in the search window of your browser. In particular, you should find links to
  - the official Mathematica documentation at http://reference.wolfram.com/language/, which brings up on line the same documentation available from the HELP menu in the NBI, and
  - many other official and unofficial documents.
- a number of books and on-line materials describing a wide spectrum of uses of Mathematica in a variety of disciplines and for a variety of specific purposes. A list of such books can be found, among other places, from the links http://www.wolfram.com/mathematica/resources/ and http://www.wolfram.com/books/ at the Wolfram Research website or by searching on-line commercial book stores (Amazon, Barnes and Noble, ...) for the keyword Mathematica.
- more specific documents on particular Mathematica commands, links to which can be found by entering “CommandName in MATHEMATICA”, e.g., “integrate in MATHEMATICA” in the search window of your browser.

The user is urged to browse in these materials, being alert not only to the available functions but also to the assorted option variables and keywords that modify their behavior.

While on-line documentation has largely superseded printed documentation in the last decade, several now largely outdated but still potentially valuable printed materials can be found. Specifically, your local library may include one or more copies of

- Getting Started with Mathematica Version # under (operating system), (Wolfram Media, Inc.), which contains installation instructions and a brief description of the issues with which a user of Mathematica must be familiar from the very beginning.
- Stephen Wolfram, The Mathematica Book, Fifth (or earlier) Edition (Wolfram Media, Inc., and Cambridge University Press, 1988, 1991, 1996, 1999, 2003), which contains a tour of the capabilities of Mathematica, an introduction to the practical use of Mathematica, an exhaustive discussion of the design principles underlying Mathematica, a treatment of more advanced topics, several pages of formulae, a section of color plates showing a spectacular array of graphical displays, and a comprehensive set of appendices listing and describing essentially all commands, options, special variables, and other objects that play a role somewhere in Mathematica. This tome of nearly 1500 pages in a single binding is massive and exhaustive.
- Mathematica Version # Standard Add-on Packages (Wolfram Media, Inc., 1996, 1999), which contains detailed descriptions of all components in all of the standard add-on packages.

8.20 Exercises

8.1. Read the Mathematica manuals to find out how the commands Expand, Factor, Apart, and Simplify differ from one another and then invent several examples that will reveal those differences. To help you get started, you might find it useful to compare the effect of each of these commands on the expressions
8.20. **EXERCISES**

(a) \( \frac{x}{a+b} + \frac{y}{a+b} \) \hspace{1cm} (b) \( (a+b)(c+d) + (a+b)c \)

(c) \( (x+y)(z+w) \) \hspace{1cm} (d) \( a(b+c(d+e)) \)

(e) \( \frac{x}{a+b} + \frac{y}{c+d} \) \hspace{1cm} (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.

8.2. Use **Mathematica** to convert each of the expressions in the left-hand column in the table below into the expression in the associated right-hand column:

| (a) \( \frac{(a-x)^2}{(a^2-2ax+x^2)^{3/2}} \) | \( \frac{1}{|x-a|} \) |
| (b) \( \sinh( \ln(x+\sqrt{x^2+a^2}) - \ln(a) ) \) | \( \frac{x}{a} \) |
| (c) \( \frac{1}{x+\sqrt{y}} \) | \( \frac{x-\sqrt{y}}{x^2-y} \) |

Write two or three paragraphs in which you describe your efforts, including some indication of approaches that were not successful. Don’t be overly concerned about the order of terms within various sets of parentheses; that order is particularly difficult to control. Focus instead on creating the general form of each desired result.

8.3. Use **Mathematica** to convert each of the expressions in the left-hand column in the table below into the expression in the associated right-hand column:

| (a) \( \frac{d}{dx} \left( xe^{-x^2} \right) \) | \( -2x(x^2-1)e^{-x^2} \) |
| (b) \( \sin(\sqrt{a^2(a+3x)+x^2(3a+x)+y}) \) | \( \sin((a+x)^{3/2}+y) \) |
| (c) \( cg+cf+b^2d+2abd+a^2d+b^2c+2abc+a^2c \) | \( (a+b)^2(c+d)+c(f+g) \) |
| (d) \( ae^{-bt}(e^{-i\omega t}+e^{i\omega t}) \) | \( \frac{a e^{-bt}}{e^{-i\omega t}+e^{i\omega t}} \) |
| (e) \( ae^{-bt}(e^{-i\omega t}+e^{i\omega t}) \) | \( 2 a e^{-bt} \cos(\omega t) \) |
| (f) \( x^2+y^2+z^2-2a(x+y)+2a^2 \) | \( (x-a)^2 + (y-a)^2 + z^2 \) |

In several cases, you may need to invoke **Part**, **ReplacePart**, and/or **FullForm**, but other commands will surely also be needed. Invoke microscopic dissection of the expressions only as a last resort. Write two or three paragraphs in which you describe your efforts, including some indication of approaches that were not successful. Don’t be overly concerned about the order of terms within various sets of parentheses in the final form; that order is particularly difficult to control. Focus instead on creating the general form of each desired result.

8.4. In Section 8.8.9, we set up the differential equation for a damped harmonic oscillator but then pursued the solution only for the underdamped case. Starting with the dimensionless form\(^{58}\)

\[
\frac{d^2x}{dt^2} + 2a \frac{dx}{dt} + x = 0 ; \quad x(0) = 1 ; \quad \left. \frac{dx}{dt} \right|_{t=0} = 0
\]

\(^{58}\)The forms of the solution work out more easily if, in contrast to the choice in the text, we express the coefficient of the second term as twice a constant, so \( \alpha = b/2m\omega_0 \).
of the equation deduced towards the end of Section 8.8.9, work out the solution for the underdamped case and then find also the solutions for the critically damped and overdamped cases defined, respectively by

\[ \alpha > 0 \quad \text{and} \quad \alpha <, =, > 1 \]

Describe the differences in the physical behavior for the three cases. For simplicity, take the initial conditions to be \( x(0) = 1 \) and \( v(0) = 0 \).

8.5. The Legendre polynomials \( P_n(x) \), which are valid and useful over the interval \(-1 \leq x \leq 1\), can be defined in many ways. They emerge as the coefficients in the Taylor expansion of the generating function

\[ g(x,t) = \frac{1}{\sqrt{1 - 2xt + t^2}} = \sum_{n=0}^{\infty} P_n(x) t^n \]

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

- \( P_0(x) = 1 \)
- \( P_1(x) = x \)
- \( P_2(x) = \frac{1}{2} (3x^2 - 1) \)
- \( P_3(x) = \frac{1}{2} (5x^3 - 3x) \)
- \( P_4(x) = \frac{1}{8} (35x^4 - 30x^2 + 3) \)
- \( P_5(x) = \frac{1}{8} (63x^5 - 70x^3 + 15x) \)

a. Use the generating function and Mathematica’s capabilities for evaluating Taylor series to find the first half-dozen Legendre polynomials, extracting each as an expression bound to a variable. The commands `Collect` and `CoefficientList` may be useful.

b. Start by binding the value 1 to \( P[0] \) and the value \( x \) to \( P[1] \). Then, using the recursion relationship, find the next several Legendre polynomials. *Hint*: The Mathematica statements might be

\begin{verbatim}
In[29]:= P[0] = 1; P[1] = x;
and then
In[30]:= PP = (2*n-1)*x*P[n-1]/n - (n-1)*P[n-2]/n
\end{verbatim}

(Verify the expression on the right by using Mathematica to deduce this relationship from the standard form—the second equation in this exercise.) With these statements, you have set up \( P_0(x) \) and \( P_1(x) \) to start the recursion and then you have defined an expression \( PP \) depending on \( n \) that can be evaluated at any \( n \). Once \( P_0 \) and \( P_1 \) have been defined, you can find \( P_2 \) and then \( P_3 \) and then ... with statements like

\begin{verbatim}
In[32]:= P[3] = PP /., n -> 3
\end{verbatim}

... c. Find the first half-dozen Legendre polynomials by using the command \texttt{D} to evaluate Rodrigues’ formula. *Hint*: You might find that using a loop would simplify your approach.

d. Be clever and, using either matrices or loops constructed in Mathematica, find the values of all of the integrals

\[ \int_{-1}^{1} P_n(x) P_m(x) \, dx \]

where \( n \) and \( m \) take on independently the values 0, 1, 2, 3, 4, 5. (There are 36 integrals to be evaluated. Try to be efficient in your coding, and note that the command `Integrate` applied to a matrix will automatically apply the command element-by-element.)
e. Within Mathematica, obtain graphs of the first six Legendre polynomials over the interval $-1 \leq x \leq 1$.

f. It is known that a function $f(x)$ defined over the interval $-1 \leq x \leq 1$ can be expanded in a series of Legendre polynomials of the form $f(x) = \sum a_n P_n(x)$ where the coefficients $a_n$ are given by

$$a_n = \frac{2n + 1}{2} \int_{-1}^{1} f(x) P_n(x) \, dx$$

Find the first six coefficients for the expansion of the function $f(x) = 0$ when $-1 < x < 0$ and $f(x) = 1$ when $0 < x < 1$. Then, construct the (partial) series representing this function and obtain a graph of that approximation to compare with the graph of the original function.

g. Mathematica actually knows quite a bit about many of the important special functions of mathematical physics. In particular, the $n$-th Legendre polynomial as a function of $x$ is known by the name `LegendreP[n, x]` and is described fully in the Mathematica manuals. Take a look at that documentation and then use that function to determine the first several Legendre polynomials.

8.6. Figure 8.6 shows the circuit diagram for a Wheatstone bridge. Using Kirchhoff’s laws, set up the equations from which you could determine the currents in each branch of the circuit. Then, using Mathematica, (a) solve the equations symbolically, (b) find conditions under which the current in the cross branch (through resistor $R_5$) will be zero, and (c) find the effective resistance seen by the battery. (The effective resistance is defined by the ratio $V/I$, where $I$ is the current in the branch containing the battery.)

8.7. In the scattering of a quantum wave from a rectangular barrier in particular circumstances, we find that the wave function must be expressed in three pieces in the form

$$\psi(x) = \begin{cases} Ae^{ikx} + Be^{-ikx} & x < 0 \\ D \cosh(\kappa x) + F \sinh(\kappa x) & 0 < x < w \\ Ce^{ikx} & x > w \end{cases}$$

where $k$ and $\kappa$ are constants related to the energy of the particle and the height of the barrier, $A$ is a constant reflecting the intensity of the incident beam, and $B$, $D$, $F$, and $C$ are constants to be determined by imposing the requirement that the wave function and its first derivative be continuous both at $x = 0$ and at $x = w$, i.e., that

$$\psi(0^-) = \psi(0^+) \quad \psi(w^-) = \psi(w^+) \quad \frac{d\psi}{dx}\bigg|_{0^-} = \frac{d\psi}{dx}\bigg|_{0^+} \quad \frac{d\psi}{dx}\bigg|_{w^-} = \frac{d\psi}{dx}\bigg|_{w^+}$$

where superscript plus and minus signs identify points slightly below and slightly above the indicated value of $x$, respectively. (While $k$ and $\kappa$ can be taken to be real for this barrier, the constants $A$, $B$, $C$, $D$, and $F$ may—and probably will—be complex.) Use Mathematica’s abilities to manipulate expressions to
(a) Obtain the equations determining $B$, $C$, $D$, and $F$ by imposing the stated boundary conditions on these solutions.

(b) Solve those equations for those constants (expressing each as a multiple of the constant $A$).

(c) Show that the reflection and transmission coefficients $R$ and $T$ defined by $R = |B/A|^2$ and $T = |C/A|^2$ are given by

$$R = \left| \frac{B}{A} \right|^2 = \frac{(\kappa^2 + k^2)^2 \sinh^2 \kappa w}{4\kappa^2 k^2 + (\kappa^2 + k^2)^2 \sinh^2 \kappa w}; \quad T = \left| \frac{C}{A} \right|^2 = \frac{4k^2 \kappa^2}{4\kappa^2 k^2 + (\kappa^2 + k^2)^2 \sinh^2 \kappa w}.$$

Here, the vertical bars symbolize the absolute value of the complex number enclosed by them.

(d) Verify that $R + T = 1$.

8.8. The distribution of wavelengths $\lambda$ in the blackbody spectrum at (absolute) temperature $T$ is given by

$$u(\lambda, T) = \frac{8\pi c h}{\lambda^5} \frac{1}{e^{ch/\lambda k T} - 1}$$

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 Mathematica, (b) obtain a graph of $f(y)$ versus $y$, making sure to extend the graph over an interval that includes its peak and estimate the value of $y$ at which that peak occurs; (c) show that the peak occurs for values of $y$ satisfying

$$(y - 5) e^y + 5 = 0$$

(d) obtain a graph of this function versus $y$; (e) make another estimate of the value of $y$ at which the original function has its maximum; and (f) show that the wavelength $\lambda_m$ at which this maximum occurs satisfies

$$\lambda_m T = 0.28978 \times 10^{-2} \text{ m K}$$

Hints: (1) Remember that maxima in a function occur where the derivative of that function with respect to the appropriate variable is zero. (2) Because $e^y$ varies rapidly with $y$, you may have to play a bit to find a suitable range of values of $y$ over which to plot these graphs.

8.9. Use Mathematica to verify (a) that the solution

$$x(t) = e^{-\tau a} x_0 \left( \cos [at] + \frac{\gamma \sin [at]}{a} \right)$$

with

$$a = \sqrt{\omega^2 - \gamma^2} \quad \text{and} \quad \gamma = \frac{b}{2m} \quad \omega^2 = \frac{k}{m}$$

presented in Section 8.8.9 satisfies the desired differential equation and the initial conditions presented at the beginning of that section and (b) that this solution and the solution

$$y(\tau) = \frac{x_0}{2\sqrt{-1 + \beta^2}} \left( e^{-\beta - \sqrt{-1 + \beta^2} \tau} + e^{(-\beta + \sqrt{-1 + \beta^2}) \tau} \beta + e^{-\beta - \sqrt{-1 + \beta^2} \tau} \sqrt{-1 + \beta^2} + e^{(-\beta + \sqrt{-1 + \beta^2}) \tau} \sqrt{-1 + \beta^2} \right)$$

where $\tau = \omega t$ and $\beta = \gamma/\omega$ obtained at the very end of the section agree with each other.

8.10. The Laguerre polynomials $L_n(x)$ can be defined in many ways. We might, for example, set $L_0(x) = 1$ and then generate each new polynomial in turn by requiring that $L_n(x)$ for $n > 0$ be a polynomial of order $n$ that is orthogonal to all previous polynomials, i.e., that

$$\int_0^\infty e^{-x} L_n(x) L_m(x) \, dx = 0 \quad \text{for all} \quad m < n.$$
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
\[
\begin{align*}
L_0(x) &= 1 \quad L_3(x) = 1 - 3x + \frac{3}{2}x^2 - \frac{1}{6}x^3 \\
L_1(x) &= 1 - x \quad L_4(x) = 1 - 4x + 3x^2 - \frac{3}{2}x^3 + \frac{1}{24}x^4 \\
L_2(x) &= 1 - 2x + \frac{1}{2}x^2 \quad L_5(x) = 1 - 5x + 5x^2 - \frac{5}{3}x^3 + \frac{5}{24}x^4 - \frac{1}{120}x^5
\end{align*}
\]
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. \textit{Hint:} The \textit{Mathematica} statements might be
\[
\begin{align*}
\text{In}[29]:= & L[0] = 1; \quad L[1] = 1-x; \\
\text{In}[30]:= & LL = (2*n-1-x)*L[n-1]/n - (n-1)*L[n-2]/n
\end{align*}
\]
(Verify the expression on the right.) With these statements, you have set up \( L_0(x) \) and \( L_1(x) \) to start the recursion and then you have defined an expression \( LL \) involving \( n \) that can be evaluated at any \( n \). Once \( L_0 \) and \( L_1 \) have been defined, you can find \( L_2 \) and then \( L_3 \) and then \( \ldots \) with statements like
\[
\begin{align*}
\text{In}[31]:= & L[2] = LL /. n -> 2 \\
\text{In}[32]:= & L[3] = LL /. n -> 3 \\
& \ldots
\end{align*}
\]
c. Find the first half-dozen Laguerre polynomials by using the command \texttt{D} to evaluate Rodrigues’ formula.

d. Be clever and, using either matrices or loops constructed in \textit{MATHEMATICA}, find the values of all of the integrals
\[
\int_0^{\infty} e^{-x} L_n(x) L_m(x) \, dx
\]
where \( n \) and \( m \) take on independently the values 0, 1, 2, 3, 4, 5. (There are 36 integrals to be evaluated. Try to be efficient in your coding, and remember the command \text{map}, which will be useful in constructing a single statement that evaluates the integral of each element in a multi-element structure.)

e. \textit{Mathematica} actually knows quite a bit about many of the important special functions of mathematical physics. In particular, the \( n \)-th Laguerre polynomial as a function of \( x \) is known by the name \texttt{LaguerreL[n,x]} and is described fully in the \textit{Mathematica} manuals. Take a look at that documentation and then use that function to determine the first several Laguerre polynomials.

8.11. Find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
0 & 1 & 0 \\
1 & 0 & 1 \\
0 & 1 & 0
\end{pmatrix}
\]
8.12. (a) Use the function `Eigensystem` to find the eigenvalues and eigenvectors of the matrix
\[
\begin{pmatrix}
2 & -1 & 0 & 0 \\
-1 & 2 & -1 & 0 \\
0 & -1 & 2 & -1 \\
0 & 0 & -1 & 2
\end{pmatrix}
\]
(b) Use the function `FullForm` to explore the structure of the entity returned by the function `Eigensystem` used in part (a).

8.13. When a (weak) constant external electric field of magnitude $F$—we reserve $E$ for energy in this exercise—is imposed on a hydrogen atom, the energy of the states with principal quantum number $n$ shift from the energy given by the Bohr model by amounts determined by the eigenvalues of the matrix whose elements are $\langle nlm|eF|nl'm'\rangle$, 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
\[
3e_aF \begin{pmatrix}
0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]
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
\[
3e_aF \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -9/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -9/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -3\sqrt{6} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -3\sqrt{6} & 0 & -9/\sqrt{3} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -9/\sqrt{3} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -9/2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -9/2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Find the eigenvalues and eigenvectors of these matrices. The eigenvalues give the energy shifts for the Stark effect for $n=2$ and $n=3$ and the eigenvectors give the linear combinations of the base states (i.e., the states in the absence of the external field) out of which the states in the presence of the field emerge as the field is turned on. *Hint:* The simplest way to create a sparse matrix is to begin by executing the statements

```
<< LinearAlgebra`MatrixManipulation`
mat = ZeroMatrix[4,4];
```

which load the add-on package that defines the command `ZeroMatrix` and create a $4 \times 4$ matrix of zeros. Continuing, we then set the nonzero entries with statements like

```
mat[[1,3]] = desired value
```

For verification, the final matrix can be displayed in matrix form with the statement

```
MatrixForm[ mat ]
```

8.14. As shown in Fig. 8.7, an object of mass $m$ is connected to the center of each of the four sides of a square of sides $2\ell$ with a spring of constant $k$ and moves on a horizontal frictionless surface in the plane defined by the square. Take the equilibrium position of the object at the center of the square to be the origin of an $xy$ coordinate system, and let the springs have unstretched length $\ell_0$. For this situation, the kinetic and potential energies $KE$ and $PE$ of the object are given by

\[
KE = \frac{1}{2}mx^2 + \frac{1}{2}my^2 ; \quad PE = \frac{1}{2}k\left[\sqrt{(\ell - x)^2 + y^2} - \ell_0\right] + \ldots
\]
where \( PE \) will have three additional terms, one for each of the remaining springs. (The term shown applies to the spring connected to the right side of the square.) By definition, the Lagrangian function for this problem is given by \( L = KE - PE \) and the equations of motion can be extracted from the Lagrangian by evaluating the expression

\[
\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}} \right) - \frac{\partial L}{\partial q} = 0
\]

where \( q \) stands first for \( x \) and then for \( y \). Write out the full potential energy and then use Mathematica to deduce the equations of motion

\[
m\ddot{x} = \ldots \quad \text{and} \quad m\ddot{y} = \ldots
\]

for this system. The results are quite complicated. To simplify the problem, find in detail the equations only for the cases (a) \( \ell_0 = 0 \), i.e., the springs have unstretched length of 0—admittedly unrealistic springs, and (b) \( x \) and \( y \) remain small compared to \( \ell \) throughout the motion, i.e., \( x/\ell \ll 1 \) and \( y/\ell \ll 1 \) for all time.

Note: Statements like \( D[L, D[x[t], t]] \) are perfectly acceptable to Mathematica, provided the dependence of \( x \) on \( t \) is explicitly indicated in setting up the Lagrangian \( L \). Thus, for example, in a different problem (particle of mass \( m \) falling under the gravitational attraction of the earth), for which we might set

\[
L = mD[x[t], t]^2/2 - mgx[t]
\]

we can evaluate \( \partial L/\partial x \) with the statement

\[
DLDx = D[L, x[t]]
\]

evaluate \( \partial L/\partial \dot{x} \) with the statement

\[
DLDx = D[L, \dot{x}]
\]

and then, finally, determine the equation of motion with the statement

\[
D[ DLDx, t] - DLDx = 0
\]

8.15. In the Cartesian coordinate system illustrated in Fig. 8.8, the coordinates of the two objects having mass \( m_1 \) and \( m_2 \) in the compound pendulum are given by

\[
x_1 = l_1 \sin \theta; \quad y_1 = -l_1 \cos \theta; \quad x_2 = l_1 \sin \theta + l_2 \sin \phi; \quad y_2 = -l_1 \cos \theta - l_2 \cos \phi
\]

Taking \( \theta \) and \( \phi \) to be the generalized coordinates and remembering that

\[
KE = \frac{1}{2} m_1 x_1^2 + \frac{1}{2} m_1 y_1^2 + \frac{1}{2} m_2 x_2^2 + \frac{1}{2} m_2 y_2^2; \quad PE = m_1 gy_1 + m_2 gy_2
\]

use Mathematica to find (a) (simple) expressions for \( KE, PE \), and \( L = KE - PE \), (b) the equations of motion by evaluating the Lagrange equations

\[
\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\theta}} \right) - \frac{\partial L}{\partial \theta} = 0; \quad \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\phi}} \right) - \frac{\partial L}{\partial \phi} = 0
\]
and (c) the equations of motion when both \( \theta \) and \( \phi \) are small. Suggestion: If you run into major difficulties keeping \( l_1 \neq l_2 \) and \( m_1 \neq m_2 \), make the lengths and masses the same. Note: See the note in the previous exercise.

8.16. The complete elliptic integrals of the first and second kinds are given by

\[ K(k) = \int_0^{\pi/2} \frac{d\phi}{(1 - k^2 \sin^2 \phi)^{1/2}} ; \quad E(k) = \int_0^{\pi/2} (1 - k^2 \sin^2 \phi)^{1/2} \, d\phi \]

Use Mathematica to evaluate these integrals to \( O(k^6) \) by expanding the integrands in Taylor series before evaluating the integrals.

8.17. Find the Laplace transform of each of the functions

(a) \( f(t) = t^n \)  
(b) \( f(t) = t e^{-at} \)  
(c) \( f(t) = \cosh(at) \)  
(d) \( \frac{df(t)}{dt} \)

and the inverse Laplace transform of each of the functions

(e) \( \tilde{f}(s) = \frac{a + bs}{s^2 + \omega^2} \)  
(f) \( \tilde{f}(s) = \frac{a}{(s^2 + 9\omega^2)(s^2 + 4\omega^2)(s^2 + \omega^2)} \)

In (c), you may have to express the hyperbolic cosine in exponential form while in (f) you may have to help Mathematica’s routine \texttt{InverseLaplaceTransform} by first invoking the command \texttt{Apart} to expand the desired function in partial fractions.

8.18. Given the three points \((x_i, y_i)\), \(i = 1, 2, 3\), (a) find symbolic expressions for the coefficients \(a, b, \) and \(c\) of the parabola \( y = ax^2 + bx + c \) that passes through these three points and then (b) find a symbolic expression for the value of \( x \) at which the extreme point of the parabola occurs. Finally, (c) determine numerically the angle at which the maximum range of a projectile occurs if the ranges at \( \theta = 39^\circ, 40^\circ, \) and \( 41^\circ \) are 0.7251744, 0.7259383, and 0.7258887, respectively.

8.19. Find symbolic expressions for the coefficients \(a, b,\) and \(c\) that will cause the parabola \( y = ax^2 + bx + c \) to pass through the three points \((x_1 = x_2 - \delta x, y_1), (x_2, y_2),\) and \((x_3 = x_2 + \delta x, y_3)\). Then integrate the parabola over the interval \( x_1 \leq x \leq x_3 \) and deduce Simpson’s rule

\[ \int_{x_1}^{x_3} y(x) \, dx = \frac{\delta x}{3} (y_1 + 4y_2 + y_3) \]

8.20. The midpoint rule \( M \) and Simpson’s rule \( S \) for evaluating integrals numerically start with the assumptions that

\[ \int_a^b f(x) \, dx \approx M = (b - a) f \left( \frac{a + b}{2} \right) ; \quad \int_a^b f(x) \, dx \approx S = \frac{b - a}{6} \left( f(a) + 4 f \left( \frac{a + b}{2} \right) + f(b) \right) \]
8.20. EXERCISES

149

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 endpoints 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 \( f(x) = cx^3 + dx^2 + ex + g! \) (Optional: Try to construct geometric arguments that would provide insight into the correctness of these analytic results.

8.21. Using Mathematica, obtain graphs of the potential

\[
U(x) = -\frac{U_0 a^2 (a^2 + x^2)}{8a^4 + x^4}
\]

and the associated force. Note that the expression for the potential is simpler to plot if you plot not \( U(x) \) versus \( x \) but rather \( U(x)/U_0 \) versus \( \varphi = x/a \), i.e., rewrite the function in the form

\[
\frac{U(x)}{U_0} = 1 + \frac{x^2}{8 + x^4}
\]

You should be able to find a similar dimensionless version of the expression giving the force.

8.22. Using Mathematica, find complete (symbolic) solutions to each of the following problems, and use Mathematica also to verify that the solutions you obtain actually do satisfy the original ODE and initial conditions.

a. \( \frac{d^2x}{dt^2} = a \), \( x(0) = x_0 \), \( v(0) = v_0 \), 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 \( v = v_0 \hat{i} + v_0 \hat{k} \).

8.23. Using the functions Grad, Laplacian, Div, and Curl defined in Section 8.8.10,

a. Evaluate the gradient of (i.e., the negative of the force field associated with) each of the functions

\[
V_1(x, y, z) = \frac{1}{(x^2 + y^2 + z^2)^{1/2}} \quad V_2(x, y, z) = \frac{z}{(x^2 + y^2 + z^2)^{3/2}}
\]

\[
V_3(x, y, z) = \frac{2z^2 - x^2 - y^2}{(x^2 + y^2 + z^2)^{5/2}} \quad V_4(x, y, z) = \frac{e^{-\sqrt{x^2+y^2+z^2}}}{(x^2 + y^2 + z^2)^{1/2}}
\]

of the Cartesian variables \( (x, y, z) \).

b. Verify that each field obtained in part (a) is conservative by showing that the curl of each is zero.
c. With \( \mathbf{r} = x \hat{i} + y \hat{j} + z \hat{k} \), show (1) that \( \nabla \times \mathbf{r} = 0 \) and (2) that \( \nabla \cdot \mathbf{r} = 3 \).
d. With \( V_5(x,y,z) = x^2 + y^2 + z^2 \), show that \( \nabla^2 V_5(x,y,z) = \nabla \cdot \nabla V_5(x,y,z) = 6 \).
e. Show that \( \nabla^2 V_1(x,y,z) = 0 \). (Note: Except at \( x = y = z = 0 \))
f. Evaluate the Laplacian of \( V_4(x,y,z) \), \( \nabla^2 V_4(x,y,z) \).

8.24. Using the functions \texttt{Grad}, \texttt{Laplacian}, \texttt{Div}, and \texttt{Curl} defined in Section 8.8.10,

a. Evaluate the gradient of (i.e., the negative of the force field associated with) each of the functions

\[
V_1(r, \phi, z) = \frac{1}{r} \quad V_2(r, \phi, z) = \frac{z}{(r^2 + z^2)^{3/2}}
\]

\[
V_3(r, \phi, z) = \frac{2r^2 - r^2}{(r^2 + z^2)^{3/2}} \quad V_4(r, \phi, z) = \frac{e^{-a(r^2 + z^2)^{1/2}}}{(r^2 + z^2)^{3/2}}
\]

of the cylindrical variables \( r \) (radial), \( \phi \) (azimuthal), and \( z \) (axial).

b. Verify that each field obtained in part (a) is conservative by showing that the curl of each is zero.
c. With \( \mathbf{r} = r \hat{\mathbf{r}} + z \hat{k} \), show (1) that \( \nabla \times \mathbf{r} = 0 \) and (2) that \( \nabla \cdot \mathbf{r} = 3 \).
d. With \( V_5(r, \phi, z) = r^2 + z^2 \), show that \( \nabla^2 V_5(r, \phi, z) = \nabla \cdot \nabla V_5(r, \phi, z) = 6 \).
e. Show that \( \nabla^2 V_1(r, \phi, z) = 0 \). (Note: Except at \( r = z = 0 \))
f. Evaluate the Laplacian of \( V_4(r, \phi, z) \), \( \nabla^2 V_4(r, \phi, z) \).

8.25. Using the functions \texttt{Grad}, \texttt{Laplacian}, \texttt{Div}, and \texttt{Curl} defined in Section 8.8.10,

a. Evaluate the gradient of (i.e., the negative of the force field associated with) each of the functions

\[
V_1(r, \theta, \phi) = \frac{1}{r} \quad V_2(r, \theta, \phi) = \frac{z}{r^3} \quad V_3(r, \theta, \phi) = \frac{r^2(3 \cos^2 \theta - 1)}{r^5} \quad V_4(r, \theta, \phi) = \frac{e^{-ar^2/2}}{r}
\]

of the spherical coordinates \( r \) (radial), \( \theta \) (polar), and \( \phi \) (azimuthal).

b. Verify that each field obtained in part (a) is conservative by showing that the curl of each is zero.
c. With \( \mathbf{r} = r \hat{\mathbf{r}} \), show (1) that \( \nabla \times \mathbf{r} = 0 \) and (2) that \( \nabla \cdot \mathbf{r} = 3 \).
d. With \( V_5(r, \theta, \phi) = r^2 \), show that \( \nabla^2 V_5(r, \theta, \phi) = \nabla \cdot \nabla V_5(r, \phi, z) = 6 \).
e. Show that \( \nabla^2 V_1(r, \theta, \phi) = 0 \). (Note: Except at \( r = z = 0 \))
f. Evaluate the Laplacian of \( V_4(r, \theta, \phi) \), \( \nabla^2 V_4(r, \theta, \phi) \).

8.26. Create four three-component vectors \( \mathbf{A}, \mathbf{B}, \mathbf{C}, \text{ and } \mathbf{D} \) with statements like\(^{59}\)

\[
\mathbf{A} = \{\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3\}; \quad \mathbf{B} = \{\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3\}; \quad \mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3\}; \quad \mathbf{D} = \{\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3\};
\]

Then, using the file \texttt{crossdot.m} as described in Section 8.11, show that

a. \( \mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A} \), or, equivalently, that \( \mathbf{A} \times \mathbf{B} + \mathbf{B} \times \mathbf{A} = 0 \).
b. \( \mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C})\mathbf{B} - (\mathbf{A} \cdot \mathbf{B})\mathbf{C} \)
c. \( (\mathbf{A} \times \mathbf{B}) \times \mathbf{C} = (\mathbf{A} \cdot \mathbf{C})\mathbf{B} - (\mathbf{B} \cdot \mathbf{C})\mathbf{A} \)
d. \( \mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C} \)
e. \( (\mathbf{A} \times \mathbf{B}) \cdot (\mathbf{C} \times \mathbf{D}) = (\mathbf{A} \cdot \mathbf{C})(\mathbf{B} \cdot \mathbf{D}) - (\mathbf{A} \cdot \mathbf{D})(\mathbf{B} \cdot \mathbf{C}) \)

\(^{59}\)The variables \( \mathbf{C} \) and \( \mathbf{D} \) in \textsc{matemathica} are protected.
Chapter 9

Introduction to Programming

In this chapter, we address the task of composing sets of statements\(^1\) that will cause an obedient, efficient, and very literal servant (i.e., computer) to perform exactly the task intended by the master (i.e., programmer). A general strategy by which a particular task can be accomplished in a finite (though perhaps large) number of steps is called an *algorithm*. A specific expression of an algorithm in a suitable language is called a *program*. Hence, this chapter is about designing algorithms and implementing them in programs. More specifically, it is about designing algorithms for the performance of tasks that will ultimately be assigned to a computer and about implementing those algorithms in several possible computer languages.

Designing algorithms and implementing them in programs that direct a computer to perform various tasks is really not very different from designing algorithms and implementing them in programs that direct a baker to bake a cake, a knitter to knit a sweater, or a cab driver to drive from the airport to the hotel. In the aggregate, an algorithm will obtain all necessary *inputs*, manipulate those inputs in some way, and produce the required *outputs*. Each step in the process must be systematically specified using elementary statements, each of which means only one thing to the servant (person or computer) that will perform the task. Some of the appropriate elementary statements—particularly the *action statements* that specify simple actions—will vary with the general type of task to be performed. “Mix” and “whip”, for example, are among the elementary action statements that must be understood by the baker of cakes; “knit”, “purl”, and “cast on” must be part of the vocabulary of a knitter of sweaters; while “turn right” and “follow interstate 90 west” are action statements for the cab driver.

The performance of even simple tasks, however, entails the execution of suitable action statements *in the right order*. The complete description of an algorithm must therefore indicate not only which actions are to be performed but also the order of their performance. Those aspects of an algorithm that specify the order of performance of action statements are called *control structures*.

A language for specifying algorithms must therefore provide not only elementary action statements, which vary with the general type of task to be performed, but also control structures, which are more universal than the action statements. In the first two sections of this chapter, we focus particularly on identifying fundamental control structures and illustrating their role in several simple algorithms. In later sections, we explain how the general structures introduced in the first section are implemented in standard languages. More detailed information about various languages may be found in the references listed in Section 9.17.

\(^1\)Individual instructions in a computer language are usually called *statements*, and we use that word throughout this book.
9.1 Components of a Programming Language

While we could code programs directly in the binary language that computers use internally, the process would be slow, tedious, prone to error, and resistant to debugging. As an alternative, a wide assortment of high-level languages has been developed. Using a chosen high-level language, the programmer creates an ASCII file containing the source code, which lays out what the computer must do to accomplish the desired task. Then, that source code must be translated into machine code—a task that a computer equipped with the proper translating program can do for itself.

Actually, there are two approaches to this translation. In the first, the computer interprets the source code, which means that each statement in the code is translated by the computer each time the statement is executed; execution is slow but the process for us is simple: we write the source code and run the program. In the second approach, the computer first compiles the source code into binary object code, and then links the object code with an assortment of system routines and possibly additional user-written routines. Only after compilation and linking have been completed can the resulting binary or executable file actually be run. This second approach is more complicated for us, since we must both write the source code and then compile and link it. The advantages are twofold: (1) the final executable file is stored in the machine and can be run any number of times without repeating compilation and linking; and (2) execution of compiled code is faster than execution when the code is interpreted.

The prospective user of a computer need only learn how to construct the source code in whatever language is to be used. While the detailed syntax of the statements that can be constructed depends very much on the language chosen, certain common elements can be identified. In particular, all languages provide ways to make use of computer memory, to obtain input and display output, and to control the flow of execution within the program. In this section, we describe those common elements as a prelude to laying out the details for specific languages in later sections.

9.1.1 Variables, Variable Names, and Use of Memory

Computers provide a means for working with data. The data, however, must be accessible to the central processing unit (CPU), which is the part of the hardware that directs and controls all actions carried out by the computer. Let us, therefore, endow the CPU with a capacity to reserve an available memory cell (or a contiguous succession of them if one cell is not large enough) and assign to it (them) a variable name, which we specify. Throughout the execution of the program, the assigned name then refers to this unique cubbyhole in the machine’s memory.

While the details vary from language to language, we suppose here that we can choose variable names freely, subject only to the constraints that

- the first character must be a letter \(^2\) (a–z; A–Z),
- subsequent characters can be chosen from the letters (a–z; A–Z), the digits (0–9), and the underscore character (_), and
- the several special words (keywords, e.g., IF, DO, UNTIL, ...) that we shall introduce in subsequent sections are reserved (or protected) and cannot be used as variable names.

While we do not explicitly impose a length limit on variable names, shorter names are preferred, simply because longer names take longer to type, take more space in the line, and, taken to extremes, result in program listings that are difficult to read. Generally, of course, we want to choose variable names that help us remember the significance of the quantity the name identifies.

\(^2\)Some languages are case-sensitive and others are not. For our general discussion, we shall assume case insensitivity, so a and A are indistinguishable in variable names.
Even in the generic discussion of this section, however, we shall pay attention to the data type of each quantity represented. In Chapter 1, we distinguished floating point numbers, integers, and character strings from one another in several ways. In structuring programs, we must remain always aware of this distinction. Further, the computer in interpreting or compiling our source code must either make assumptions or be told the data type of each variable we use. Some languages (e.g., Pascal, C) require an explicit declaration of variable names and associated data types before the name can be used. Other languages (e.g., BASIC, FORTRAN) exploit implicit data typing by looking to the composition of the variable name to determine the intended data type—though more recent versions of FORTRAN, for example, also admit explicit data typing. Still others (e.g., IDL, MATLAB, OCTAVE) assign a data type dynamically and automatically on the basis of the value assigned to the variable. In our general discussion, we shall adopt a convention that keeps us consciously aware of data type every time we use a variable name. Following the pattern actually used in BASIC, we shall take variable names with no appended suffix to represent floating point numbers, variable names with an appended percent sign \% to represent integers, and variable names with an appended dollar sign $ to represent character strings. Occasionally, we shall use an appended at sign @ to stand for any one of these three possibilities (no suffix, \%, or $). Thus, in the pseudocode we are defining, the names x, position, and field identify memory cells for floating point values, the names I%, count%, and lower_limit% are valid names for cells storing integers, the names word$ 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

```
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 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.\footnote{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.} We shall use the following special symbols and words for these operations:

- Assign a value to a named memory location:

  \[(\text{variable}) \leftarrow (\text{expression determining value to be assigned})\]

  e.g.,\footnote{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).}

  \begin{align*}
  \text{val} & \leftarrow \sin(\pi x) \\
  \text{count}\% & \leftarrow \text{count}\% + 1\% \\
  \text{first}\_\text{name}\$ & \leftarrow "David"
  \end{align*}

  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 (\textit{sin}, \textit{cos}, \textit{atan}, \textit{sqrt}, \textit{exp}, ...) to facilitate scientific computation.

- Obtain a value from the keyboard and store it in a (named) memory cell:

  \begin{verbatim}
  READ PROMPT="Enter first name: ", \textit{first}\_\text{name}\$
  \end{verbatim}

- Transmit a value from a (named) memory cell and/or a (quoted) literal message to the screen:

  \begin{verbatim}
  WRITE "The value of the integral is ", \textit{value}, "."
  \end{verbatim}

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.
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>True 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 NOT ⟨condition⟩ which we define to be true if ⟨condition⟩ is false and false if ⟨condition⟩ 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

EQUAL? ←− N? = M?

in response to which the computer will assess whether N? is equal to M? and set EQUAL? to true or false, depending on the outcome of that assessment. With this expansion of our language, conditions might be expressed not only by the comparisons illustrated in Table 9.1 but also by the
simple assertion of a single Boolean variable or by a logical expression involving combinations of comparisons and/or assertions of Boolean variables. Note, incidentally, that the statement in this paragraph reveals why computer languages must have a different symbol for assignment than for testing equality.\(^5\)

### 9.1.4 Syntactic Wrinkles

In some programming contexts, we will find it convenient—and sometimes even necessary—to convey additional information about the *structure* of the code to the CPU. We might wish

- to place more than one statement on a single physical line. In our generic code, we shall use the character ‘;’ to separate individual statements on a single line. The three lines of code early in Section 9.1.2 might be combined in one line with the coding
  \[
  \text{val} \leftarrow \sin(\pi \times \text{x}); \quad \text{count}\% \leftarrow \text{count}\% + 1\%; \quad \text{first\_name}\$ \leftarrow \text{"David"}
  \]
- to spread a single statement over more than one physical line. In our generic code, we shall use the symbol \(\rightarrow\) at the end of a line to indicate that the statement continues on the next line. Thus, for example, the two-line statement
  \[
  U(I\%, J\%) \leftarrow 0.25 \times (U(I\%+1\%, J\%) + U(I\%-1\%, J\%) \rightarrow \quad + U(I\%, J\%+1\%) + U(I\%, J\%-1\%))
  \]
  is to be seen as a single statement.\(^6\)
- to group several statements together to form a *block* of statements so that they can be properly treated when the block is placed in a context in which the interpreter or compiler is expecting a single statement. In our generic code, we shall use the keywords \texttt{BEGIN\_BLOCK} and \texttt{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
  \[
  \text{BEGIN\_BLOCK}
  \begin{align*}
  \text{val} & \leftarrow \sin(\pi \times \text{x}) \\
  \text{count}\% & \leftarrow \text{count}\% + 1\% \\
  \text{first\_name}\$ & \leftarrow \text{"David"}
  \end{align*}
  \text{END\_BLOCK}
  \]
  would group the three statements as a unit.

Different languages differ significantly in the way these three situations are coded. We shall be more explicit in subsequent sections as we address each language in turn.

### 9.1.5 Essential Control Structures

Four main control structures\(^7\) are generally provided in standard computer languages and figure significantly in the smooth expression of algorithms:

- **Sequence**, which is conveyed by the order of the statements in the algorithm.

---

\(^5\)We have used \(\leftarrow\) and \(\times\). In later sections, we will identify the symbols used in other languages.

\(^6\)Remember that spaces not in quoted strings are ignored by the compiler.

\(^7\)Since the \texttt{CASE} structure can be expressed in terms of nested \texttt{IF-THEN-ELSE} structures and the \texttt{IF-THEN-ELSE} structure is simply a \texttt{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.
9.1. COMPONENTS OF A PROGRAMMING LANGUAGE

- Two-way selection, expressed generically with a statement like
  \[
  \text{IF } \langle \text{condition} \rangle \text{ THEN } \langle \text{block1 of statements} \rangle \text{ ELSE } \langle \text{block2 of statements} \rangle \text{ END_IF}
  \]
  Here \langle block1 of statements \rangle is executed when \langle condition \rangle is true and \langle block2 of statements \rangle is executed when \langle condition \rangle is false. The \text{ELSE} clause may be omitted altogether if \langle block2 of statements \rangle is null. The flow diagram in Fig. 9.1(a) is often used to convey this structure.

- Multi-way selection
  \[
  \text{CASE } \langle \text{condition1} \rangle \text{ DO } \langle \text{block1 of statements} \rangle \text{ OF } \langle \text{condition2} \rangle \text{ DO } \langle \text{block2 of statements} \rangle \text{ OF } \langle \text{condition3} \rangle \text{ DO } \langle \text{block3 of statements} \rangle \ldots \text{ OF OTHERS DO } \langle \text{block0 of statements} \rangle \text{ END_CASE}
  \]
  In executing this overall statement, the computer will test each condition in turn, execute the block of statements associated with the first true condition it encounters, and then jump out of the \text{CASE} structure. The \text{OTHERS} clause may be omitted altogether if \langle block0 of statements \rangle is null. The flow diagram in Fig. 9.1(b) is often used to convey this structure.

- Repetition
  \[
  \text{LOOP } \langle \text{block1 of statements} \rangle \text{ EXIT_LOOP WHEN } \langle \text{condition} \rangle \langle \text{block2 of statements} \rangle \text{ END_LOOP}
  \]
  Here, the statements in the body of the loop are executed repeatedly until \langle condition \rangle, which is tested at the indicated point in the loop, becomes \text{true}. Looping will continue forever unless repeated execution of the statements eventually causes \langle condition \rangle to become \text{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 \text{X} containing \text{N} elements and we wish to create a second vector \text{Y}, each of whose elements is—say—the sine of the corresponding element in \text{X}. In the coding we have so far described, we would achieve this objective with an explicit loop involving the statements

\[
\text{CNT} \leftarrow 0 \text{! Set CNT to 0}
\]

The structure of this single statement is sufficiently distinctive—it is not complete until the keyword \text{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.

9 We use the exclamation point to introduce comments. Judicious use of such comments, irrelevant though they may be to the computer, can clarify the algorithm for the human reader of the listing. To be useful, a comment should amplify the meaning of, or clarify the role of, the commented statement. The statement

\[
\text{CNT} \leftarrow 0 \text{! Set CNT to 0}
\]

for example, might as well be uncommented, since the comment really says no more than can be inferred from the statement itself.

10 As with the previous structure, the structure of this single statement is sufficiently distinctive—it is not complete until the keyword \text{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.

11 Not all compilers adopt this strategy. Some will execute the blocks associated with every true condition in the structure.
(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

\[
Y \leftarrow \sin( X )
\]

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 \texttt{THEN} and \texttt{ELSE} clauses, the statements between \texttt{CASE} and \texttt{END\_CASE}, or the body of the loop, the (outer) \texttt{IF-END\_IF} statement, the (outer) \texttt{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 \texttt{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.,

\[
\text{WHILE} \langle \text{condition} \rangle \text{ DO } \langle \text{block of statements} \rangle
\]

and the block is not executed at all if \langle condition \rangle is \texttt{false} when the loop is entered. In the second version, the test is at the end of the loop, e.g.,

\[
\text{REPEAT} \langle \text{block of statements} \rangle \text{ UNTIL } \langle \text{condition} \rangle
\]

and the block will be executed at least once, whether \langle condition \rangle is \texttt{true} or \texttt{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.,

\[
\text{FOR } I\% \leftarrow \texttt{IMIN}\% \text{ THRU } \texttt{IMAX}\% \text{ DO } \langle \text{block of statements} \rangle
\]

When this loop is executed, the body of the loop is executed for \texttt{I\%} having the value \texttt{IMIN}\%, then for \texttt{I\%} having the value \texttt{IMIN}%+1%, then for \texttt{I\%} having the value \texttt{IMIN}%+2%, and so on, continuing until the loop has been executed for the largest value of \texttt{I\%} that does not exceed \texttt{IMAX}\%. The more sophisticated form

\[
\text{FOR } I\% \leftarrow \texttt{IMIN}\% \text{ THRU } \texttt{IMAX}\% \text{ STEP } \texttt{INC}\% \text{ DO } \langle \text{block of statements} \rangle
\]
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 \( I_{\text{MIN}} \) equals \( I_{\text{MAX}} \) and not at all if \( I_{\text{MIN}} \) exceeds \( I_{\text{MAX}} \)—but beware; different languages may behave differently in this regard.

Though it is not essential, one further control structure is standard in all scientific programming languages. Conveyed in our generic code by the statement

\[
\text{EXECUTE } \langle \text{procedure name} \rangle
\]
this single statement makes possible the invocation of a (properly defined) procedure in a program (or for that matter in another procedure); it both facilitates a modular approach to the design of programs and provides a means to avoid duplication of program modules that must be invoked more than once to specify a task completely.

9.2 Sample Short Algorithms

Large algorithms for accomplishing complex tasks are frequently constructed by combining various smaller algorithms. In this section, we enumerate several program fragments that may serve as building blocks for the construction of larger algorithms.

1. Exchange the values stored in A@ and B@:\(^{12}\)

\[
\begin{align*}
\text{TEMP} &\leftarrow A@ \\
A@ &\leftarrow B@ \\
B@ &\leftarrow \text{TEMP@}
\end{align*}
\]

The temporary memory cell TEMP@ saves the original value of A@ so that it is still retrievable after the value of B@ has been copied into cell A@, overwriting the original contents of cell A@.

2. Stopping a loop with a sentinel [and counting]:

\[
\begin{align*}
\text{SENTINEL@} &\leftarrow \langle\text{agreed-upon special value}\rangle \\
[\text{COUNT%} &\leftarrow 0%]\end{align*}
\]

LOOP

\[
\begin{align*}
\text{READ ITEM@} \\
\text{EXIT LOOP WHEN ITEM@ = SENTINEL@} \\
\langle\text{Statements processing ITEM@}\rangle \\
[\text{COUNT%} &\leftarrow \text{COUNT%} + 1%]
\end{align*}
\]

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:

\[
\begin{align*}
\text{NUMBER OF TIMES%} &\leftarrow \langle\text{desired number of executions}\rangle \\
\text{COUNTER%} &\leftarrow 0% \\
\text{LOOP}
\end{align*}
\]

\[
\begin{align*}
\text{EXIT LOOP WHEN COUNTER% = NUMBER OF TIMES%} \\
\langle\text{block of statements}\rangle \\
\text{COUNTER%} &\leftarrow \text{COUNTER%} + 1%
\end{align*}
\]

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:

\(^{12}\)Note again the use of the character ! to flag comments. See footnote 9.
TIMES_REMAINING% ← (desired number of executions)
LOOP
   EXIT_LOOP WHEN TIMES_REMAINING% = 0%
   (block of statements)
   TIMES_REMAINING% ←− TIMES_REMAINING% - 1%
END_LOOP

This coding has a small advantage over the previous coding because it requires only one variable (TIMES_REMAINING) to control the loop. Note, however, that the initial value of that variable is irrecoverably lost by the time execution of the loop has been completed.

5. Summing [and counting]:

SENTINEL ← (agreed-upon special value)
SUM ←− 0.0
[COUNT% ←− 0%]
LOOP
   READ ITEM
   EXIT_LOOP WHEN ITEM = SENTINEL
   SUM ←− SUM + ITEM
   [COUNTR% ←− COUNTR% + 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:

```plaintext
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:

```plaintext
ITEM_Sought@ ← ⟨item to be found⟩
NO_ELEMENTS% ← ⟨number of items in ITEMS@⟩
POINT% ← 0% ! Start at beginning
LOOP
  POINT% ← POINT% + 1% ! Advance pointer to next item
  EXIT_LOOP WHEN ITEMS@(POINT%) = ITEM_Sought@ OR POINT%=NO_ELEMENTS%
END_LOOP
```

We assume that, by the time this procedure is invoked, the list of items to be examined has been stored in the one-dimensional array `ITEMS` and that we know the number of elements in that array. Note that the pointer `POINT%` is stepped through the records one at a time as the search unfolds. Finally, when this loop terminates, the item sought has been found if and only if the value in the record identified by `POINT%` matches the item sought.

### 9.3 Two Larger Algorithms

In this section, we discuss two algorithms in some detail, partly to illustrate the general features described in Section 9.1 more fully and partly to lay out the algorithms generically before we implement them in specific languages.

#### 9.3.1 Solving Laplace’s Equation

Among the more important equations in mathematical physics, Laplace’s equation appears in the study of electromagnetic fields, steady state heat flow, fluid mechanics, and many other contexts. In two-dimensions and in Cartesian coordinates \((x, y)\), the equation assumes the form

\[
\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0
\]  

(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° and the fourth edge is maintained at 100° or the electrostatic potential within the square when three edges are maintained at a potential of 0 Volts and the fourth edge is maintained at a potential of 100 Volts.

The basis for a simple algorithm for solving this problem numerically involves imposing an $N \times N$ regular grid of points $(x_i, y_j)$, $1 \leq i, j \leq N$, on the region. Then, we declare that we have found a solution when we know appropriate values for $U_{i,j} = U(x_i, y_j)$ at each grid point. The simplest boundary conditions, of course, tell us the values at all points for which $i$ and/or $j$ is either 1 or $N$. The values we seek for the other points ($2 \leq i, j \leq N - 1$) must in some sense reflect the differential equation. If, however, we know the values of $U$ at three consecutive points along a line parallel to the $x$ axis, say, we can use the approximation (see exercises)

$$\frac{\partial^2 U}{\partial x^2} \approx \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{\Delta x^2}$$

(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

$$U_{i,j} \approx \frac{1}{4} \left( U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1} \right)$$

(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 \textit{when} satisfactory convergence has been achieved, however, is difficult. Because we are here
focusing on programming aspects rather than algorithmic refinements, we shall ignore the issue of convergence for the
moment. We shall return to it at appropriate points in later chapters.} Supposing that we seek a solution on a $15 \times 15$ grid
of points, we make a first pass at constructing an algorithm to implement this overall strategy by
listing the statements

\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 ITCN% ← 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 ITCN% ← 1% THRU 30% DO Conduct one iteration
  FOR I% ← 2% THRU 14% DO
    FOR J% ← 2% THRU 14% DO U(I%,J%) ← 0.25 \times ( 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.

\subsection{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.
Table 9.3: Sample data file having the structure described in Section 9.3.2. The line numbers in the first column and the comments in the third column of this table are not in the actual file.

<table>
<thead>
<tr>
<th>Line Number</th>
<th>Line in file</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>01:</td>
<td>Example Data File; Author: DMC; Date: 7-21-01</td>
<td>ID of file</td>
</tr>
<tr>
<td>02:</td>
<td>First line of comments.</td>
<td></td>
</tr>
<tr>
<td>03:</td>
<td>Second line of comments.</td>
<td></td>
</tr>
<tr>
<td>04:</td>
<td>Third line of comments.</td>
<td></td>
</tr>
<tr>
<td>05:</td>
<td>Fourth line of comments.</td>
<td></td>
</tr>
<tr>
<td>06:</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>07:</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>08:</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>09:</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10:</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>11:</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>12:</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>13:</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>14:</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>15:</td>
<td>0.57</td>
<td></td>
</tr>
</tbody>
</table>

At base, we imagine that the data files we create will store some number \( n \) of identically dimensioned three-dimensional arrays \( A_{ijk} \) whose three indices will assume \( n_x, n_y, \) and \( n_z \) values,\(^{15}\) respectively, though at times \( n_z \) will have the value 1 so that the storage of two-dimensional arrays can be accommodated. We adopt the following standard format for all data files:

- five lines of comments, possibly including title, author, and date;
- one line specifying the number \( n \) of arrays in the file;
- one line specifying the number \( n_x \) of values assumed by the first index \( i \) of \( A_{ijk} \);
- one line specifying the number \( n_y \) of values assumed by the second index \( j \) of \( A_{ijk} \);
- one line specifying the number \( n_z \) of values assumed by the third index \( k \) of \( A_{ijk} \); and finally
- \( n_xn yn_z \) 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 \( n_* - 1 \); in other languages, the values will range from 1 to \( n_* \). (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

\begin{verbatim}
DIMENSION ARRAY(n_x,n_y)       ! Prepare array for values
NARR% ←− 1%                     ! Set number of arrays
NX% ←− n_x                     ! Set number of grid points in each coordinate
NY% ←− n_y
NZ% ←− 1%
DX ←− (x_{\text{max}} - x_{\text{min}})/(n_x - 1)  ! Set increments
DY ←− (y_{\text{max}} - y_{\text{min}})/(n_y - 1)
\end{verbatim}

Then, we evaluate the function $f(x,y)$ with the double loop expressed in the code

\begin{verbatim}
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
END
\end{verbatim}

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\footnote{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.}\footnote{Be aware that some languages reserve a few channel numbers for special "files", e.g., the keyboard or the screen.}

\begin{verbatim}
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, "***"
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
\end{verbatim}

Here, the first statement identifies the file and assigns to it a number—any number will do\footnote{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.}—to be used for subsequent reference to the file. Next, the first five WRITE statements send five lines of
comments, all five of which must be physically present even if fewer than five are needed to contain necessary information about the file. Then, the sixth WRITE statement writes the number of arrays to the file, the seventh and eighth write the $x$ and $y$ dimensions to the file, and the ninth writes the $z$ dimension to the file. Finally, the double loop writes the elements of the array to the file in the proper order, and the last statement makes sure the file is properly detached from the program so that it can be accessed by other programs.

Instead of writing data to a file, we frequently will need to read data from a file, perhaps as a means to import data generated by one program for use in a different program. Slight modification of the ATTACH and READ statements introduced earlier provide for that action. Specifically, the statements

```
ATTACH FILE ⟨filename⟩ ON CHANNEL 1 FOR READING
READ FROM CHANNEL 1, Comma-separated list of variables for storage of values
;
CLOSE FILE ON CHANNEL 1
```

Here, the first statement opens an existing file for reading (and displays an error if the specified file does not exist), the remaining statements save the last read the information in the file and stores it in the specified variables, and the last statement detaches the file when, presumably all data have been read. Note that the structure of the READ statements must reflect the structure of the file, and the variables specified must have the data types of the values to be read. Normally, the READ statement reads a line at a time, but in some languages each READ statement reads as many values as needed to complete the list of variables, however these values are distributed over lines.

In the remaining sections of this chapter, we shall implement the two algorithms described in this section in different languages and, especially for the second algorithm, in a variety of different contexts.

### 9.8 Solving Laplace’s Equation with IDL

The algorithm described in Section 9.3.1 is readily implemented in IDL with the statements

```
u = fltarr(15,15) & u[14,*]=100.0
for itcnt = 1, 30 do$
   for i = 1, 13 do$
      for j = 1, 13 do$
         u[i,j]=0.25*(u[i+1,j]+u[i-1,j]+u[i,j+1]+u[i,j-1])
```

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.\(^{18}\) Then, the remaining statement (which contains three nested for loops and occupies four physical lines) executes 30 iterations (the outermost for loop) of the double loop it contains. The innermost statement simply expresses the substance of Eq. (9.3). Note that the inner loops do not include points on the boundary of the region in which a solution is sought, so the values in the array $u$ at its “edges” are left alone throughout the 30 iterations.

Once the solution has been completed, of course, we could display it on the screen with any number of statements, the simplest of which probably are

```
IDL> surface, u
```

\(^{18}\)Remember that indices in IDL start at 0, so a 15-element array has indices running from 0 to 14.
though these could be embellished by exploiting optional arguments and a variety of keywords to rescale the axes, 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.

9.12.3 ... with IDL

Note: All IDL programs (*.pro) and all IDL-created data files (*.idl.dat) in this chapter can be copied from the directory $HEAD/idl, where (as defined in the Local Guide) $HEAD must be replaced by the appropriate path for your site.

In IDL, creation of an internal array containing the values of the function of Eq. (9.4) is facilitated by the Lawrence-written IDL procedure \texttt{lugen.grid}. As described in Section 2.12.1, 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\(^20\)

IDL> \texttt{lugen.grid, x, xrange=[-3.0*!pi, 3.0*!pi], nx=49, y, yrange=[-3.0*!pi, 3.0*!pi], ny=49}

IDL> I = ( (\sin(x)/x)^2 ) * ( (\sin(y)/y)^2 )

This array can be written to a file having the format described in Section 9.3.2 by using the Lawrence-written pro-file \texttt{lugen.list.pro}. The statement\(^21\)

IDL> \texttt{lugen.list, I}

calls \texttt{lugen.list} and tells IDL to write the array \( I \), which must already exist in the IDL workspace, to a file. When this command is entered, a window similar to that in Fig. 9.4 appears. The user

\(^{19}\)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.

\(^{20}\)Remember that, in IDL, array indices start at 0, so division of the interval into 49 segments will result in 50 grid points along each axis and in indices running from 0 to 49.

\(^{21}\)With indices starting at zero, the entry in the upper left corner of the array \( I \) is the (0, 0) entry. Further—and this wrinkle is extremely important—remember that values for which the first index of a two-dimensional array in IDL stays fixed are in a column while values for which the second index stays fixed are in a row. That is, as IDL views two-dimensional arrays internally, the first index is the column index and the second index is the row index. IDL is entirely consistent in this internal interpretation, but we must be careful both in writing internal arrays out into files and—later—in reading those files into internal IDL arrays. Otherwise, we may effect an inadvertent transposition of the array. The procedure \texttt{lugen.list} pays proper attention to this wrinkle.
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 by

\[
p(x, y, z) = \frac{R_{nlm}(x, y, z)}{N} = \frac{1}{\pi^6} \left( \frac{2}{3} \right)^{3/2} \frac{1}{2l+1} \frac{1}{4\pi^2} \frac{1}{N} \sum_{m=-l}^{l} P^l_m \sin(m\phi) \sin(l\theta) 
\]

where \( R_{nlm}(x, y, z) \) is the radial part of the electron wave function and \( N \) is a normalization constant.

The default filename is ‘untitled.dat’ and the default comments are ‘**’. The user should edit these entries as appropriate.

---

\(^{22}\) The default filename is ‘untitled.dat’ and the default comments are ‘**’. The user should edit these entries as appropriate.

\(^{23}\) A Lawrence-written IDL procedure for reading files written by `lugen_list.pro` is described in Section 9.16.3.
density is given as a function of Cartesian coordinates \((x, y, z)\) with the nucleus located at the origin by

\[
p(x, y, z) = \frac{1}{2\pi(27)^3} e^{-2\rho/3} \left( \frac{3z^2}{\rho^2} - 1 \right)^2 = \frac{1}{2\pi(27)^3} e^{-2\rho/3} (9z^4 - 6z^2\rho^2 + \rho^4)
\]  \tag{9.5}

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

\[
\rho = \sqrt{x^2 + y^2 + z^2}
\]  \tag{9.6}

This function of three variables is commonly visualized either by focusing on the function in various planes intersecting the three-dimensional volume (thereby reducing the display to a family of two-dimensional displays) or by displaying various contour surfaces. Programs producing such displays need a three-dimensional array of values as input. Suppose, then, we anticipated using some graphical visualization program to explore the quantum probability given by Eq. (9.6). Let us decide to determine values of \(p(x, y, z)\) over the region \(-10 \leq x, y, z \leq 10\), dividing each axis into 29 segments, 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.

9.13.3 ... with IDL

The procedure for using IDL to generate three-dimensional scalar arrays is similar to the method described in Section 9.12.3. The array representing the probability density in Eq. (9.5) is created using the IDL statements

\[
\begin{align*}
&\text{IDL} > \text{lugen\_grid, x, xrange=[-10.0, 10.0], nx=29, } \\
&\text{IDL} > \quad \quad \text{y, yrange=[-10.0, 10.0], ny=29, } \\
&\text{IDL} > \quad \quad \text{z, zrange=[-10.0, 10.0], nz=29} \\
&\text{IDL} > \quad \quad \text{P = fltarr(30,30,30)} \\
&\text{IDL} > \quad \quad \text{pfactor = 1.0/(2.0*!pi*27.0*27.0*27.0)} \\
&\text{IDL} > \quad \quad \text{zs = z*z} \\
&\text{IDL} > \quad \quad \text{rhos = x^2 + y^2 + zs} \\
&\text{IDL} > \quad \quad \text{rho = sqrt(rhos)} \\
&\text{IDL} > \quad \quad \text{P = pfactor * exp(-2*rho/3) * (9*zs^2 - 6*zs*rhos + rhos^2)} \\
\end{align*}
\]

The first three lines produce a grid of values for the evaluation of the scalar field.\(^{24}\) Here, \(x\), \(y\), and \(z\) each ranges from \(-10.0\) to \(10.0\), and the interval is divided into 29 segments. The fourth line creates a three-dimensional \(30 \times 30 \times 30\) array \(P\). The remaining lines evaluate several intermediate quantities and then evaluate \(P\). Finally, with the statement

\[
\text{IDL} > \text{lugen\_list, P}
\]

we write the data into a file in the desired format.\(^{25,26}\) After the filename—we here choose \texttt{pdensidl.dat}\(^{27}\)—and optional information have been typed in the \texttt{lugen\_list} window (see Fig. 9.4) and ML has been clicked on the ‘SAVE’ button, the data file will be saved in the default directory.

\(^{24}\)We here use the optional arguments to \texttt{lugen\_grid}. These arguments specify ranges on the \(z\) coordinate. Their presence triggers the creation of three-dimensional arrays containing the \(x\), \(y\), and \(z\) coordinates of a grid of points in three dimensions.

\(^{25}\)The procedure \texttt{lugen\_list} invoked here is identical for both two- and three-dimensional scalar and vector arrays and is described in Section 9.12.3.

\(^{26}\)A Lawrence-written IDL procedure for reading files written by \texttt{lugen\_list.pro} is described in Section 9.16.3.
9.14 Creating and Storing Two-Dimensional Vector Arrays

Display of a \emph{vector} field in two dimensions requires a two-dimensional \emph{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 \emph{pair} of two-dimensional \emph{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. 9.5, we take the waveguide to be oriented so that the wave propagates in the positive $x$ direction. Let the guide have a rectangular cross-section with dimensions $(b,d)$, i.e., $(0 < y < b, 0 < z < d)$. Further, let the wave be polarized in the $z$ direction and consider the particular TE waves whose electric field does not depend on $z$. In mksa units, the (complex) fields in this guide are given by

$$
\mathbf{E}(\mathbf{r},t) = A \sin \frac{n \pi y}{b} e^{i(n \pi y/x - \omega t)} \mathbf{k} \tag{9.7}
$$

$$
\mathbf{H}(\mathbf{r},t) = \nabla \times \mathbf{E} = A \omega \mu_0 \left( \frac{n \pi}{b} \cos \frac{n \pi y}{b} \mathbf{i} - i \kappa x \sin \frac{n \pi y}{b} \mathbf{j} \right) e^{i(n \pi y/x - \omega t)} \tag{9.8}
$$

where $n = 1, 2, 3, \ldots$ and, with $c$ standing for the speed of light,

$$
\kappa^2 = \frac{\omega^2}{c^2} - \left( \frac{n \pi}{b} \right)^2 \tag{9.9}
$$

If we focus on the physical fields (real parts of $\mathbf{E}$ and $\mathbf{H}$) specifically at $t = 0$, we find that

$$
\mathbf{E}(\mathbf{r}) = A \cos \kappa x \sin \frac{n \pi y}{b} \mathbf{k}
$$

$$
\mathbf{H}(\mathbf{r}) = \frac{An}{\omega \mu_0 b} \sin \kappa x \cos \frac{n \pi y}{b} \mathbf{i} - \frac{A\kappa x}{\omega \mu_0} \cos \kappa x \sin \frac{n \pi y}{b} \mathbf{j} \tag{9.10}
$$

\footnote{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.}
To simplify these equations further, we divide \( E \) by \( A \) and \( H \) by \( \frac{A\pi/\omega_\mu_0 b}{b} \), and we elect to measure lengths in units of \( b \) by introducing the variables \( y/b = y \) and \( x/b = x \). Equation (9.10) then becomes

\[
E(r) = \frac{E(r)}{A} = \cos \kappa_x b x \sin n\pi y \hat{k}
\]

\[
H(r) = \frac{H(r)}{\frac{A\pi/\omega_\mu_0 b}{b}} = \sin \kappa_x b x \cos n\pi y \hat{i} - \frac{\kappa_x b}{n\pi} \cos \kappa_x b x \sin n\pi y \hat{j}
\]

and we find that the field of interest depends on two parameters \( n \) and \( \kappa_x b \). Further, Eq. (9.9) imposes the constraint

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

on these parameters. The only non-zero components of the fields in Eq. (9.11) are

\[
\begin{align*}
E_z &= \cos \kappa_x b x \sin n\pi y \\
H_x &= \sin \kappa_x b x \cos n\pi y \\
H_y &= -\frac{\kappa_x b}{n\pi} \cos \kappa_x b x \sin n\pi y
\end{align*}
\]

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

\[
\begin{align*}
E_z &= \cos 2\pi x \sin 2\pi y \\
H_x &= \sin 2\pi x \cos 2\pi y \\
H_y &= -\cos 2\pi x \sin 2\pi y
\end{align*}
\]

where \( x \) can range over any values—we choose \( 0 \leq x \leq 1 \)—but, to be inside the guide, \( y \) is confined to the region \( 0 \leq y \leq 1 \). Each component of these fields can now be represented by a two-dimensional array. The \( H \) 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. To be explicit, we determine values of \( H_x(x,y) \), \( H_y(x,y) \), and \( E_z(x,y) \) over the region \( 0 \leq x, y \leq 1 \), dividing each axis into 30 segments of length 1/29, 0.0, which will entail evaluating the field components at 30 values of \( x \) and 30 values of \( y \)—a total of \( 30 \times 30 = 900 \) values. Normally, these field components would be stored in three \( 30 \times 30 \) two-dimensional arrays. Because of our declared file format, however, we must view ourselves as needing storage for three \( 30 \times 30 \times 1 \) three-dimensional arrays. We shall, however, view the structure as a single \( 30 \times 30 \times 1 \times 3 \) array \( H \), with \( H(*,*,1,1) \) storing \( H_x \), \( H(*,*,1,2) \) storing \( H_y \), and \( H(*,*,1,3) \) storing \( E_z \).

9.14.3 ... with IDL

The procedure for generating two-dimensional vector arrays is very similar to the method described in Section 9.12.3. 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 IDL commands

---

28We have supposed indices starting at 1. In languages where indices start at 0, the associations would, of course, be \( H(*,*,0,0) \) storing \( H_x \), \( H(*,*,0,1) \) storing \( H_y \), and \( H(*,*,0,2) \) storing \( E_z \).
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

$$\frac{[E_x(x, y, z), E_y(x, y, z), E_z(x, y, z)]}{q/4\pi\varepsilon_0 a^2} = \left[ \frac{[x - 1, y - 1, z]}{[(x - 1)^2 + (y - 1)^2 + z^2]^{3/2}} - \frac{[x + 1, y - 1, z]}{[(x + 1)^2 + (y - 1)^2 + z^2]^{3/2}} \right] + \left[ \frac{[x + 1, y + 1, z]}{[(x + 1)^2 + (y + 1)^2 + z^2]^{3/2}} - \frac{[x - 1, y + 1, z]}{[(x - 1)^2 + (y + 1)^2 + z^2]^{3/2}} \right]$$

(9.15)

To be explicit, we determine values of the field components over the region $-2.0 \leq x, y, z \leq 2.0$, dividing each axis into 29 segments, which will entail evaluating the components at 30 values of x, 30 values of y, and 30 values of z—a total of $30 \times 30 \times 30 = 27000$ values. Because of our declared file

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 $30 \times 30 \times 1$ three-dimensional arrays.

The procedure `lugen_list` invoked here is identical for both two-dimensional scalar and vector arrays and is described in Section 9.12.3.

A Lawrence-written IDL procedure for reading files written by `lugen_list.pro` is described in Section 9.16.3.
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 \(E(*,*,*,1)\) 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.

9.15.3 ... with IDL

The procedure for using IDL to generate three-dimensional vector arrays is very similar to the method described in Section 9.14.3. 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 IDL statements

```idl
IDL> lugen_grid, x, xrange=[-2.0, 2.0], nx=29, $
IDL> y, yrange=[-2.0, 2.0], ny=29, $
IDL> z, zrange=[-2.0, 2.0], nz=29
IDL> E = fltarr(30,30,30,3)
IDL> xp = x + 1.0 & xm = x - 1.0 & yp = y + 1.0 & ym = y - 1.0
IDL> xps = xp*xp & xms = xm*xm & yps = yp*yp & yms = ym*ym
IDL> zs = z*z
IDL> rmm = (xms+yms+zs)^1.5 & rmp = (xms+yps+zs)^1.5
IDL> rpm = (xps+yms+zs)^1.5 & rpp = (xps+yps+zs)^1.5
IDL> E(*,*,*,0) = xm*(1.0/rmm - 1.0/rmp) + xp*(1.0/rpp - 1.0/rpm)
IDL> E(*,*,*,1) = ym*(1.0/rmm - 1.0/rmp) + yp*(1.0/rpp - 1.0/rpm)
IDL> E(*,*,*,2) = z*(1.0/rmm - 1.0/rmp + 1.0/rpp - 1.0/rpm)
```

The first three 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. The fourth line creates a four-dimensional array \(E\) consisting of three \(30 \times 30 \times 30\) three-dimensional arrays. 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

```idl
IDL> lugen_list, E
```

we write the data to a file in the desired format.\(^{33,34}\) After the filename—we here choose `quadpole_idl.dat`—and optional information have been typed in the `lugen_list` window (see Fig. 9.4) and ML has been clicked on the ‘SAVE’ button, 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.

\(^{32}\)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 \(E_z\).

\(^{33}\)The procedure `lugen_list` invoked here is identical for both two- and three-dimensional scalar and vector arrays and is described in Section 9.12.3.

\(^{34}\)A Lawrence-written IDL procedure for reading files written by `lugen_list.pro` is described in Section 9.16.3.
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

9.16.3 ... with IDL

The Lawrence-generated IDL function luread_list parallels the IDL function lugen_list first described in Section 9.12.3 and provides the facility for reading into IDL any file created in the format described in Section 9.3.2, whether the files were created with IDL or some other program. We simply execute the statement

luread_list, 'FileName', Var

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

IDL> luread_list, 'irrad_idl.dat', I

will recreate the variable I as it was when the file irradiidl.dat was created in Section 9.12.3. 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.pro that defines this function is presented in Section 9.Q.

9.17 References

For IDL, see Section 2.18.

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 9.6 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 $\text{LARGEST}_k$ 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.
9.18. EXERCISES

Table 9.5: Algorithm for Exercise 9.5.

START % ←⟨agreed-upon special value⟩
LARGEST % ←⟨assumed fictitious integer known to occur before any possible real integer in the list⟩
LOOP
READ ITEM %
EXIT_LOOP WHEN ITEM % = START %
IF ITEM % > LARGEST %
THEN LARGEST % ← ITEM %
END_IF
END_LOOP
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 $A % * A % + B % * B % = C % * C %$, subject to the restriction that $A %, B %,$ and $C %$ shall all be smaller than some value MAXNUM % supplied as input. Hint: Systematically examine all possibilities, but do so thoughtfully. For example, there is no point in examining cases for which $C % <= A %$ or $C % <= B %$. Express your loops so that these cases (and any others that you can reject a priori) will not even be considered. Optional: For MAXNUM % = 20%, determine the number of executions of your innermost loop.

9.9. Suppose you have % cards laid out in a row on a table. On each card is a single word. Determine the end result of applying the mystery procedure laid out in Table 9.6 to that array of cards and choose a suitable name for the procedure.

9.10. Let the digits in an integer be counted from the left end of the integer, i.e., in the four-digit number “4358”, call “4” digit 1, “3” digit 2, “5” digit 3, and “8” digit 4. Determine the function of the mystery procedure in Table 9.7 and choose a suitable name for the procedure.

9.11. Starting with the approximation

$$\frac{du}{dx} \bigg|_{x+\Delta x} \approx \frac{u(x+\Delta x) - u(x)}{\Delta x}$$

deduce the finite difference approximation

$$\frac{d^2u}{dx^2} \bigg|_{x} \approx \frac{u(x+\Delta x) - 2u(x) + u(x-\Delta x)}{\Delta x^2}$$

for the second derivative.

9.12. Write, compile, and test a program that asks for the input of a temperature in Celsius and prints out the corresponding temperature in Fahrenheit. To make it a bit more of a challenge, write the 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

```plaintext
PROCEDURE ?????
SCANEND% ← N%
LOOP
CARD% ← 1%
Obtain word on card CARD% and store in WORD$
LATEST_WORD$ ← WORD$
LATEST_CARD% ← CARD%
LOOP
CARD% ← CARD% + 1%
Obtain word on card CARD% and store in WORD$
IF WORD$ occurs after LATEST_WORD$
THEN BEGIN_BLOCK
  LATEST_WORD$ ← WORD$
  LATEST_CARD% ← CARD%
END_BLOCK
END_IF
EXIT_LOOP WHEN CARD% = SCANEND%
END_LOOP
Exchange card LATEST_CARD% with card SCANEND%
SCANEND% ← SCANEND% - 1%
EXIT LOOP WHEN SCANEND% = 1%
END_LOOP
END_PROCEDURE
```

given by the equation

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

where \( B_0 \) is the magnetic field at the origin. This field can be considered as a function not only of \( z \), the coordinate of a point on the \( z \) axis, but also of \( b \), (half) the separation of the two loops. Create a file conforming to the structure described in Section 9.3.2 and containing values of this function seen as a two-dimensional scalar function of \( z/a \) and \( b/a \). 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
Table 9.7: Procedure for Exercise 9.10.

PROCEDURE ??????
Obtain a positive integer A\% from a friend
Obtain a (second) positive integer B\% from a friend
CASE
   OF A\% has more digits than B\% DO
      LOOP
         Add digit 0 in front of B\% and call result B\%
      EXIT_LOOP WHEN B\% and A\% have same number of
digits
   END_LOOP
   OF B\% has more digits than A\% DO
      LOOP
         Add digit 0 in front of A\% and call result A\%
      EXIT_LOOP WHEN A\% and B\% have same number of
digits
   END_LOOP
END_CASE
WRITE A\% on a piece of paper
WRITE B\% under A\% with corresponding digits in same
column
Draw a line under B\%
DIGIT\% ← number of digits in either number
CARRY\% ← 0\%
LOOP
   SUM\% ← digit DIGIT\% of A\% + digit DIGIT\% of B\% +
   CARRY\%
   IF SUM\% < 10\% THEN CARRY\% ← 0\%
   ELSE BEGIN_BLOCK
      CARRY\% ← 1\%
      SUM\% ← SUM\% - 10\%
   END_BLOCK
END_IF
WRITE SUM\% under digit DIGIT\% of B\%
DIGIT\% ← DIGIT\% - 1\%
EXIT_LOOP WHEN DIGIT\% = 0\%
END_LOOP
IF CARRY\% = 1\%
   THEN WRITE CARRY\% in front of all digits beneath line
END_IF
END_PROCEDURE

language of your choice, testing your program(s) with the trajectory given by

\[ \mathbf{r} = \cos t \hat{i} + \sin t \hat{j} + 0.1t \hat{k} \]

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

\[ p_{3,1,0}(x, y, z) = \frac{8}{(27)^2 \pi} \rho^2 \left( 1 - \frac{\rho}{6} \right)^2 e^{-2\rho/3} \cos^2 \theta \]

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

\[ p_{3,2,1}(x, y, z) = \frac{3}{(27)^3 \pi} \rho^4 e^{-2\rho/3} \cos^2 \theta (1 - \cos^2 \theta) \]

\[ p_{3,2,2}(x, y, z) = \frac{3}{(27)^3 \pi} \rho^4 e^{-2\rho/3} (1 - \cos^2 \theta)^2 \]

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

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

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 \(E\) at the point \(r = x \hat{i} + y \hat{j}\) in the \(xy\) plane is given in mks units by

\[ E(x, y) = q \frac{1}{4 \pi \epsilon_0} \left[ \frac{x \hat{i} + (y - a) \hat{j}}{x^2 + (y - a)^2} - \frac{x \hat{i} + (y + a) \hat{j}}{x^2 + (y + a)^2} \right] \]

Expressing coordinates in terms of the dimensionless variables \(\tilde{x} = x/a\) and \(\tilde{y} = y/a\) and measuring \(E\) in the unit \(q/(4 \pi \epsilon_0 a^2)\), create a file conforming to the structure described in Section 9.3.2 and containing the \(x\) and \(y\) components of this field over the interval \(-2a \leq x, y \leq 2a.\) Divide the interval in each coordinate direction into about 25 segments, but choose the precise number carefully so as to avoid division by zero in evaluating the field at any point. Optional: Read your file into a suitable graphical display program and produce graphs showing the character of this field.

9.20. The velocity field of an object rotating about the \(z\) axis with angular momentum \(\omega\) is given in terms of the angular velocity and the position vector \(r\) by the expression

\[ \mathbf{v} = \omega \times \mathbf{r} = \omega \hat{k} \times (x \hat{i} + y \hat{j} + z \hat{k}) = \omega (-y \hat{i} + x \hat{j}) \]

Choosing a unit of length \(a\) and expressing coordinates in units of \(a\) and velocities in units of \(a \omega,\) 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 \(E\) at the point \(r = x \hat{i} + y \hat{j} + z \hat{k}\) is given in mks units by

\[ E(x, y, z) = q \frac{1}{4 \pi \epsilon_0} \left[ \frac{x \hat{i} + y \hat{j} + (z - a) \hat{k}}{x^2 + y^2 + (z - a)^2} - \frac{x \hat{i} + y \hat{j} + (z + a) \hat{k}}{x^2 + y^2 + (z + a)^2} \right] \]

Expressing coordinates in terms of the dimensionless variables \(\tilde{x} = x/a, \tilde{y} = y/a,\) and \(\tilde{z} = z/a\) and measuring \(E\) in the unit \(q/(4 \pi \epsilon_0 a^2)\), create a file conforming to the structure described in Section 9.3.2 and containing the \(x, y,\) and \(z\) components of this field over the interval \(-2a \leq
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° 5′ 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].
9.D Listing of lugen_list.pro (IDL)

pro lugen_list, arr

; Lugen_list and lugen_list_event work together to facilitate writing
; data stored in an array in IDL 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
; arr.

; Establish a COMMON block named tblock for storing the array and
; the ID's of several text-entry boxes so they can be communicated
; to the event handler lugen_list_event.

COMMON tblock, array, fname, ti, au, da, c1, c2, c3, c4

; Move array supplied in the input variable as arr to the COMMON
; block.

array=arr

; Set default filename

filename='untitled.dat'

; Create outermost widget named base (i.e., with its unique ID
; stored in base).

base=widget_base(title=" L U G E N _ L I S T ",
                 /row, xoffset=150, yoffset=250)

; Create child widgets lcol and rcol within base.

lcol=widget_base(base, /column, /frame)
rcol=widget_base(base, /column, /frame, space=50)

; Create widget with ID stored in b within lcol and, within b, create
; labeled text-entry boxes. Note that the text-entry boxes have
; ID's stored in fname and ti, whose values reside in the COMMON block.
; Note also that the uvalues 'FILENAME' and 'TITLE' are user-assigned
; here for later convenient identification of the referenced widgets.

b=widget_base(lcol, /row, /frame)
fnl=widget_label(b, value='Filename: ')
fname=widget_text(b, /editable, xsize=25, ysize=1, value=filename,$
                 uvalue='FILENAME')
til=widget_label(b, value='Title : ')
ti=widget_text(b, /editable, xsize=25, ysize=1, uvalue='TITLE')

; Repeat, creating labeled text entry windows for the author and
; the date. Since we have no further use for the ID previously written
; into b, we can here reuse the variable b.

b=widget_base(lcol, /row, /frame)
aul=widget_label(b, value='Author :')
au=widget_text(b, /editable, xsize=25, ysize=1, uvalue='AUTHOR')
dal=widget_label(b, value='Date :')
da=widget_text(b, /editable, xsize=25, ysize=1, uvalue='DATE')

; Repeat, creating labeled text entry windows for four lines of comments.

b=widget_base(lcol, /row, /frame)
c1l=widget_label(b, value='Comments 1:')
c1=widget_text(b, /editable, xsize=65, ysize=1, value='**', uvalue='COMM1')
b=widget_base(lcol, /row, /frame)
c2l=widget_label(b, value='Comments 2:')
c2=widget_text(b, /editable, xsize=65, ysize=1, value='**', uvalue='COMM2')
b=widget_base(lcol, /row, /frame)
c3l=widget_label(b, value='Comments 3:')
c3=widget_text(b, /editable, xsize=65, ysize=1, value='**', uvalue='COMM3')
b=widget_base(lcol, /row, /frame)
c4l=widget_label(b, value='Comments 4:')
c4=widget_text(b, /editable, xsize=65, ysize=1, value='**', uvalue='COMM4')

; Create SAVE and CANCEL buttons within rcol.

button1=widget_button(rcol, value='Save', uvalue='SAVE')
button1=widget_button(rcol, value='Close', uvalue='CLOSE')

; Display the created widget.

widget_control, base, /realize

; Invoke xmanager to sense the mouse events occurring within the base widget. With each event, xmanager calls lugen_list_event to provide the proper response.

xmanager, 'lugen_list', base

END
9.E Listing of lugen_list_event.pro (IDL)

pro lugen_list_event, event
+ ; This pro-file works in conjunction with the profile lugen_list.
; Everytime an event (mouse click) occurs in the base widget created by
; lugen_list, the IDL routine xmanager, which is invoked at the end of
; lugen_list, itself invokes lugen_list_event and supplies the input
; value for the variable event. Event is a structured variable
; containing information about the event that has occurred. That
; information is used here to determine what action should be taken
; in response to the event.
;-

; Create COMMON block parallel to the one in lugen_list.
COMMON tblock, array, fname, ti, au, da, c1, c2, c3, c4

; Fetch the uvalue assigned in lugen_list to identify the particular
; widget in which the event occurred, storing that uvalue in the
; variable eventval, and then execute the proper response determined
; by the uvalue associated with the event that occurred.

widget_control, event.id, get_uvalue=eventval

case eventval of

; Respond to mouse clicks in several text-entry boxes
; by doing nothing

'FILENAME':
'TITLE':
'AUTHOR':
'DATE':
'COMM1':
'COMM2':
'COMM3':
'COMM4':

; Respond to mouse click on SAVE button by reading data in
; base widget, processing it, writing it to the specified
; output file, and terminating execution of lugen_list. Note that
; the widget ID’s written into the COMMON block by lugen_list are
; here read out of that COMMON block to identify the widget
; to be read by each widget_control statement.

'SAVE' : begin

; Read data in text-entry windows.

widget_control, fname, get_value=filename
widget_control, ti, get_value=title
widget_control, au, get_value=author
widget_control, da, get_value=date
widget_control, c1, get_value=comment1
widget_control, c2, get_value=comment2
widget_control, c3, get_value=comment3
widget_control, c4, get_value=comment4

; Determine dimensions of array

dim = size(array)
narr = 1
xdim = fix( dim[1] )
ydim = fix( dim[2] )
zdim = fix( dim[3] )
case dim[0] of
  2: zdim = 1
  3:
endcase

; Write output file, inform user of success, ; return to IDL main level

openw, 1, filename(0)
printf, 1, title + ' Author: ' + author + ' Date: ' + date
printf, 1, comment1
printf, 1, comment2
printf, 1, comment3
printf, 1, comment4
printf, 1, narr
printf, 1, xdim
printf, 1, ydim
printf, 1, zdim
for n = 0, narr-1 do$
  for k = 0, zdim-1 do$
    for j = 0, ydim-1 do$
      for i = 0, xdim-1 do$
        printf, 1, array[i,j,k,n]
close, 1
print, "File ", filename(0), ", saved successfully!"

widget_control, event.top,/destroy
end

; Respond to mouse click on CLOSE button.

'CLOSE' : widget_control, event.top, /destroy
endcase

END
9.Q Listing of luread_list.pro (IDL)

```idl
pro luread_list, filename, U

; This pro-file reads a file created in the format described in
; CPSUP and stores the array internal to IDL. It is invoked
; with the command
;
; luread_list, 'filename', U
;
; where 'filename' specifies the file and U identifies the name of the
; variable into which the data will be read.
;
;++++ Block 1: Print starting message
; Specify string variables for reading first lines of file ++++

print, format = '($,"Reading file")'
head1 = ' ' & head2 = ' ' & head3 = ' ' & head4 = ' ' & head5 = ' '

;++++ Block 2: Open file for reading, read headers and dimensions
; into IDL, and write headers to terminal ++++

openr, 1, filename
readf, 1, head1 & print, format = '($,".")'
readf, 1, head2 & print, format = '($,".")'
readf, 1, head3 & print, format = '($,".")'
readf, 1, head4 & print, format = '($,".")'
readf, 1, head5 & print, format = '($,".")'
readf, 1, narr & narr = fix(narr) & print, format = '($,".")'
readf, 1, xdim & xdim = fix(xdim) & print, format = '($,".")'
readf, 1, ydim & ydim = fix(ydim) & print, format = '($,".")'
readf, 1, zdim & zdim = fix(zdim) & print, format = '($,".")'
print, ".

print, " ", head1 & print, " ", head2 & print, " ", head3
print, " ", head4 & print, " ", head5

;++++ Block 3: Read U into IDL and close file ++++

U = fltarr( xdim, ydim, zdim, narr )
for n = 0, narr-1 do $
  for k = 0, zdim-1 do $
    for j = 0, ydim-1 do $
      for i = 0, xdim-1 do $
        begin readf,1, tmp & U[i,j,k,n] = tmp & endfor

close, 1

;++++ Block 4: Print reassuring message ++++

print, "..........File ", filename, "; read successfully!"

end```

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\(^1\) \(-mg\mathbf{k}\) and the viscous force \(\mathbf{F}_v\), from the medium in which it moves. The former is directed downward and the latter is directed

\(^1\)We choose a coordinate system in which positive \(z\) is directed upward, and we take \(m\) and \(g\) to be positive.
11.1. SAMPLE PROBLEMS

opposite to the velocity \( \mathbf{v} \). Usually, the magnitude of the viscous force is a function of the speed with which the projectile moves, symbolically \( |\mathbf{F}_v| = f(|\mathbf{v}|) \). In general, if the projectile has mass \( m \), Newton’s second law yields the equation of motion

\[
m \frac{d^2 \mathbf{r}}{dt^2} = -mg \hat{k} + \mathbf{F}_v(|\mathbf{v}|) = -mg \hat{k} - f(|\mathbf{v}|) \frac{\mathbf{v}}{|\mathbf{v}|}
\]

(11.1)

which we are to solve subject to the general initial conditions,

\[
\mathbf{r}(0) = \mathbf{r}_0 = x_0 \hat{i} + y_0 \hat{j} + z_0 \hat{k} ; \quad \mathbf{v}(0) = \mathbf{v}_0 = v_{x0} \hat{i} + v_{y0} \hat{j} + v_{z0} \hat{k}
\]

(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

\[
m \frac{d^2 \mathbf{r}}{dt^2} = -mg \hat{k} - b \mathbf{v} = -mg \hat{k} - b \frac{d \mathbf{r}}{dt}
\]

(11.3)

or, in component form, to

\[
m \frac{d^2 x}{dt^2} = -b \frac{dx}{dt} ; \quad m \frac{d^2 y}{dt^2} = -b \frac{dy}{dt} ; \quad m \frac{d^2 z}{dt^2} = -mg - b \frac{dz}{dt}
\]

(11.4)

This system of equations is uncoupled, since each of the three independent variables \( x, y, \) and \( z \) satisfies its own private equation that does not involve either of the other variables. They are second-order and linear, and the coefficients are constant. The first two are homogeneous and, because of the term \(-mg\), the third is inhomogeneous. The equations involve the parameters \( m, b, \) and \( g \), and their solutions will depend on these parameters and on the six initial values in Eq. (11.2).

When the speed is too large for a linear approximation to the viscous damping, we can sometimes take the viscous force to be proportional instead to the square of the speed, \( f = b |\mathbf{v}|^2 \) \( (b \) a positive constant, though not the same constant as in the previous paragraph). This time, the equation of motion reduces to

\[
m \frac{d^2 \mathbf{r}}{dt^2} = -mg \hat{k} - b|\mathbf{v}| \mathbf{v} = -mg \hat{k} - b \frac{d \mathbf{r}}{dt} \frac{\mathbf{v}}{|\mathbf{v}|} \sqrt{\left( \frac{dx}{dt} \right)^2 + \left( \frac{dy}{dt} \right)^2 + \left( \frac{dz}{dt} \right)^2}
\]

(11.5)

or, in component form, to

\[
m \begin{bmatrix} \frac{d^2 x}{dt^2} & \frac{d^2 y}{dt^2} & \frac{d^2 z}{dt^2} \end{bmatrix} = [0, 0, -mg] - b \begin{bmatrix} \frac{dx}{dt} & \frac{dy}{dt} & \frac{dz}{dt} \end{bmatrix} \sqrt{\left( \frac{dx}{dt} \right)^2 + \left( \frac{dy}{dt} \right)^2 + \left( \frac{dz}{dt} \right)^2}
\]

(11.6)

This system of equations is second-order and distinctly non-linear, and its members are coupled because each of the equations involves all three of the dependent variables. Even if the motion occurs in only the vertical dimension (projectile tossed straight up or, simply, released from rest and allowed to drop), the equation, which then is

\[
m \frac{d^2 z}{dt^2} = -mg - b \frac{dz}{dt} \frac{dz}{dt}
\]

(11.7)

is complicated by the absolute value in the viscous term. These equations involve the parameters \( m, b, \) and \( g \), and their solutions will depend on these parameters and on the six initial values in Eq. (11.2).

To cast these equations in a dimensionless form, we would start by choosing a unit of length, say \( a \)—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 \( \mathbf{r} = x/a, \mathbf{g} = y/a, \) and \( \mathbf{z} = z/a \) (i.e., \( \mathbf{r} = r/a \)), we find that Eqs. (11.3) and (11.5) become

\[
\frac{d^2 \mathbf{r}}{dt^2} = -\frac{g}{a} \mathbf{k} - \frac{b}{m} \frac{d\mathbf{r}}{dt}
\]

(11.8)

and

\[
\frac{d^2 \mathbf{r}}{dt^2} = -\frac{b}{m} \frac{d\mathbf{r}}{dt} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2} \left(\frac{ba}{m}\right)
\]

(11.9)

respectively. Next, we recognize that \( \sqrt{g/a} \) is dimensionally a frequency. Hence, the variable \( \tilde{t} = t/\sqrt{g/a} \) provides a suitable rescaling of the independent variable. In terms of \( \tilde{t} \), Eqs. (11.8) and (11.9) become

\[
\frac{d^2 \mathbf{r}}{d\tilde{t}^2} = -\frac{b}{m} \frac{d\mathbf{r}}{d\tilde{t}} \sqrt{\left(\frac{dx}{d\tilde{t}}\right)^2 + \left(\frac{dy}{d\tilde{t}}\right)^2 + \left(\frac{dz}{d\tilde{t}}\right)^2} \left(\frac{ba}{m}\right)
\]

(11.10)

and

\[
\frac{d^2 \mathbf{r}}{d\tilde{t}^2} = -\frac{b}{m} \frac{d\mathbf{r}}{d\tilde{t}} \left(\frac{ba}{m}\right)
\]

(11.11)

Finally, we introduce the symbol \( \beta \) for the dimensionless quantity \( b\sqrt{a/g}/m \) in the first of these equations and for the dimensionless quantity \( ba/m \) in the second. In the end, the equations in dimensionless form are

\[
\frac{d^2 \mathbf{r}}{d\tilde{t}^2} = -\frac{b}{m} \beta \frac{d\mathbf{r}}{d\tilde{t}}
\]

(11.12)

and

\[
\frac{d^2 \mathbf{r}}{d\tilde{t}^2} = -\frac{b}{m} \sqrt{\left(\frac{dx}{d\tilde{t}}\right)^2 + \left(\frac{dy}{d\tilde{t}}\right)^2 + \left(\frac{dz}{d\tilde{t}}\right)^2} \left(\frac{ba}{m}\right)
\]

(11.13)

These equations are, of course, to be solved subject to the dimensionless initial conditions

\[
\mathbf{r}(0) = \frac{r_0}{a}; \quad \mathbf{v}(0) = \frac{d\mathbf{r}}{d\tilde{t}}(0) = \frac{d(r/a)}{d(t/\sqrt{g/a})}(0) = \frac{v_0}{\sqrt{g/a}}
\]

(11.14)

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

\[
m \frac{dv_x}{dt} = -bv_x; \quad m \frac{dv_y}{dt} = -bv_y; \quad m \frac{dv_z}{dt} = -mg - bv_z
\]

(11.15)

the three members of Eq. (11.6) become

\[
m \begin{bmatrix} \frac{dv_x}{dt}, \frac{dv_y}{dt}, \frac{dv_z}{dt} \end{bmatrix} = [0, 0, -mg] - b[v_x, v_y, v_z] \sqrt{v_x^2 + v_y^2 + v_z^2}
\]

(11.16)

and Eq. (11.7) becomes

\[
m \frac{dv_z}{dt} = -mg - bv_z |v_z|
\]

(11.17)

Once these first-order equations have been solved, the components of the position vector can then be found by solving the differential equations

\[
\frac{dx}{dt} = v_x; \quad \frac{dy}{dt} = v_y; \quad \frac{dz}{dt} = v_z
\]

(11.18)

---

2Using \([\_\_]\) to indicate “the dimensions of \(\ldots\),” we argue \([g] = \text{length}/\text{time}^2, [a] = \text{length} \Rightarrow [g/a] = \text{time}^{-2} \Rightarrow [\sqrt{g/a}] = \text{time}^{-1}\).  
3Remember that \(b\) is not the same in the two instances.  
4Here (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.

\[ A \xrightarrow{k_A} B \xrightarrow{k_B} C \]

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

\[
\frac{dA}{dt} = -k_A A \quad ; \quad \frac{dB}{dt} = k_A A - k_B B \quad ; \quad \frac{dC}{dt} = k_B B
\]

where \( k_A \) and \( k_B \) are decay constants (parameters); \( A(t) \), \( B(t) \), and \( C(t) \) are the number of nuclei of each species present; and nucleus \( C \) is assumed to be stable. These equations are linear, first-order, and homogeneous, and they have constant coefficients. They are, however, coupled, since each of the second and third of them involves two of the dependent variables. They also support a conservation law: adding the three equations yields that

\[
\frac{d}{dt}(A + B + C) = 0 \quad \implies \quad A + B + C = \text{constant} \quad (11.20)
\]

As always we, of course, need initial values, e.g., \( A(0) = A_0, B(0) = 0, \) and \( C(0) = 0, \) before the differential equations have a unique solution, and that solution will depend on the parameters \( k_A \) and \( k_B \) and on the three initial values.

To cast these equations in a dimensionless form, we choose a reference amount—here, conveniently, \( A_0 \), the initial amount of \( A \)—as the unit for measuring the quantities of \( A, B, \) and \( C \). Then, we rescale the values for these quantities by introducing the dimensionless variables \( \bar{A} = A/A_0, \bar{B} = B/A_0, \) 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

\[
\frac{d\bar{A}}{d\bar{t}} = -\bar{A} \quad ; \quad \frac{d\bar{B}}{d\bar{t}} = \bar{A} - \bar{k} \bar{B} \quad ; \quad \frac{d\bar{C}}{d\bar{t}} = \bar{k} \bar{B} \quad (11.21)
\]

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.

---

5We shall later see that conservation laws can sometimes prove valuable in assessing the accuracy of solutions.
11.1.3 Exponential and Logistic Growth

An important illustration of a non-linear first-order equation occurs in population biology, where—in the absence of predation—the population of a species grows at a rate proportional to that population until the population becomes so large that individual organisms compete significantly with one another for space and/or food. In a simple model, the effects of competition are proportional to the likelihood that one organism will encounter another—a likelihood that is proportional to the square of the population. Thus, a population subject to both effects will evolve in accordance with the first-order, non-linear equation

\[
\frac{dN}{dt} = kN \left(1 - \frac{N}{N_c}\right)
\]

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 \(\tilde{N} = N/N_c\) and the dimensionless time \(\tilde{t} = kt\), and find that the equation and initial condition reduce to

\[
\frac{d\tilde{N}}{d\tilde{t}} = \tilde{N}(1 - \tilde{N}) ; \quad \tilde{N}(0) = \frac{N_0}{N_c}
\]

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.

\[\text{Logistic growth is also sometimes said to follow a sigmoid curve because of the stylized ‘S’ shape that solutions exhibit when the initial population is much smaller than the carrying capacity. Graphs of this shape will be found in subsequent sections of this chapter.}\]
To cast this equation in dimensionless form, we introduce a unit of length, say $\ell$, rescale position with the expression $\bar{x} = x/\ell$, introduce a unit of time, say $\tau$, rescale the physical time with the expressions $\bar{t} = t/\tau$, and thereby transform Eq. (11.24) to

$$\frac{d^2\bar{x}}{d\bar{t}^2} = -\frac{\tau^2 k}{m} \bar{x} - \frac{\tau b}{m} \frac{d\bar{x}}{d\bar{t}} + \frac{\tau^2 F(\tau t)}{m\ell}$$

(11.26)

The term on the left in this equation is now dimensionless. Thus, all terms on the right must be dimensionless as well. In particular, the combination $\tau^2 k/m$ must be dimensionless. Remember, however, that $\tau$ is at the moment merely a symbol; we have not yet made a well defined choice for the unit of time, and we are free to choose $\tau$ any way we like. Clearly, the choice $\tau^2 k/m = 1$ or $\tau = \sqrt{m/k}$ is judicious. With this choice, the equation of motion becomes

$$\frac{d^2\bar{x}}{d\bar{t}^2} = -\bar{x} - \bar{\beta} \frac{d\bar{x}}{d\bar{t}} + \bar{F}(\bar{t})$$

(11.27)

where, in the final form, we have set $\bar{\beta} = b/\sqrt{mk}$ and $\bar{F}(\bar{t}) = F(\tau t)/k\ell$. The single, dimensionless parameter $\bar{\beta}$ contains the essential influence of the three parameters $m$, $b$, and $k$ once differences attributable to scaling have been removed. The dimensionless force $\bar{F}$ expresses the physical force in units of $k\ell$, which—note—is the force that the spring would exert if extended by the chosen unit of length!

We must, of course, also translate the initial conditions of Eq. (11.25) into dimensionless form, finding that

$$\bar{x}(0) = \frac{x_0}{\ell} \quad ; \quad \bar{v}(0) = \frac{d\bar{x}}{d\bar{t}}(0) = \frac{d(x/\ell)}{d(t/\tau)} = \frac{v_0}{\ell/\tau}$$

(11.28)

and we conclude that dimensionless velocities will be measured in units of $\ell/\tau$, which is the speed of an object that moves the reference distance $\ell$ in the reference time $\tau$.

Alternatively (and, in some approaches to solution, necessarily), we would recast this single second-order differential equation as a pair of first-order equations, either

$$\frac{dx}{dt} = v \quad ; \quad m \frac{dv}{dt} = -kx - bv + F(t)$$

(11.29)

in the original dimensional form, or

$$\frac{d\bar{x}}{d\bar{t}} = \bar{v} \quad ; \quad \frac{d\bar{v}}{d\bar{t}} = -\bar{x} - \bar{\beta} \bar{v} + \bar{F}(\bar{t})$$

(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

$$ V(t) - L \frac{di}{dt} - \frac{q}{C} - iR = 0 \implies L \frac{di}{dt} + iR + \frac{q}{C} = V(t) \tag{11.31} $$

As it stands, this equation looks to be first order but it involves two variables $i(t)$ and $q(t)$. We can complete the statement of a problem having a unique solution by recognizing the relationship

$$ i = \frac{dq}{dt} \tag{11.32} $$

between $i$ and $q$ and supplementing what is now a pair of coupled first-order, linear, constant coefficient, inhomogeneous equations with the general initial conditions

$$ q(0) = q_0 ; \quad \frac{dq}{dt}(0) = i(0) = i_0 \tag{11.33} $$

Alternatively, we could substitute Eq. (11.32) into Eq. (11.31) to find the equivalent, single, second-order, linear, constant coefficient, inhomogeneous equation

$$ L \frac{d^2q}{dt^2} + R \frac{dq}{dt} + \frac{q}{C} = V(t) \tag{11.34} $$

The recasting of these equations in dimensionless form is left as an exercise.

Adopting the correspondences, $q \leftrightarrow x$, $L \leftrightarrow m$, $R \leftrightarrow b$, $1/C \leftrightarrow k$, $V(t) \leftrightarrow F(t)$, and $t \leftrightarrow t$, we can turn Eq. (11.34) into Eq. (11.24). Thus, mathematically, the driven, damped mechanical oscillator and the RLC circuit exhibit exactly analogous behavior, and the behavior of the RLC circuit simulates the behavior of the mechanical oscillator.

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.

Years ago, when digital computers were not as fast as they have come to be and smooth graphical output from a digital computer was unheard of, correspondences such as this one were the basis of the analog computer, whereon we could easily set up electronic circuits whose behavior simulated the behavior of more expensive mechanical systems. With an analog computer, we could learn about mechanical systems by observing the real-time variation of the voltages and currents at various points in an analogous electronic circuit.
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

$$\begin{align*}
    m \frac{d^2x_1}{dt^2} &= -kx_1 + k'(x_2 - x_1) \\
    m \frac{d^2x_2}{dt^2} &= -kx_2 - k'(x_2 - x_1)
\end{align*}$$

(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 $\tilde{t} = \omega t$, where $\omega = \sqrt{k/m}$, and setting $\kappa = k'/k$. The equations then become

$$\begin{align*}
    \frac{d^2\bar{x}_1}{d\tilde{t}^2} &= -\bar{x}_1 + \kappa (\bar{x}_2 - \bar{x}_1) \\
    \frac{d^2\bar{x}_2}{d\tilde{t}^2} &= -\bar{x}_2 - \kappa (\bar{x}_2 - \bar{x}_1)
\end{align*}$$

(11.36)

Initial conditions such as

$$\begin{align*}
    \bar{x}_1(0) &= \bar{x}_{10} \\
    \frac{d\bar{x}_1}{dt}(0) &= \bar{v}_{10} \\
    \bar{x}_2(0) &= \bar{x}_{20} \\
    \frac{d\bar{x}_2}{dt}(0) &= \bar{v}_{20}
\end{align*}$$

(11.37)

complete the statement of the problem—a problem that involves a pair of coupled, linear, second-order, constant coefficient, homogeneous differential equations containing one internal parameter $\kappa$.

### 11.1.7 Motion under Central Forces

Consider next an object of mass $m$ moving in the $xy$ plane under the action of a central force$^8 \mathbf{F}$, as shown in Fig. 11.6. According to Newton’s second law, the position vector $\mathbf{r}$ of this object satisfies

$^8$A central force is one whose direction is always away from or towards a fixed point—the force center—and whose magnitude depends only on the distance from that point.
the differential equation
\[ m \frac{d^2 \mathbf{r}}{dt^2} = \mathbf{F} = f(r) \hat{r} \]  
(11.38)
where \( f(r) > 0 \) corresponds to a repulsive force and \( f(r) < 0 \) corresponds to an attractive force.

Two articulations of this equation are in order. A more specific expression in polar coordinates is treated in an exercise. Here, we extract its components in Cartesian coordinates \((x, y)\). Since
\[ r^2 = x^2 + y^2, \quad \mathbf{r} = x \hat{i} + y \hat{j}, \quad \text{and} \quad \hat{r} = \frac{r}{r} = \frac{x \hat{i} + y \hat{j}}{\sqrt{x^2 + y^2}} \]  
(11.39)
the Cartesian components of Eq. (11.38) are
\[ m \frac{d^2 x}{dt^2} = f \left( \sqrt{x^2 + y^2} \right) \frac{x}{\sqrt{x^2 + y^2}} ; \quad m \frac{d^2 y}{dt^2} = f \left( \sqrt{x^2 + y^2} \right) \frac{y}{\sqrt{x^2 + y^2}} \]  
(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, 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
\[ \frac{d^2 x}{dt^2} = -GM \left( \frac{x}{(x^2 + y^2)^{3/2}} ; \quad \frac{d^2 y}{dt^2} = -GM \left( \frac{y}{(x^2 + y^2)^{3/2}} \right) \]  
(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
\[ \mathbf{r} = \frac{x}{\ell} ; \quad \mathbf{y} = \frac{y}{\ell} ; \quad \mathbf{r} = \frac{r}{\ell} \]  
(11.44)
With this change, Eq. (11.43) becomes
\[ \frac{d^2 \mathbf{r}}{dt^2} = -GM \left( \frac{\mathbf{r}}{(\mathbf{r}^2 + \mathbf{y}^2)^{3/2}} ; \quad \frac{d^2 \mathbf{y}}{dt^2} = -GM \left( \frac{\mathbf{y}}{(\mathbf{r}^2 + \mathbf{y}^2)^{3/2}} \right) \]  
(11.45)
Finally, we introduce the dimensionless time variable \( \tilde{t} = t \sqrt{GM/\ell^3} \) to find that
\[ \frac{d^2 \mathbf{r}}{d\tilde{t}^2} = -\frac{\mathbf{r}}{(\mathbf{r}^2 + \mathbf{y}^2)^{3/2}} ; \quad \frac{d^2 \mathbf{y}}{d\tilde{t}^2} = -\frac{\mathbf{y}}{(\mathbf{r}^2 + \mathbf{y}^2)^{3/2}} \]  
(11.46)
Interestingly, in Cartesian coordinates, all parameters have disappeared from the equations.

\[ \text{9} \text{Remember, too, that the problem of a charged particle moving in the field of another charged particle is mathematically identical to the planetary problem. (See exercises.)} \]
Translation of the initial conditions into dimensionless form is easier than translation of the differential equations. The positions, of course, become

\[ \bar{x}(0) = \frac{x_0}{\ell} \quad ; \quad \bar{y}(0) = \frac{y_0}{\ell} \]  

(11.47)

To translate the velocities, we argue that

\[ \frac{d\bar{x}}{dt}(0) = \frac{d(x/\ell)}{d(t\sqrt{GM/\ell^3})}(0) = \frac{v_x}{\ell} \]  

(11.48)

\[ \frac{d\bar{y}}{dt}(0) = \frac{d(y/\ell)}{d(t\sqrt{GM/\ell^3})}(0) = \frac{v_y}{\ell} \]

We conclude that dimensionless velocities in the present context are measured in units of \( \sqrt{GM/\ell} \).

Before leaving this important problem, we note two additional features. First, if the planet happens to be moving at a distance \( r_{\text{circ}} \) from the sun with a velocity of magnitude \( v_{\text{circ}} \) directed perpendicular to the radius line, then the planet will move in a circular orbit if its speed and radius are related so that the centripetal force \( m v_{\text{circ}}^2 / r_{\text{circ}} \) 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

\[ \frac{m v_{\text{circ}}^2}{r_{\text{circ}}} = \frac{GMm}{r_{\text{circ}}^2} = \Rightarrow v_{\text{circ}}^2 = \frac{GM}{r_{\text{circ}}} \]  

(11.49)

Translated into dimensionless form, this special relationship becomes

\[ \left( \sqrt{\frac{GM}{\ell}} v_{\text{circ}} \right)^2 = \frac{GM}{\ell r_{\text{circ}}} = \Rightarrow v_{\text{circ}}^2 = \frac{1}{r_{\text{circ}}} \]  

(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

\[ E = \frac{1}{2} m \left( v_x^2 + v_y^2 \right) - \frac{GMm}{\sqrt{x^2 + y^2}} = \text{constant} \]  

(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

\[ L = m \left( xv_y - yv_x \right) = \text{constant} \]  

(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

\[ \frac{1}{2} \left( \bar{v}_x^2 + \bar{v}_y^2 \right) - \frac{1}{\sqrt{\bar{x}^2 + \bar{y}^2}} = \text{constant} ; \quad \bar{x} \bar{v}_y - \bar{y} \bar{v}_x = \text{constant} \]  

(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,

\[
  \frac{d^2 \psi}{dx^2} + (2\epsilon - x^2) \psi = 0
\]

where \( \omega = \sqrt{k/m} \), \( \epsilon = E/\hbar\omega \) and \( x = x\sqrt{\hbar/m\omega} \). 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 $x$. 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. 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

$$\frac{d^2\psi}{dx^2} + (2\epsilon - x^2) \psi = 0 \quad ; \quad \psi(0) = 1.0 \quad ; \quad \lim_{x \to \infty} \psi(x) = 0$$

and yielding even solutions and the other defined by the equation and initial values

$$\frac{d^2\psi}{dx^2} + (2\epsilon - x^2) \psi = 0 \quad ; \quad \psi(0) = 0 \quad ; \quad \frac{d\psi}{dx}(0) = 1 \quad ; \quad \lim_{x \to \infty} \psi(x) = 0$$

and yielding odd solutions. The interesting aspect of these problems is that we will be able to find solutions satisfying the boundary requirement at $x = +\infty$ 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

$$L(f(t)) = \tilde{f}(s) = \int_0^\infty e^{-st} f(t) dt$$

We can best illustrate the value of the Laplace transform in solving a linear, ordinary differential equation with constant coefficients by presenting a short example. We would—and symbolic manipulators may well—apply this technique to the third member of Eq. (11.4) as follows. First we use Eqs. (1.16) and (1.17) and entries in Table 1.2 to evaluate its Laplace transform, finding that

$$m[s^2 \tilde{z}(s) - sz_0 - v_{z0}] = -\frac{mg}{s} - b[s \tilde{z}(s) - z_0]$$

Then we solve this algebraic equation for $\tilde{z}(s)$, finding that

$$\tilde{z}(s) = \frac{z_0}{s} + \frac{v_{z0}s - g}{s^2(s + b/m)}$$

Finally, we invert this Laplace transform to find the solution itself. Unfortunately, inverting Laplace transforms is rarely easy. We can, of course, read a table of transforms such as Table 1.2 backwards,

---

10It involves only second derivatives and the potential energy involves only $x^2$.

11Reducing 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.

12The 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.

13Murphy’s law, in the version that says that transforms may move difficulties around but may not eliminate them altogether, applies. The price we pay for the conversion to an algebraic equation is that we now have to invert the transform we found so easily.
but that strategy will work only if the entry we need is in fact included in the table. In the present case, however, we can recast the result we have obtained by using the technique of partial fractions to find the equivalent expression

\[ \tilde{z}(s) = \frac{z_0}{s} - \frac{mg/b}{s^2} + \frac{m}{b} \left( v_{z0} + \frac{mg}{b} \right) \left( \frac{1}{s} - \frac{1}{s + b/m} \right) \]  

(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.

### 11.5 Solving ODEs Symbolically with \textit{Mathematica}

Fundamentally, solving an ordinary differential equation involves specifying the equation, finding a solution containing the appropriate number of arbitrary constants, and imposing the appropriate initial or boundary conditions. A few features of the command \texttt{DSolve}, which is the primary ODE-solving command in \textit{Mathematica}, were discussed briefly in Section 8.8.9. In that section, we used \texttt{DSolve} only for solving a single ODE, and we learned that—to that end—the command has two forms, each of which itself has two (sub)forms. To use the first form, we submit either the statement

\texttt{DSolve[ ODE, DVar[IVar], IVar ]}

or the statement

\texttt{DSolve[ ODE, DVar, IVar ]}

In either case, the command returns the solution of the single ordinary differential equation \textit{ODE} for the dependent variable \textit{DVar} as a function of the single independent variable \textit{IVar}. In the first subform, the solution is presented as a transformation rule for \( x[t] \); in the second subform, \textit{Mathematica} instead returns a rule defining \( x \) as a function of the variable \( t \). In the form of this paragraph, \texttt{DSolve} does not impose initial conditions; they haven’t even been specified. Instead, it uses the symbols \( C[1], C[2], \ldots \) (however many are needed) for the integration constants. These constants must then be determined by subsequent imposition of appropriate initial conditions.

In the second form, the command \texttt{DSolve} accepts as its first argument a \textit{list} which contains both the differential equation to be solved and an appropriate number of initial conditions. The statement would then have either the first subform

\texttt{DSolve[ \{ ODE, IC1, IC2, \ldots \}, DVar[IVar], IVar ]}

or the second subform

\texttt{DSolve[ \{ ODE, IC1, IC2, \ldots \}, DVar, IVar ]}

where these statements return results with the same features as described in the previous paragraph for the first form. These forms return solutions on which the initial conditions \textit{IC1}, \textit{IC2}, \ldots have been imposed and which therefore contain no undetermined constants. If, for example, the function to be found is \( y[t] \), then an initial condition on the function itself would have the form \( y[0] == y_0 \) while an initial condition on its first derivative would have the form \( y'[0] == v_0 \), where \( y_0 \) and \( v_0 \)
are the desired values—and may be symbolic or explicitly numeric. Note, in particular, the use of the symbol \(==\) to specify those initial conditions.

The command \texttt{DSolve} is also capable of solving (some) systems of ordinary differential equations. In this case, the first argument will be a list containing all of the differential equations as well as any initial conditions to be imposed, the second argument will be a list containing the several dependent variables, and the third argument will identify the independent variable. In brief, the statement invoking \texttt{DSolve} will have the subform

\[
\text{DSolve}[ \{ \text{ODE1, ODE2, \ldots, IC1, IC2, \ldots} \}, \\
\{ \text{DVar1[IVar], DVar2[IVar], \ldots} \}, \text{IVar} ]
\]

or, alternatively, the subform

\[
\text{DSolve}[ \{ \text{ODE1, ODE2, \ldots, IC1, IC2, \ldots} \}, \\
\{ \text{DVar1, DVar2, \ldots} \}, \text{IVar} ]
\]

### 11.5.1 Projectile in a Viscous Medium

For the sake of a comparison with the result obtained in Section 11.2, we begin by using \textit{Mathematica} to solve the third member of Eq. (11.4). The necessary commands are\(^\text{14}\)

\[
\text{In}[1]:= \text{eq} = m z''[t] == -m g - b z'[t] \\
\text{Out}[1]= m z''[t] == -mg - bz'[t] \\
\text{In}[2]:= \text{soln} = \text{DSolve}[ \text{eq, z, t } ][[1,1]]
\]

Define differential equation. Solve the equation for \(z\) as a function of \(t\). \texttt{DSolve} returns a solution buried two-levels down in lists. The suffix \([[1,1]]\) elevates the solution to the top level.

\[
\text{Out}[2]= z \rightarrow \text{Function}\left[\{t\}, \frac{-gmt}{b} - \frac{e^{-bt/m}}{b} m c[1] + c[2]\right]
\]

Note that \textit{Mathematica} has here tacitly assumed that \(b \neq 0\). More careful treatment would be required if \(b = 0\), i.e., in the case of no damping.

To complete the solution, we must of course impose appropriate initial conditions. Thus, we continue with the statements

\[
\text{In}[3]:= \text{eq1} = z[0] == z_0 /. \text{soln}; \\
\text{In}[4]:= \text{eq2} = z'[0] == v_0 /. \text{soln}; \\
\text{In}[5]:= \text{Solve}[ \{ \text{eq1,eq2} \}, \{c[1],c[2]\} ][[1]]$
\]

Solve for integration constants, elevating the result one-level to be a simple list rather than a one-element list of a one-element list. \texttt{Solve} returns a solution buried three-levels down in lists. The suffix \([[1]]\) elevates the solution to the top level.

\[
\text{In}[6]:= \text{soln} = \text{soln} /. %; \\
\text{In}[7]:= z[t] /. \text{soln}; \\
\text{In}[8]:= z'[t] /. \text{soln};
\]

Substitute constants into solution. Evaluate position. Evaluate velocity.

The end result of these operations is the solution

\[
z(t) = z_0 - \frac{mg}{b} t + \frac{m}{b} \left( v_0 + \frac{mg}{b} \right) \left( 1 - e^{-bt/m} \right) \tag{11.69}
\]

\(^\text{14}\)As in Chapter 8, we abbreviate the presentation of \textit{Mathematica} dialogs by making liberal use of terminating semicolons to suppress intermediate output. You are urged to duplicate the dialog in an actual session with \textit{Mathematica}, omitting all terminating semicolons.
for the position and the solution

\[
v_z(t) = -\frac{mg}{b} + \left(v_{z0} + \frac{mg}{b}\right) e^{-bt/m}
\]  \hspace{1cm} (11.70)

for the velocity, though we have rearranged the solutions presented by Mathematica to improve their readability. This result for \(z(t)\) agrees fully with the result obtained in Eq. (11.68).

Wisdom suggests that we should whenever possible actually verify any solution that Mathematica generates. To this end, we substitute the solution into the original equation, evaluate the derivatives explicitly, and simplify the result with the statement

\[
\text{In[9]} := \text{Simplify}[\text{eq} /. \text{soln}]
\]

\[
\text{Out[9]} = \text{True}
\]

To complete the verification, we check satisfaction of the initial conditions with the statement

\[
\text{In[10]} := \text{Simplify}[\{z[0], z'[0]\} /. \text{soln}]
\]

\[
\text{Out[10]} = \{z_0, v_{z0}\}
\]

All is indeed well.

One further check on this solution involves examining its limit as \(b\)—the viscous damping—becomes small. For the coordinate and velocity, we find respectively that

\[
\text{In[11]} := \text{Collect}[\text{Series}[z[t] /. \text{soln}, \{b, 0, 1\}], t]
\]

\[
\text{Out[11]} = z_0 + v_{z0}t - \frac{1}{2}gt^2 - \frac{bt^2}{2m}\left(v_{z0} - \frac{1}{3}gt\right) + O[b]^2
\]

\[
\text{In[12]} := \text{Collect}[\text{Series}[z'[t] /. \text{soln}, \{b, 0, 1\}], t]
\]

\[
\text{Out[12]} = v_{z0} - gt - \frac{bt}{m}\left(v_{z0} - \frac{1}{2}gt\right) + O[b]^2
\]

The terms free of \(b\) (i.e., terms in \(b^0 = 1\)) clearly agree with the known results for free fall in the absence of viscous resistance.

The solution of this problem when motion occurs in two or three of the coordinate directions is explored in one of the exercises.

We have, of course, taken a bit longer route than necessary so as to illustrate the use of Mathematica to find the undetermined constants in the general solution to an ODE. We could combine the first several of our statements into an invocation of \texttt{DSolve} that includes the incorporation of the initial conditions with the statements

\[
\text{In[13]} := \text{Quit[]}
\]

\[
\text{In[1]} := \text{eq} = \text{m}\cdot z''[t] == -m\cdot g - b\cdot z'[t];
\]

\[
\text{In[2]} := \text{soln} = \text{DSolve}\{\text{eq}, z[0]==z_0, z'[0]==v_{z0}\}, z, t\};
\]

\[
\text{In[3]} := \text{Quit[]}
\]

These statements lead us in rather fewer steps to the result already presented in Eq. (11.69), though it may require extraction of the expression for the solution from \text{soln} and a subsequent application of \texttt{Expand} and \texttt{Collect} to reveal that agreement more clearly.

\[
\text{---}\hspace{1cm}^{15}\hspace{1cm}\text{---}
\]

\^[15]Again, we have taken the liberty to rearrange the terms in Mathematica’s output to facilitate recognition of the limits.
11.5.2 Logistic Growth

Consider next Eq. (11.22) governing the logistic growth of a population \( N \) in an environment with carrying capacity \( N_c \). We seek the solution with the Mathematica statements

\[
\text{In}[1]:= \text{eq} = \frac{n'[t]}{n[t]} = k \cdot n[t] \cdot (1-n[t]/n_c);
\]

Enter equation. (We use \( n \) rather than \( N \) because \( N \) is a reserved symbol in Mathematica.)

\[
\text{In}[2]:= \text{soln} = \text{DSolve}[ \{\text{eq}, n[0]==n0\}, n, t ] ;
\]

Solve equation for \( n \) as function of \( t \), imposing the general initial condition \( n(0) = n_0 \). Mathematica has produced a warning message that, in this context, we can safely ignore.

\[
\text{In}[3]:= \text{soln} = \text{soln[[1,1]]};
\]

Elevate function definition to top level.

We can verify satisfaction of the initial condition and the differential equation with the statements

\[
\text{In}[4]:= n[0] /. \text{soln}
\]

Verify satisfaction of initial condition.

\[
\text{Out}[4]= n_0
\]

\[
\text{In}[5]:= \text{Simplify}[ \text{eq} /. \text{soln} ]
\]

Verify satisfaction of ODE.

\[
\text{Out}[5]= \text{True}
\]

Verifying the limit as \( t \to \infty \) is more complicated because, in the most straightforward statement, Mathematica’s command \text{Limit} will assume \( k \) to be complex and the command has no feature allowing us to tell Mathematica otherwise. Instead, we use the statements

\[
\text{In}[6]:= \text{qn} = n[t] /. \text{soln}
\]

\[
\text{Out}[6]= \frac{e^{kt}n_0n_c}{-n_0 + e^{kt}n_0 + n_c}
\]

\[
\text{In}[7]:= \text{qn} /. \text{Exp}[k*t] \rightarrow \text{tmp}
\]

\[
\text{Out}[7]= \frac{\text{tmp} n_0 n_c}{-n_0 + \text{tmp} n_0 + n_c}
\]

to obtain an expression giving the solution and then cast that expression in terms of the single variable \( \text{tmp} \) for \( e^{kt} \). Then, since \( k > 0 \) and—hence—\( e^{kt} = \text{tmp} \to \infty \) as \( t \to \infty \), we can determine the solution at large \( t \) with the statement

\[
\text{In}[8]:= \text{Limit}[ \%, \text{tmp} \rightarrow \text{Infinity} ]
\]

\[
\text{Out}[8]= n_c
\]

As expected, the solution approaches the carrying capacity, regardless of the initial population. The solution therefore passes all three tests.

Let us conclude the discussion of this example by displaying the evolution of the population in a graph. We begin by casting the expression to be plotted in a dimensionless form, choosing \( n_c \) as the unit in terms of which to express the population. We have in line \text{Out}[6] already obtained the expression to be plotted. We proceed with the statements

\[
\text{In}[9]:= \text{qn} = \text{qn}/n_c
\]

\[
\text{Out}[9]= \frac{e^{kt}n_0}{-n_0 + e^{kt}n_0 + n_c}
\]

Measure population \( n \) in units of \( n_c \).
Chapter 11. Solving Ordinary Differential Equations

In[10]:= qn=Simplify[ qn/.n0->\[Alpha]*nc ]    Recast to measure initial population in units of \( n_c \).

Out[10]= \( e^{kt} \alpha_1 + (\frac{-1 + e^{kt} \alpha}{1 + (-1 + e^{kt}) \alpha}) \)

Finally, we make one further recasting to measure time in units of \( 1/k \) with the statement

In[11]:= qn = qn /. t -> \[Tau]/k

Out[11]= \( e^{\tau} \alpha_1 + (\frac{-1 + e^{\tau} \alpha}{1 + (-1 + e^{\tau}) \alpha}) \)

With these rescalings, we have but one parameter [physically (biologically?) the initial population measured in units of the carrying capacity]. We produce the graph in Fig. 11.7 with the statements

In[12]:= q1 = qn /. \[Alpha] -> 0.25;
In[13]:= q2 = qn /. \[Alpha] -> 0.5;
In[14]:= q3 = qn /. \[Alpha] -> 1.0;
In[15]:= q4 = qn /. \[Alpha] -> 2.0;
In[16]:= q5 = qn /. \[Alpha] -> 4.0;
In[17]:= Plot[ {q1,q2,q3,q4,q5}, {\[Tau], 0.0, 5.0}, PlotRange -> {0.0, 4.0},
               PlotStyle -> Thickness[0.01], PlotLabel->StyleForm[ "Logistic Behavior",
               FontFamily->"Times", FontSize -> 14,
               FontWeight->"Bold"], AxesLabel->{"kt", "n/nc"} ]

Regardless of the initial population, all solutions converge monotonically on the carrying capacity (1.0 in the units we are using). Further, if we execute the statements

In[18]:= q6 = qn /. \[Alpha] -> 0.05;
In[19]:= Plot[ q6, {\[Tau], 0.0, 10.0},
               PlotRange -> {0.0, 1.0}, PlotLabel->StyleForm[ "Logistic Behavior",
               FontFamily->"Times", FontSize -> 14,
               FontWeight->"Bold"], AxesLabel->{"kt", "n/nc"} ]

we generate the graph in Fig. 11.8, which shows the initial exponential growth when the population is well below the carrying capacity but then reveals the leveling off as the population approaches the carrying capacity.

11.5.3 Damped Harmonic Oscillator

A third example of the use of Mathematica to solve a single second-order, linear, constant-coefficient, homogeneous, ordinary differential equation was presented in Chapter 8. You are invited to review the discussion of the damped harmonic oscillator in Sections 8.8.9 and 11.1.4.

11.5.4 Chain Radioactive Decay

As a first example of the use of Mathematica to solve a system of linear equations, let us determine the behavior of the radioactive decay chain described by Eq. (11.19). For simplicity, we suppose the initial values \( A(0) = A_0 \) and \( B(0) = C(0) = 0 \), i.e., we start with some \( A \) and no \( B \) or \( C \). The temporal evolution of this system is found with the Mathematica statements
Figure 11.7: Logistic growth or decay of a population when, starting with the highest graph, the initial population is 4.0, 2.0, 1.0, 0.5, and 0.25 times the carrying capacity of the environment.

In[1]:= eqA = A'[t] == -kA*A[t];
In[2]:= eqB = B'[t] == kA*A[t] - kB*B[t];
In[3]:= eqC = C'[t] == kB*B[t];
In[4]:= soln = DSolve[ {eqA, eqB, eqC,
  A[0]==A0, B[0]==0, C[0]==0},
  {A,B,C}, t ]

Enter equations, suppressing display.
Request solution, incorporating initial conditions in request.

Mathematica displays the fairly complicated result
Out[4] := \{\{ A \rightarrow \text{Function}[\{t\}, A_0 e^{-k_A t}], \\
B \rightarrow \text{Function}[\{t\}, -\frac{A_0 e^{-k_A t - k_B t}(-e^{k_A t} + e^{k_B t})k_A}{k_A - k_B}], \\
C \rightarrow \text{Function}[\{t\}, \frac{A_0 e^{-k_A t - k_B t}(-e^{k_A t}k_A + e^{k_A t}k_B - e^{k_A t + k_B t}k_B)}{k_A - k_B}]\}\}\)

To cast these results in a more interpretable form, let us extract each solution as an expression with the statements

Out[5] = A_0 e^{-k_A t}

Out[6] = A_0 (e^{k_A t} - e^{-k_B t})k_A

In[7] := qC = Simplify[ C[t] /. soln[[1]] ]
Out[7] = A_0 (k_A - e^{-k_B t}k_A + (-1 + e^{-k_A t})k_B)

These forms make evident one important conclusion: in this chain decay, we see a linear combination of two exponential decays, each with its own distinct halflife.

We conclude the discussion of this example by generating a graph of A(t), B(t), and C(t) for a specific set of values. Anticipating that we will set k_A = k_B for the graphs and noting that the expressions obtained for B(t) and C(t) are indeterminate when k_A = k_B, we prepare for the graph by evaluating the limits of qB and qC as k_B \rightarrow k_A. The statements

In[8] := qBeq = Limit[ qB, k_B \rightarrow k_A ]
Out[8] = A_0 e^{-k_A t}k_A t

In[9] := qCeq = Limit[ qC, k_B \rightarrow k_A ]
Out[9] = A_0 e^{-k_A t}(-1 + e^{k_A t} - k_A t)

will evaluate those limits, leaving each expressed in terms of the parameters k_A and A_0. Then, to obtain the desired graph, we specify values for A_0 and k_A with the statements

In[10] := A_0 = 1000; k_A = 0.1;

and produce the graph with the single statement

In[11] := Plot[ {qA, qBeq, qCeq}, \{t, 0.0, 50.0\}, PlotRange -> \{0.0,1000.0\},
PlotStyle -> Thick, PlotLabel->StyleForm[ "Radioactive Decay",
FontFamily->"Times", FontSize -> 14,
FontWeight->"Bold"], AxesLabel->\{"t", "A,B,C"\} ]

The resulting graph is shown in Fig. 11.9.
11.5. SOLVING ODES SYMBOLICALLY WITH MATHEMATICA

Figure 11.9: Radioactive decay of A, B, and C. The graph starting at 1000 is A; the graph rising to 1000 at $t = 50.0$ is C; and the remaining graph is B.

11.5.5 Coupled Oscillators

Let us next solve for the motion of the coupled oscillators described in Section 11.1.6. We shall address the problem in the dimensionless form presented in Eq. (11.36), imposing the general initial conditions in Eq. (11.37). For simplicity, however, we will take the initial velocities both to be zero. First, suppressing the display of Mathematica’s response, we enter the equations with the statements:

In[1]:= eq1 = x1''[t] + x1[t] - \[Kappa]*(x2[t]-x1[t]) == 0;
In[2]:= eq2 = x2''[t] + x2[t] + \[Kappa]*(x2[t]-x1[t]) == 0;

Then we create a list containing the initial values with the statement

In[3]:= ics = { x1[0]==x10, x2[0]==x20, x1'[0]==0, x2'[0]==0 };

and, finally, we request the solution with the statement

In[4]:= soln = DSolve[ Union[{eq1,eq2}, ics], {x1, x2}, t ]
Out[4]= {{x1 \to mess, x2 \to mess}}

While correct, the (suppressed) output from this statement is unfortunately in a form that is difficult to fathom. To rectify that situation, we

- execute the statement

In[5]:= soln1 = soln /. Sqrt[-1-2*\[Kappa]] -> I*Sqrt[1+2*\[Kappa]];

\footnote{You are urged to explore other initial conditions.}
\footnote{In the derivation of the dimensionless equations, we used overbars to identify the dimensionless quantities. We here drop that refinement in the notation, omitting the overbars but understanding that the quantities represented are still dimensionless.}
to replace several occurrences of $\sqrt{-1-2\kappa}$ with $i\sqrt{1+2\kappa}$, thus making the imaginary character of $\sqrt{-1-2\kappa}$ explicit. (Remember that $\kappa > 0$.)

- extract expressions for the position of each oscillator as a function of time with the statements

\[
\text{In}[6]:= qx1 = x1[t] /. \text{soln1}[[1]]; \\
\text{In}[7]:= qx2 = x2[t] /. \text{soln1}[[1]]; \\
\]

- execute the statements

\[
\text{In}[8]:= qx1a = \text{Expand}[qx1 ]; \\
\text{In}[9]:= qx2a = \text{Expand}[qx2 ]; \\
\]

to simplify the resulting expressions by multiplying several exponentials together.

- convert complex exponentials to trigonometric functions with the statements

\[
\text{In}[10]:= qx1b = \text{ExpToTrig}[qx1a ]; \\
\text{In}[11]:= qx2b = \text{ExpToTrig}[qx2a ]; \\
\]

At this point, we recognize that the time dependence of each term in these solutions has either $\cos t$ or $\cos [t\sqrt{1+2\kappa}]$ as a factor. Thus, we collect those terms with the nested statements

\[
\text{In}[12]:= qx1c = \text{Collect}[\text{Collect}[qx1b, \cos t], \cos [t\sqrt{1+2\kappa}]] \bigg] \\
\text{Out}[12]= \left(\frac{x_{10}}{2} + \frac{x_{20}}{2}\right) \cos t + \left(\frac{x_{10}}{2} - \frac{x_{20}}{2}\right) \cos [t\sqrt{1+2\kappa}] \\
\text{In}[12]:= qx2c = \text{Collect}[\text{Collect}[qx2b, \cos t], \cos [t\sqrt{1+2\kappa}]] \bigg] \\
\text{Out}[12]= \left(\frac{x_{10}}{2} + \frac{x_{20}}{2}\right) \cos t + \left(-\frac{x_{10}}{2} + \frac{x_{20}}{2}\right) \cos [t\sqrt{1+2\kappa}] \\
\]

At this point, we finally recognize that the solution is a superposition of different sinusoidal oscillations, one at frequency 1 (in units of $\omega = \sqrt{k/m}$) and the other at frequency $\sqrt{2\kappa+1}$ (again in units of $\omega$). Clearly also, only the term at frequency 1 is present if $x_{20} = x_{10}$ (the lower-frequency normal mode, in which the two objects oscillate with equal amplitude and in phase), and only the term at frequency $\sqrt{2\kappa+1}$ is present if $x_{20} = -x_{10}$ (the higher-frequency normal mode, in which the objects oscillate with equal amplitude but out of phase).

### 11.5.6 Standing Waves in a String

Boundary value problems need to be treated differently from initial value problems. To find the standing waves in a string that satisfy Eqs. (11.56) and (11.57), for example, we might enter and solve the equation subject to zero boundary conditions at each end of the string with the statements

\[
\text{In}[1]:= \text{eq} = f''[x] + k^2 * f[x] == 0; \\
\text{In}[2]:= \text{soln} = \text{DSolve}[\{\text{eq, f[0]==0, f[l]==0}\}, f, x] \\
\text{Out}[2]= \{f \rightarrow \text{Function}\left[x, \begin{cases} \\
C[1] \sin \sqrt{k^2} x & n \in \text{Integers} \& n \geq 1 \& k^2 = \frac{n^2\pi^2}{l^2} \& \ell > 0 \\
0 & \text{True} \end{cases}\right]\} \\
\]
11.5. SOLVING ODES SYMBOLICALLY WITH MATHEMATICA

The version of Mathematica here in use recognizes that only the trivial solution exists unless \( k \) has one of the special values \( k = n\pi/l, n \in \text{integer} \). In the special case, we learn that

\[
f(x) = C_1 \sin \frac{n\pi x}{\ell}
\]

where \( C_1 \) is undetermined and \( n \) assumes any of the values 1, 2, 3, \ldots (but not the value 0—because the solution reduces to zero in that case—and not negative values—because they really change only the overall sign of the solution and could be seen simply as an alternative choice of \( C_1 \)). Further, since \( \omega = kc \), we would associate with this solution the frequency

\[
\omega_n = kn\frac{c}{\ell} = n\omega_1
\]

where \( \omega_1 \) is the fundamental frequency. We find, in particular, that all allowed frequencies are integer multiples of the fundamental frequency.

Earlier versions of Mathematica concluded that only the trivial solution to the ODE can be made to satisfy both boundary conditions and returned only the solution

\[
\text{Out}[2]= \{ \{ f \rightarrow \text{Function}[\{x\}, 0] \} \}
\]
to the statement at \text{In}[2] above. These versions of Mathematica failed to recognize that we can find non-trivial solutions, but only if we are prepared to constrain \( k \) to a limited set of values.

In those versions, we must evidently be more deliberate. To help Mathematica recognize that the problem is a bit more subtle, we simply solve the differential equation without specifying the boundary conditions, thereby generating a solution that includes two initially undetermined integration constants. Then, we impose the boundary conditions. First, we solve the equation with the statement

\[
\text{In}[3]:= \text{soln} = \text{DSolve}[\text{eq}, f, x][[1,1]]
\]


\[
\]
elevating the solution with \([1,1]\) to the top level in its expression. Next, we determine the consequences of the boundary conditions with the statements

\[
\text{In}[4]:= \text{eq1} = 0 == f[0] /. \text{soln} \quad \text{Impose condition at } x = 0.
\]

\[
\text{Out}[4]= 0 == C[1]
\]

\[
\text{In}[5]:= \text{eq2} = 0 == f[l] /. \text{soln} \quad \text{Impose condition at } x = \ell.
\]

\[
\text{Out}[4]= 0 == C[1] \cos[k\ell] + C[2] \sin[k\ell]
\]

At this point, we recognize that we have a pair of equations to be solved simultaneously for \( C[1] \) and \( C[2] \). We also recognize, however, that these equations are homogeneous and will in fact have non-trivial solutions only if the determinant of the coefficient matrix happens to be zero. To complete the problem, then, we must extract that coefficient matrix, set its determinant to zero, and solve the resulting equation to find any values of \( k \) that may admit non-trivial solutions for \( C[1] \) and \( C[2] \). To find the coefficient matrix, we invoke the Mathematica statement

\[
\text{In}[5]:= \text{tmp} = \text{CoefficientArrays}[\{\text{eq1}, \text{eq2}\}, \{C[1], C[2]\}]
\]

whose output is a bit obscure. Even so, the statement

\[
\text{Out}[5]= \{ \text{SparseArray}[ \text{mess}\], \text{SparseArray}[ \text{mess}\] \}
\]

\[18\] In some earlier versions of Mathematica, the necessary functionality resided in the older command \text{LinearEquationsToMatrices} in the package \text{LinearAlgebra}. While still available, that entire package is now obsolete and may not be included in future versions of Mathematica.

\[19\] In the output from Mathematica, all braces will be the same size. We have here used different sizes to clarify the structure of that output.
In[6]:= tmp1 = Normal[%]
Out[6]=
\{\{0, 0\}, \{-\cos[k\ell], -\sin[k\ell]\}\}\}

yields a more transparent result. Here, the output is a list of lists, the first of which conveys the inhomogeneties in the equations while the second, itself a list of lists, provides the coefficient matrix (though the minus signs in several places are a bit mysterious). We can extract and display the coefficient matrix with the statements

In[7]:= coefmat = tmp1[[2]];
In[8]:= MatrixForm[ coefmat ]
Out[8]=
\[-1 0
-\cos[k\ell] -\sin[k\ell]\]

Next, we seek values of \(k\) for which the determinant of this matrix will be zero by invoking the statements

In[9]:= charpol = Det[ coefmat ]
Out[9]= \(\sin[k\ell]\)
In[10]:= rts = Solve[ charpol==0, \(k\) ]
Evaluate determinant.
Solve characteristic equation for \(k\).
Out[10]= \{\{\(k\) \rightarrow \text{ConditionalExpression} \left[ \frac{2\pi C[1]}{\ell}, C[1] \in \text{Integers} \right],

\(k\) \rightarrow \text{ConditionalExpression} \left[ \frac{\pi + 2\pi C[1]}{\ell}, C[1] \in \text{Integers} \right] \}\}\}

Mathematica recognizes that this equation has numerous roots.\(^{20}\) Note that the \(C[1]\) in these statements is not the same as the \(C[1]\) at Out[2] and In[4]–Out[5] and in Eq. (11.71). To avoid confusion, let’s simplify the appearance here by replacing this \(C[1]\) with \(n\) with the statement

In[11]:= rts1 = rts /. C[1] -> \(n\)
Out[11]= \{\{k \rightarrow \text{ConditionalExpression} \left[ \frac{2\pi n}{\ell}, n \in \text{Integers} \right],

k \rightarrow \text{ConditionalExpression} \left[ \frac{\pi + 2\pi n}{\ell}, n \in \text{Integers} \right] \}\}\}

In this form, it is easier to see that, with \(n\) an integer, \(k = 2n\pi/\ell\) or \(k = (2n + 1)\pi/\ell\), i.e., \(k\) is either an even or an odd multiple of \(\pi/\ell\). More simply, \(k = n\pi/\ell\) for \(n\) any integer. With this recasting, we use the statement

In[12]:= rl = \(k \rightarrow n\pi/\ell;\)
to define a transformation rule for substituting \(n\pi/\ell\) for \(k\) in any expression. With this rule, we quickly find the corresponding values of \(C[1]\) and \(C[2]\) by returning to equations eq1 and eq2 and executing the statements

In[13]:= eq1 /. rl
Out[13]= 0 == \(C[1]\)
In[14]:= Simplify[ eq2 /. rl, \(\text{Element}[n,\text{Integers}]\) ]
Out[14]= 0 == \(C[1]\)

\(^{20}\)An earlier version of Mathematica returned the single (correct) root \(k = 0\). The behavior has changed in newer versions of Mathematica.
Clearly, we must choose $C[1]$ to be 0—a fact we really knew at the very beginning, but we learn nothing about $C[2]$, which evidently remains arbitrary insofar as the conditions of the problem are concerned.

In the end, we discover that we can find acceptable solutions to the original boundary value problem only when $k$ has one of the values $n\pi/\ell$ and $C[1]$ is set to zero. Within *Mathematica*, we then find that the solution to the original boundary value problem is given by

\[
\text{In}[15]:= \text{stdwv} = f[x] /. \text{solin}
\]

\[
\]

\[
\text{In}[15]:= \text{stdwv} = \text{stdwv} /. \{\text{r1, C[1]}\to 0\}
\]

\[
\text{Out}[15]= C[2] \sin[n\pi x/\ell]
\]

where $n$ assumes any of the values 1, 2, 3, ... (but not the value 0—because the solution reduces to zero in that case—and not negative values—because they really change only the overall sign of the solution and could be seen simply as an alternative choice of $C[2]$). After this more arduous route, we arrive again at the solution expressed in Eqs. (11.71) and (11.72).

### 11.5.7 Infinite Depth Quantum Well

As pointed out in Section 11.1.9, the quantum problem of a particle in an infinitely deep, one-dimensional, square potential well is mathematically identical to the problem we have just addressed. Since the allowed values of $k$ have turned out to be $n\pi/\ell$, we then conclude that the allowed energies in the quantum problem would be

\[
E_n = \frac{\hbar^2 k_n^2}{2m} = \frac{n^2 \hbar^2 \pi^2}{2m \ell^2}
\]

(11.73)

Further, we would associate the (unnormalized) wave function

\[
\psi_n(x) = N_n \sin\left(\frac{n\pi x}{\ell}\right)
\]

(11.74)

with the energy $E_n$. In the quantum problem, the allowed energies are not integer multiples of the lowest energy; rather, the allowed energies increase in proportion to the squares of the integers.

### 11.5.8 Finding Solutions by Series Methods

Particularly when analytic methods appear to be yielding little result, we are sometimes tempted to invoke the *method of Frobenius*, representing the solution to the ODE of interest as a power series of the form

\[
f(x) = \sum_{i=0}^{\infty} a_i x^{i+\gamma}
\]

(11.75)

and seeking both an *indicial equation* determining $\gamma$ and *recursion* relationships determining $a_i$. In the absence of a simple *Mathematica* command to facilitate this approach to solving an ODE, we must either write our own package (which is beyond the scope of this book) or step *Mathematica* systematically through the process. For the sake of concreteness, suppose we seek a solution to the equation

\[
\frac{d^2 y}{dx^2} + y = 0
\]

We might then define a (truncated) series for the solution with the statement

\[
\text{In}[1]:= y[x_] := \text{Sum}[a[i] x^{i+\gamma}, \{i, 0, 8\}]
\]
We could, of course, extend the series as far as we wished. We here truncate it with a few terms to keep the output to manageable size.) Then we would evaluate the right-hand side of the ODE with the statement

```
In[2]:= y''[x] + y[x]
Out[2]= (expression too long to record)
```

Not surprisingly, each term has a factor of $x^\gamma$. Since we ultimately will set this expression equal to zero, we can divide the expression by $x^\gamma$, and then—since a premultiplying factor of $1/x^2$ appears—multiply the entire expression by $x^2$. The necessary Mathematica statement and its output are

```
In[3]:= soln1 = x^2*Simplify[ %/x^\[Gamma] ]
Out[3]= (-1 + \gamma)\gamma a_0 + x\gamma(1 + \gamma)a_1 + x^2(a_0 + (2 + 3\gamma + \gamma^2)a_2 + ...)  
```

We have written out only the terms with the lowest powers of $x$ for the moment because they are the terms on which we must focus first. Since this series must be identically zero for all $x$ and since different powers of $x$ are linearly independent, the series will sum to zero only if the coefficient of each power of $x$ is separately zero. To simplify the application of this criterion, we might extract the coefficients themselves as a list with the statement

```
In[4]:= coefs = CoefficientList[ soln1, x ]
Out[4]= {(-1 + \gamma)\gamma a_0, \gamma(1 + \gamma)a_1, a_0 + (2 + 3\gamma + \gamma^2)a_2, ...}
```

Following the method of Frobenius, we determine acceptable values of $\gamma$ by requiring that the coefficient of the lowest power of $x$ (i.e., the first element in this list) be zero under the restriction that $a_0 \neq 0$. The statements

```
In[5]:= coefs[[1]]/a[0]
Out[5]= (-1 + \gamma)\gamma
In[6]:= gam = Solve[ % == 0, \[Gamma] ]
Out[6]= {{\gamma \to 0}, {\gamma \to 1}}
```

reveal that $\gamma$ can be either 0 or 1. When $\gamma = 1$, the coefficients of successive powers of $x$ are

```
In[7]:= coeffs1 = coefs /. gam[[2]]
Out[7]= {0, 2a_1, a_0 + 6a_2, a_1 + 12a_3, a_2 + 20a_4, a_3 + 30a_5, a_4 + 42a_6, a_5 + 56a_7, a_6 + 72a_8, a_7, a_8}
```

Each of these coefficients must itself be zero. Ignoring the first coefficient (which is already zero) and the last two coefficients (which are incomplete because of our truncation of the series), we extract a set of eight simultaneous linear equations and solve them for the coefficients with the statements

```
In[8]:= eqs1 = Table[ coeffs1[[i+1]]==0, {i, 8} ]
Out[8]= {2a_1 == 0, a_0 + 6a_2 == 0, a_1 + 12a_3 == 0, ...}
In[9]:= vars1 = Table[ a[i], {i, 8} ]
Out[9]= {a_1, a_2, a_3, ...}
```

---

21To shorten the expression, we use the symbol $a_i$ rather than $a[i]$ for the $i$-th coefficient in the original series.
In[10]:= Solve[ eqs1, vars1 ][[1]]
Out[10]= 
{ 
    a1 -> 0, a2 -> -\frac{a0}{6}, a3 -> 0, a4 -> \frac{a0}{120}, a5 -> 0, a6 -> -\frac{a0}{5040}, a7 -> 0, a8 -> \frac{a0}{362880} 
}

Finally, we can substitute these solutions into the original series with the statement

In[11]:= y[x] /. \[Gamma] -> 1;
In[12]:= % /. %10;
In[13]:= Collect[ %, a[0] ]
Out[13]= 
\left( x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \frac{x^9}{362880} \right) a_0

Recognizing that 3! = 6, 5! = 120, 7! = 5040, and 9! = 362880, we then also recognize that the series in parentheses is in fact the beginnings of the series for \text{Sin}[x] and this solution apparently is \text{a}_0 \text{Sin}[x].

The solution when \gamma = 0 can be found in much the same way. We begin by invoking the statement

In[14]:= coefs0 = coefs /. gam[[1]]
Out[14]= 
{0, 0, a_0 + 2a_2, a_1 + 6a_3, a_2 + 12a_4, a_3 + 20a_5, a_4 + 30a_6, a_5 + 42a_7, a_6 + 56a_8, a_7, a_8}

to determine the coefficients for this case. As before, each of these coefficients must itself be zero. Ignoring the first two coefficients (which are already zero) and the last two (which are incomplete because of our truncation of the series), we extract a set of seven simultaneous linear equations and solve them for the coefficients with the statements

In[15]:= eqs0 = Table[ coefs0[[i+2]] == 0, {i, 7} ]
Out[15]= 
{a_0 + 2a_2 == 0, a_1 + 6a_3 == 0, a_2 + 12a_4 == 0, a_3 + 20a_5 == 0, ... }
In[16]:= vars0 = Table[ a[i+1], {i, 7} ]
Out[16]= 
{a_2, a_3, a_4, ...}
In[17]:= Solve[ eqs0, vars0 ][[1]]
Out[17]= 
\left\{ 
    a_2 -> -\frac{a_0}{2}, a_3 -> -\frac{a_1}{6}, a_4 -> \frac{a_0}{24}, a_5 -> \frac{a_1}{120}, a_6 -> -\frac{a_0}{720}, a_7 -> -\frac{a_1}{5040}, a_8 -> \frac{a_0}{40320} \right\}

Here, we have noted that the conditions imposed on the solution do not constrain either \text{a}_0 or \text{a}_1. Finally, we can substitute these solutions into the original series with the statement

In[18]:= y[x] /. \[Gamma] -> 0;
In[19]:= % /. %17;
In[20]:= Collect[ %, \{a[0], a[1]\} ]
Out[20]= 
\left( 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720} + \frac{x^8}{40320} \right) a_0 + \left(x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040}\right) a_1

\text{This value of} \gamma \text{has, in fact, yielded a solution containing two arbitrary constants, exactly what we expect in the general solution to a second-order ODE. Interestingly, the series multiplied by} a_1 \text{in this result is identical to the (only) series obtained when} \gamma \text{was set to the value 1, i.e., it is again the function} \text{Sin}[x]. \text{Recognizing that 2! = 2, 4! = 24, 6! = 720, and 8! = 40320, we recognize that the other series in this result—the one multiplied by} a_0\text{ —is in fact the beginnings of the series for} \text{Cos}[x]. \text{Thus, the solution obtained by the method of Frobenius is apparently} a_0 \text{Cos}[x] + a_1 \text{Sin}[x], \text{where} a_0 \text{and} a_1 \text{are to be determined by initial or boundary conditions.
11.6 Algorithms for Solving ODEs Numerically

A numerical solution to an ordinary differential equation emerges from the application of a procedure—frequently called an algorithm—for calculating approximate values of the dependent variables at a succession of values of the independent variable. All numerical methods for solving ordinary differential equations exploit the fact that the differential equations determine the rates of change of the dependent variables from the dependent variables themselves. Because solutions obtained numerically are approximate, we must give attention not only to the methods themselves but also to means by which we can assess the accuracy of the solutions obtained.

For the sake of a simple discussion, we shall, in laying out the essence of each of several algorithms, suppose that we are dealing with a single first-order ODE and an initial condition of the form

$$\frac{dx}{dt} = f(x,t) \quad ; \quad x(0) = x_0$$

(11.76)

where \(f(x,t)\) and \(x_0\) are known from the beginning. The relatively straightforward extension of the initial discussion to systems of first-order equations and to single equations of higher order will be illustrated in several specific examples but will not be explicitly discussed in general terms.

11.6.1 Euler’s Method

Euler’s method, which embodies the simplest numerical approach to ODEs, is based on the assumption that the rates of change of the dependent variables do not themselves change very quickly. Thus, given a short enough time interval \(\Delta t\), the rates of change throughout that interval may be regarded, at least approximately, as constant and equal to the rates of change at the beginning of the interval. For example, provided \(\Delta t\) is not too large,\(^{22}\) we can write the approximation

$$\frac{dx}{dt} = f(x,t) \quad \Rightarrow \quad \frac{x(t+\Delta t) - x(t)}{\Delta t} \approx f(x,t) \quad \Rightarrow \quad x(t+\Delta t) \approx x(t) + f(x,t) \Delta t$$

(11.77)

The value \(x(t+\Delta t)\) of the dependent variable at \(t+\Delta t\) is the value \(x(t)\) at time \(t\) plus the amount \(\Delta x\) by which \(x\) changes in the interval, where we estimate \(\Delta x\) by multiplying the rate of change of \(x\) given by \(f(x(t),t)\) at the beginning of the interval by the elapsed time \(\Delta t\), i.e., \(\Delta x = f(x(t),t) \Delta t\).

In an alternative notation, if we think of \(t\) as the “old” time and \(t+\Delta t\) as the “new” time, we might express the basic stepping equations as

$$f_{\text{old}} = f(x_{\text{old}}, t_{\text{old}}) \quad ; \quad x_{\text{new}} = x_{\text{old}} + f_{\text{old}} \Delta t \quad ; \quad t_{\text{new}} = t_{\text{old}} + \Delta t$$

(11.78)

Starting from the initial condition as the first \(x_{\text{old}}\) (and a choice of time step \(\Delta t\)—see later), we can then use these stepping equations repeatedly to move from knowledge of \(x(0)\) to knowledge of \(x(\Delta t)\) to knowledge of \(x(2 \Delta t)\) to . . . , continuing as long as our patience endures. The basic strategy of Euler’s method would be summarized in the algorithm in Table 11.1.\(^{23}\) As this example shows, we can generate a complete—though approximate—solution from initial knowledge of only

1. The differential equation, which determines the rate of change of the dependent variable from the dependent and independent variables themselves,

2. The initial condition, which starts the process by providing the first row in a table containing \(t, x, \) and \(f,\)

\(^{22}\)The precise meaning of “too large” is difficult to define. In general terms, if \(T\) is a typical time during which the solution changes appreciably (say by 10–20%), then a value of \(\Delta t < T\) will probably yield an adequate solution. Each case must be examined on its own terms; no general rules can be formulated.

\(^{23}\)In many cases, the ultimate and penultimate pairs of statements in the loop can be combined into a single pair of statements. We here refrain from that more compact expression so that the two distinct operations—calculating the new to complete the step and replacing the old with the new to prepare for the next step—will remain distinct.
3. Specific values for any parameters—here there happen to be none—in the differential equation, and

4. A choice of time step $\Delta t$.

With this information as input, each pass through the loop in the above algorithm generates a new row in the table containing $t$, $x$, and $f$.

To be even more concrete (and to illustrate the simple extension to a system of ODEs), let us work out by hand the first few steps in the Euler solution of Eq. (11.19) describing a chain radioactive decay. Reflecting the differential equations, the stepping equations for this specific case are

$$A(t + \Delta t) = A(t) + \frac{dA}{dt}(t) \Delta t = A(t) - k_A A(t) \Delta t \tag{11.79}$$

$$B(t + \Delta t) = B(t) + \left[k_A A(t) - k_B B(t)\right] \Delta t \tag{11.80}$$

$$C(t + \Delta t) = C(t) + k_B B(t) \Delta t \tag{11.81}$$

For definiteness,\(^{24}\) we take $A(0) = 1000.000$, $B(0) = C(0) = 0.000$, and $k_A = k_B = 0.100$, and we select a time step of $\Delta t = 0.250$.\(^{25}\) Equations (11.79), (11.80), and (11.81) at $t = 0.0$ then become

$$A(0.00 + 0.25) = 1000.0 - 0.1 \cdot 1000.0 \cdot 0.25 = 975.000 = A(0.25) \tag{11.82}$$

$$B(0.00 + 0.25) = 0.0 + \left[0.1 \cdot 1000.0 - 0.1 \cdot 0.0\right] \cdot 0.25 = 25.000 = B(0.25)$$

$$C(0.00 + 0.25) = 0.0 + 0.1 \cdot 0.0 \cdot 0.25 = 0.000 = C(0.25)$$

Continuing the algorithm further by applying Eqs. (11.79), (11.80), and (11.81) for $t = 0.25$, we find that

$$A(0.25 + 0.25) = 975.0 - 0.1 \cdot 975.0 \cdot 0.25 = 950.625 = A(0.50) \tag{11.83}$$

$$B(0.25 + 0.25) = 25.0 + \left[0.1 \cdot 975.0 - 0.1 \cdot 25.0\right] \cdot 0.25 = 48.750 = B(0.50)$$

$$C(0.25 + 0.25) = 0.0 + 0.1 \cdot 25.0 \cdot 0.25 = 0.625 = C(0.50)$$

---

\(^{24}\) One disadvantage of numerical approaches is that they are, indeed, numerical. We cannot find solutions containing symbols representing parameters. We must seek solutions for specific numerical values. If we need to know the dependence of the solution on a particular parameter, we will have to generate a separate solution for each desired value of the parameter.

\(^{25}\) This choice at the moment is more for convenience than accuracy. The accuracy of the resulting solution is, of course, markedly influenced by the choice of $\Delta t$. We shall return to assess the suitability of this choice in later sections.
Table 11.2: Evolution of three-species radioactive decay determined by Euler’s method with $\Delta t = 0.25$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
<th>$A + B + C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>1000.000</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>

Going yet one more step, Eqs. (11.79), (11.80), and (11.81) for $t = 0.50$ yield that

\[
A(0.50 + 0.25) = 950.625 - 0.1 \cdot 950.625 \cdot 0.25 = 926.859 = A(0.75)
\]

\[
B(0.50 + 0.25) = 48.75 + \left[0.1 \cdot 950.625 - 0.1 \cdot 48.75\right] \cdot 0.25 = 71.267 = B(0.75)
\]

\[
C(0.50 + 0.25) = 0.625 + 0.1 \cdot 48.75 \cdot 0.25 = 1.844 = C(0.75)
\]

The resulting values are compiled in Table 11.2, in which the first row is provided by the initial conditions and all subsequent rows are determined by the algorithm outlined in Eqs. (11.79), (11.80), and (11.81). Knowing that an exact solution will reflect the conservation of $A + B + C$ [see Eq. (11.20)], we have added a column containing the sum of the amounts of all three species present. Clearly, carried this far, anyway, the solution appears automatically to satisfy that requirement.\(^{26}\)

### 11.6.2 Improved Euler Method

While Euler’s method is simple to motivate, describe, and implement, it unfortunately yields only a coarse approximation. Typically it will require a very small time step (which translates into a large amount of computational time and the potential accumulation of computer roundoff error) to achieve adequate accuracy. Considerable effort has been spent in devising alternative, more refined algorithms that converge more rapidly with a minimal amount of computational labor (and hence less internal roundoff error). The improved Euler method, for example, takes the Euler result at each step to be only an estimate (prediction) of the solution for that step and uses that estimate to refine the solution before going on to the next step. Starting with the values $x_{\text{old}}$ and $t_{\text{old}}$ and a chosen time step $\Delta t$, we

1. calculate $f_{\text{old}}$,

2. calculate the predicted values

\[
x_{\text{pred}} = x_{\text{old}} + f_{\text{old}} \Delta t \quad ; \quad t_{\text{new}} = t_{\text{old}} + \Delta t
\]

3. calculate $f_{\text{pred}} = f(x_{\text{pred}}, t_{\text{new}})$, and

4. calculate the final (corrected) value

\[
x_{\text{new}} = x_{\text{old}} + \frac{1}{2} \left(f_{\text{old}} + f_{\text{pred}}\right) \Delta t
\]

\(^{26}\)Actually, however, that the sum $A + B + C$ preserves its initial value is not really a check on the accuracy of the method in this example. If we simply add the three stepping equations we are using [Eqs. (11.79)–(11.81)], we find that $A(t + \Delta t) + B(t + \Delta t) + C(t + \Delta t) = A(t) + B(t) + C(t)$. Solutions generated by Euler’s method applied to this problem automatically satisfy the conservation law, regardless of $\Delta t$. 
11.6. ALGORITHMS FOR SOLVING ODES NUMERICALLY

Table 11.3: Improved Euler algorithm.

\[
\begin{align*}
dt & \leftarrow \Delta t \\
t_{old} & \leftarrow 0.0 \\
x_{old} & \leftarrow x_0 \\
\text{loop} & \\
\quad \text{fold} & \leftarrow f(x_{old}, t_{old}) \\
\quad \text{print}, t_{old}, x_{old}, \text{fold} & \\
\quad \text{exit} & \text{loop when done} \\
x_{pred} & \leftarrow x_{old} + \text{fold} \times dt \\
t_{new} & \leftarrow t_{old} + dt \\
f_{pred} & \leftarrow f(x_{pred}, t_{new}) \\
x_{new} & \leftarrow x_{old} + \frac{1}{2}(\text{fold} + f_{pred}) \times dt \\
x_{old} & \leftarrow x_{new} \\
t_{old} & \leftarrow t_{new} \\
\text{end_loop} & \\
\end{align*}
\]

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 \( \Delta t = 0.25 \).

\[
\begin{array}{cccc}
\text{t} & A & B & C \\
0.00 & 1000.000 & 0.000 & 0.000 \\
0.25 & 975.313 & 24.375 & 0.312 \\
0.50 & 951.234 & 47.547 & 1.219 \\
0.75 & 927.751 & 69.559 & 2.690 \\
\vdots & \vdots & \vdots & \vdots \\
\end{array}
\]

In this last step, we in effect estimate the average rate of change over the interval from \( t \) to \( t + \Delta t \) as the average of (a) its value \( f_{old} \) at the beginning of the interval and (b) the best estimate \( f_{pred} \) we have of its value at the end of the interval. In so doing, we admit that the rate of change may change in the interval. Intuitively, for a given \( \Delta t \), the improved Euler method will be more accurate than Euler’s method, which presumes that the average rate of change over the interval is adequately approximated by its value at the beginning of the interval.\(^{27}\)

A full laying out of this improved algorithm differs from the algorithm presented in the previous section in only a few lines, having the expression laid out in Table 11.3. As with Euler’s method, knowledge of the differential equation, the initial condition, and any parameters, together with a choice of a time step, starts a process that leads to (approximate) knowledge of the solution indefinitely into the future.

By way of example (but leaving the arithmetic to the reader), we present in Table 11.4 the results of applying the improved Euler method for the example treated in the previous section. Values in this table should be compared with those in Table 11.2.

\(^{27}\) Phrased as we have described it, the improved Euler method entails first calculating a \textit{predicted} value at the end point of an interval and then uses that value to determine a \textit{corrected} final value at that point before going on to the next point. Such methods are often called \textit{predictor-corrector} methods.
11.6.3 Runge-Kutta Methods

A popular alternative viewpoint, which leads to the deduction of what are called Runge-Kutta stepping algorithms, requires at base that two different Taylor series expansions of the solution match to a chosen number of terms. On the one hand, we know that

\[ x(t + \Delta t) = x(t) + \frac{dx(t)}{dt} \Delta t + \frac{1}{2} \frac{d^2x(t)}{dt^2} \Delta^2 t + O(\Delta^3 t) \]

\[ = x(t) + f(x, t) \Delta t + \frac{1}{2} \left( \frac{\partial f(x, t)}{\partial x} \frac{dx(t)}{dt} + \frac{\partial f(x, t)}{\partial t} \right) \Delta^2 t + O(\Delta^3 t) \]

\[ = x(t) + f(x, t) \Delta t + \frac{1}{2} \left( f(x, t) \frac{\partial f(x, t)}{\partial x} + \frac{\partial f(x, t)}{\partial t} \right) \Delta^2 t + O(\Delta^3 t) \quad (11.87) \]

Motivated by the improved Euler method on the other hand, we are tempted to suppose that we might build a stepping algorithm by\(^{28}\)

1. introducing a judicious time \(t' = t + a \Delta t\), where \(a\), constrained by \(0 \leq a \leq 1\), is yet to be specified,
2. introducing a tentative solution \(x' = x(t) + b \Delta t f(x(t), t)\), where \(b\), constrained by \(0 \leq b \leq 1\), is yet to be specified, and
3. taking \(x(t + \Delta t) = x(t) + \left( w_1 f(x(t), t) + w_2 f(x', t') \right) \Delta t\), where the weights \(w_1\) and \(w_2\) are yet to be specified.

The essential idea of the Runge-Kutta approach is to expand the assumption of item 3 as a power series in \(\Delta t\) and require that its first few terms agree with those in the series expressed in Eq. (11.87). To deduce this second Taylor series, we first invoke the two-dimensional Taylor series to find that

\[ f(x', t') = f \left( x + b \Delta t f(x, t), t + a \Delta t \right) \]

\[ = f(x, t) + \frac{\partial f(x, t)}{\partial x} b \Delta t f(x, t) + \frac{\partial f(x, t)}{\partial t} a \Delta t + O(\Delta^2 t) \quad (11.88) \]

where—noting at item 3 above that this term will ultimately be multiplied by \(\Delta t\)—we have included only terms through those first order in \(\Delta t\). Then, we find that

\[ x(t + \Delta t) = x(t) + w_1 f(x(t), t) \Delta t \]

\[ + w_2 \left( f(x, t) + \frac{\partial f(x, t)}{\partial x} b \Delta t f(x, t) + \frac{\partial f(x, t)}{\partial t} a \Delta t + O(\Delta^2 t) \right) \Delta t \]

\[ = x(t) + (w_1 + w_2) f(x(t), t) \Delta t \]

\[ + \left( w_2 b f(x(t)) \frac{\partial f(x, t)}{\partial x} + w_2 a \frac{\partial f(x, t)}{\partial t} \right) \Delta^2 t + O(\Delta^3 t) \quad (11.89) \]

which will match Eq. (11.87) through terms of order \(O(\Delta^2 t)\) if we choose

\[ w_1 + w_2 = 1 \quad axw_2 = bw_2 = \frac{1}{2} \quad \text{or} \quad w_1 + w_2 = 1 \quad a = b \quad w_2 = \frac{1}{2a} \quad (11.90) \]

\(^{28}\)Note that the steps laid out here reduce to the improved Euler method if we take \(a = b = 1\) and \(w_1 = w_2 = 1/2\).
11.6. ALGORITHMS FOR SOLVING ODES NUMERICALLY

Table 11.5: A Runge-Kutta Method.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( dt )</td>
<td>Choose time step.</td>
</tr>
<tr>
<td>( t_{old} )</td>
<td>Initialize ( t_{old} ).</td>
</tr>
<tr>
<td>( x_{old} )</td>
<td>Initialize ( x_{old} ).</td>
</tr>
<tr>
<td>fold</td>
<td>Evaluate ( f_{fold} ).</td>
</tr>
<tr>
<td>print</td>
<td>Display results.</td>
</tr>
<tr>
<td>k1</td>
<td>Find change using rate at ( t_{old} ).</td>
</tr>
<tr>
<td>k2</td>
<td>Find change using rate at ( t_{new} ).</td>
</tr>
<tr>
<td>tnew</td>
<td>Calculate ( t_{new} ).</td>
</tr>
<tr>
<td>xnew</td>
<td>Calculate ( x_{new} ).</td>
</tr>
<tr>
<td>xold</td>
<td>Replace old values with new.</td>
</tr>
</tbody>
</table>

We have discovered a multitude of Runge-Kutta schemes, one corresponding to each possible choice of \( a \), \( b \), \( w_1 \), and \( w_2 \). One common choice embodies the values in footnote 28—values that reveal that the improved Euler method is a member of this Runge-Kutta family. A second common choice takes \( w_1 = 0 \), \( w_2 = 1 \), and \( a = b = \frac{1}{2} \), in which case the stepping equations become

\[
t' = t + \frac{1}{2} \Delta t \quad; \quad x' = x + \frac{1}{2} f(x, t) \Delta t \quad; \quad x(t + \Delta t) = x(t) + f(x', t') \Delta t
\] (11.91)

With this scheme, we first step half way over the interval with Euler’s method, then we use that result to estimate the rate of change at the midpoint of the interval and, finally, we use the rate of change at the midpoint to project the solution at the end of the interval. Clearly, this route provides yet another means to estimate the average rate of change over the interval; it is called the midpoint method.

Even though the Runge-Kutta algorithm for \( w_1 = w_2 = \frac{1}{2} \) and \( a = b = 1 \) coincides with the improved Euler method, it is usually presented in the form shown in Table 11.5. This form is more compatible with the most convenient expressions of other algorithms in the broad Runge-Kutta family.

The algorithm described in the previous two paragraphs is known as a second-order Runge-Kutta algorithm because its deduction entailed matching terms in two Taylor expansions through those of order \( \Delta t^2 \). We could, of course, replace the expression in item 3 above with a more elaborate expression, expand it to include higher-order terms in \( \Delta t \), and insist on agreement with a higher-order Taylor expansion deduced from the differential equation. The calculational labor becomes increasingly complicated. Partly because of its popularity, we present without derivation the essence of a fourth-order Runge-Kutta algorithm, limiting ourselves only to the steps that would replace the ones calculating \( k_1 \), \( k_2 \), and \( x_{new} \) in the above algorithm:

\[
\begin{align*}
\text{k1} & \leftarrow f(x_{old}, t_{old}) \Delta t & \text{Find change using rate at beginning.} \\
\text{k2} & \leftarrow f(x_{old}+k1/2, t_{old}+dt/2) \Delta t & \text{Estimate change using rate at midpoint.} \\
\text{k3} & \leftarrow f(x_{old}+k2/2, t_{old}+dt/2) \Delta t & \text{Estimate change using refined rate at midpoint.} \\
\text{k4} & \leftarrow f(x_{old}+k3, t_{old}+dt) \Delta t & \text{Estimate change using rate at end.} \\
x_{new} & \leftarrow x_{old} + (k1+2*k2+2*k3+k4)/6 & \text{Calculate final value.}
\end{align*}
\]

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

\footnote{Actually, we choose only \( a \). Then \( b = a \), \( w_2 = 1/2a \), and \( w_1 = 1 - w_2 \) are fixed.}
given time step. Equivalently, a lower-order method requires a smaller time step to give the same degree of accuracy as a higher-order method.

### 11.6.4 Assessing Accuracy

Numerical evaluations, of course, only approximate the solution to ordinary differential equations. Thus, we cannot complete a solution without also assessing its accuracy. Furthermore, this task must be accomplished without knowledge of the exact solution. The importance of being aware that numerical methods are always approximate cannot be overstressed.

Two distinctly different sorts of errors can occur. Truncation errors arise because the solution has been based on a finite-difference approximation to the derivatives appearing in the equation; roundoff errors arise because computers do not store non-integers to 100% precision and, in an iterated calculation where each step depends on the previous step, the imprecision with which each component is represented within the computer can accumulate as the number of steps increases. Truncation errors become smaller as the step size is reduced. Roundoff errors, unfortunately, become more significant as step size is reduced (because, with smaller steps, more arithmetic must be done). Usually, roundoff errors are negligible, the more so as the sophistication of the algorithm increases (and, hence, the amount of arithmetic decreases). Provided we do not strive for accuracy greater than about 1 part in $10^5$ (with single precision floating point arithmetic), we can usually ignore roundoff errors. Thus, provided the solution we seek does not vary too rapidly on the time scale defined by the time step in use, the quickest way to obtain a reasonably reliable estimate of truncation error is to solve the equation with two different step sizes, the second being half of the first, and compare the two results. Presuming that roundoff error has not begun to be important, we can be confident that the second result is more accurate than the first. Thus, if the two agree to 1 part in $10^3$, say, we can with reasonable confidence assume that the second value is good to one part in $10^3$. Indeed, the second value is probably better than that, but assessing its accuracy by this method would entail obtaining a third value by using a step size half of that used to determine the second value. Indeed, one strategy for achieving a desired accuracy with reasonable certainty is to solve the ODE repeatedly by a particular method, halving the step size each time, and continuing until the new value received differs from its predecessor by less than the desired accuracy (though we must be careful not to push this approach so far that roundoff problems within the computer begin to become significant). We will illustrate this approach in the context of a specific example as soon as we are ready to use the computer to do the arithmetic.

From a more sophisticated perspective, numerical analysts have deduced expressions for the error in various approaches to solving ODEs numerically. To assess the error in Euler’s method, for example, we begin by noting the Taylor theorem with remainder, which asserts that

$$x(t + \Delta t) = x(t) + \frac{dx}{dt}(t) \Delta t + \frac{1}{2} \frac{d^2x}{dt^2}(\xi) \Delta t^2$$

(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

$$x_{\text{exact}}(t + \Delta t) = x_{\text{Euler}}(t + \Delta t) + \frac{1}{2} \frac{d^2x}{dt^2}(\xi) \Delta t^2$$

(11.93)

or that

$$\left| x_{\text{exact}}(t + \Delta t) - x_{\text{Euler}}(t + \Delta t) \right| \leq \text{Max} \left( \frac{1}{2} \frac{d^2x}{dt^2}(\xi) \right) \Delta t^2$$

(11.94)

which provides yet another motivation for casting problems in dimensionless form.

Warning: This approach is not entirely secure if the procedure converges slowly as the time step is reduced. Remember that $\sum_{n=1}^{\infty} 1/n$ diverges even though the effect of the millionth term added to the partial sum starts in the sixth decimal place.

---

30 This assertion is refined in the next paragraphs.
31 Warning: This approach is not entirely secure if the procedure converges slowly as the time step is reduced. Remember that $\sum_{n=1}^{\infty} 1/n$ diverges even though the effect of the millionth term added to the partial sum starts in the sixth decimal place.
Thus, we learn that, with Euler’s method, the (truncation) error per step varies as the square of the step size; halving the step size will reduce the error per step by a factor of four.

Statements similar to Eq. (11.93) can be deduced for all of the methods that we have described in this section. Without proof,\(^{33}\) 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)

\[
x_{\text{RK2}}(t + \Delta t) + O(\Delta t^3)
\]

(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 the various estimates of \(x(t + \Delta t)\) starts with the exact solution at time \(t\). Except for the first step, which moves away from (exact) initial conditions, that assumption is invalid. Truncation errors per step compound as more and more steps are taken, and a full assessment of the error in a solution obtained by one or another numerical means must recognize this cumulation. Beyond the error per step, we must be aware of the global truncation error, which attempts to estimate how much error accumulates in the course of working out a solution over the entire desired range of the independent variable. The task of assessing global error is extremely difficult. Crudely, however, if we simply add up expressions like those in the previous paragraph for the \(N\) steps in an entire solution and suppose—without much justification—that each step contributes about the same amount, we would conclude that the global truncation error would be order \(N O(\Delta t^p)\) when the error per step is of order \(O(\Delta t^p)\). For a fixed interval, \(N\) is itself of order \(1/\Delta t\). Thus, we infer that the global truncation error is of order \(O(\Delta t^{p-1})\). This result does not help us much in determining the global error. It does, however, support the conclusion that, for the four methods in the previous paragraph, halving the step size will reduce the global error by a factor of two, four, four, and sixteen, respectively. In particular, halving the step size with the fourth-order Runge-Kutta method will add at least one more decimal digit to the accuracy of the solution overall (and may well do much better).

We have already mentioned that conserved quantities like energy, linear momentum, and angular momentum can sometimes also be used as a check on the accuracy of an evolving solution. Whenever such a conserved quantity exists, its initial value must, of course, be preserved (within some limits). Failure of a particular solution to conform to that requirement signals a need for a smaller time step or a more sophisticated algorithm. Note, however, that some algorithms (e.g., the Euler algorithm for the three-species radioactive decay) automatically preserve one or more conserved quantities; preservation in such cases does not provide any information about the accuracy of the solution itself.

### 11.6.5 Adaptive Methods

In the previous subsections, we assumed that the user of a particular algorithm would actually view the values obtained for different time steps and decide personally when to stop by examining the changes that occur as the time step is successively halved. One can, of course, program a computer to make those decisions. In essence, the program generates the solution with one time step, then

---

repeatedly generates it with a succession of ever smaller time steps, comparing each new solution with its immediate predecessor and stopping when the absolute value of the difference is smaller than a tolerance—either absolute or relative—prescribed in advance. As a guard against an infinite loop, these algorithms should also stop if the desired tolerance has not been achieved in some maximum number of refinements and should print a warning when the desired tolerance has not in its judgment been achieved.

Another family of algorithms aims to minimize computational labor by estimating—though the methods for doing so are often themselves approximate—the accuracy obtained at each step along the way to a solution.\(^{34}\) A trial row in the table of data is generated, and the error is assessed. If the error is within a user-specified tolerance, the program moves on to the next row in the data table. If the error exceeds the desired tolerance, the program repeats the calculation with progressively smaller time steps until the desired tolerance is achieved and only then is a new row added to the table. The procedure also contains means by which the time step is increased when the estimated error is less than the specified tolerance. Thus, the time step fluctuates as the solution unfolds, being small when the solution is changing rapidly and large when the solution is changing slowly. Because of this feature, the procedure is said to be adaptive. Without the overlay of an elaborate interpolation, adaptive methods have the disadvantage of generating solutions at irregularly spaced times. That disadvantage, however, is frequently outweighed by the substantial advantage of concentrating the computational effort in regions where the solution changes rapidly.\(^{35}\)

### 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 \textit{single} earlier point, though many have interpolated solutions at several points \textit{between} the initially known point and the desired end point; such methods are called \textit{single-step methods}. Some of the solvers available in some software packages use \textit{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 \(f_i = f(x_i, t_i)\). Then, supposing we already had in hand estimates of the solution at four (say) earlier points, we might estimate the solution at \(t_{i+1}\) by the formula

\[
x_{i+1} = x_i + \frac{\Delta t}{24} \left( 55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3} \right)
\]

(11.99)

(which represents a particular fourth-order member of a family of methods known as Adams-Bashforth methods).\(^{36}\) 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 \textit{not} self starting. We must adopt some other method to obtain from the initial values however many values are needed to support the multistep method. To start a solution based on the stepping formula of Eq. (11.99), we might use a fourth-order Runge-Kutta method to obtain \(x_1, x_2,\) and \(x_3\) from \(x_0\). Then we could shift to the fourth-order Adams-Bashforth formula for all subsequent points.

\(^{34}\text{Sometimes the assessment of accuracy will be based not on the generated solution itself but on some property, e.g., the energy, computed from that solution.}\)

\(^{35}\text{In contrast, Euler's method, for example, would require using throughout a time step small enough to generate an accurate solution in regions where the solution varies most rapidly. Consequently, too much computational effort would be focused in regions where the solution changes slowly enough that a larger time step would be acceptable.}\)

\(^{36}\text{See \textit{Elementary Numerical Analysis}, S. D. Conte (McGraw-Hill, New York, 1965), Section 6.6.}\)
11.7. SOLVING ODES NUMERICALLY WITH IDL

We could, of course, simply take the result given by Eq. (11.99) as the solution at \( t_{i+1} \). For a more sophisticated solution, we could regard that result as a prediction and use it in a paired formula to “correct” the prediction. An appropriate fourth-order corrector formula is a member of the Adams-Moulton family, namely

\[
x_{i+1} = x_i + \frac{\Delta t}{24} \left( 9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2} \right)
\]

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.

11.7 Solving ODEs Numerically with IDL

Note: All IDL program (.pro) files referred to in this chapter are available in the directory $HEAD/idl, 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 IDL’s default search path. If so, the files can be found by IDL 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 IDL’s elementary commands or, more simply, by invoking one of the built-in routines, which include IDL-supplied routines rk4 and lsode and Lawrence adapted routines ludiffeq_23 and ludiffeq_45.

11.7.1 Using Elementary Commands

To carry the process worked out by hand in Section 11.6.1 much beyond the first few steps is a job for a computer. Suppose, for example, we wanted to determine the Euler approximation to the solution of Eq. (11.19) for chain radioactive decay over the time interval \( 0.0 \leq t \leq 50.0 \) with a time step of \( \Delta t = 0.25 \) and \( k_A = k_B = 0.1 \). We could exploit the capabilities of IDL as follows. First, we set up arrays for \( t \), \( A \), \( B \), and \( C \), where 201 elements are needed (time interval divided by time step = 50.0/0.25 = 200 steps, but we need an element also for the initial value). Possible statements are

\[
\text{IDL> } t = \text{fltarr}(201) \quad \text{Create vector for } t.
\]

\[
\text{IDL> } A = \text{fltarr}(201) \quad \text{Create vectors for } A(t), B(t), C(t).
\]

\[
\text{IDL> } B = \text{fltarr}(201) \quad \text{IDL> } C = \text{fltarr}(201)
\]

Since \( t \), \( B \), and \( C \) start with the value 0.0 (which has already been established in the corresponding vectors), we need set the initial value explicitly only for \( A \). We also set values for the two parameters and for the time step. These values are established with the commands

\[
\text{IDL> } A[0] = 1000.0 \quad \text{Set } A(0) \text{ to initial value.}
\]

\[
\text{IDL> } kA = 0.1 \& kB = 0.1 \quad \text{Set parameters.}
\]

\[
\text{IDL> } dt = 0.25 \quad \text{Set time step.}
\]

The solution is then calculated, time instant by time instant, with the loop\(^{37}\)

\(^{37}\)As described at the end of Section 2.8, when we work interactively, the commands in an IDL for loop must be assembled so that IDL sees the entire loop as a single command. Thus, the ampersands and dollar signs are critical. The same constraint applies when a for loop is used in a command file executed with the $ command. When a for loop is used in a procedure or function file, however, the loop is delimited by the begin and endfor commands and no ampersands or dollar signs are necessary.
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>544.641</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>

IDL> for i = 1,200 do begin $$
    IDL> B[i] = B[i-1] + kA*A[i-1]*dt - kB*B[i-1]*dt & $$
    IDL> C[i] = C[i-1] + kB*B[i-1]*dt & $$
    IDL> t[i] = t[i-1] + dt & $$
    IDL> endfor

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 $t$.\footnote{\textsuperscript{38}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 $t[i] = i \cdot dt$, though doing so would increase the execution time. (Multiplication takes more time than addition.)} The simple statement

IDL> for i = 0,3 do print, [ t[i], A[i], B[i], C[i], [ A[i]+B[i]+C[i] ] ]

<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>0.0000</td>
<td>1000.00</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1000.00</td>
</tr>
<tr>
<td>0.2500</td>
<td>975.000</td>
<td>25.000</td>
<td>0.0000</td>
<td>1000.00</td>
</tr>
<tr>
<td>0.5000</td>
<td>950.625</td>
<td>48.7500</td>
<td>0.62500</td>
<td>1000.00</td>
</tr>
<tr>
<td>0.7500</td>
<td>926.859</td>
<td>71.2969</td>
<td>1.84375</td>
<td>1000.00</td>
</tr>
</tbody>
</table>

displays the resulting solution and the conserved quantity $A + B + C$ for the first few time steps. Reassuringly, the values here agree with those in Table 11.2. The more interpretable graphical output of Fig. 11.10 is produced by the statements

IDL> plot, t, A, thick=3.0, title='Three-Species Decay Chain', $$
IDL> xtitle='Time', ytitle='Number of Atoms Present', charsize=1.3
IDL> oplot, t, B, thick=3.0, linestyle=1
IDL> oplot, t, C, thick=3.0, linestyle=2

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 $\pm 1.0$ in these values. For some purposes, of course, that accuracy may well be adequate. In particular, to the resolution of the graph in Fig. 11.10, for example, a variation of $\pm 1.0$ in a particular vertical coordinate would hardly show. At the same time, we must remember that these differences may cumulate and may well be larger at $t = 50.0$, for example, than they are at $t = 6.0$. To assess that possibility, we
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 11.7: 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>

generate the entire solution via Euler’s method for three time steps ($\Delta t = 2.0$, 0.5, and 0.0625), plotting those solutions in Fig. 11.11. The comparison of the resulting graphs is described in the caption to the figure.

A similar demonstration using the improved Euler method is left to an exercise. In particular, the values presented in Table 11.7 reflect the result of that exercise when the procedures leading to Table 11.6 are repeated with the improved Euler method. Clearly, the results in Table 11.7 are converging more quickly on stable values to more decimal places than we observed in Table 11.6. 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.
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.

![Three-Species Decay Chain](image)

Table 11.8: A generic function to define ODEs for IDL.

```idl
FUNCTION FunctionName, IndVar, DepVar
;+ ;      Explanatory comments.
;+ ;- common params, NameOfParam1, NameOfParam2, ...
;: Intermediates statements.
;: RETURN, [DerivOfFirstVar, DerivOfSecondVar,...]
END
```

11.7.2 Defining ODEs for IDL

Before using any of the routines IDL supplies for solving a system of first-order differential equations numerically, we must create a function pro-file whose execution by IDL 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 pro-file, which—in broad outline—will have the general form shown in Table 11.8. In this structure,

39See Section 2.8.
11.7 SOLVING ODES NUMERICALLY WITH IDL

- The function name must 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 must be named decay.pro.

- The variables must be given in the order independent variable, dependent variable.

- If the differential equations of interest have no internal parameters, then the common area named params should be omitted from the file defining the function.\(^{40}\)

Explicit pro-files for several examples will be presented in the subsequent subsections.

11.7.3 The Procedures rk4, lsode, ludiffeq\.23, and ludiffeq\.45

IDL contains two built-in procedures for solving ODEs. The procedure\(^{41}\) rk4 is adaptive and uses a fourth-order Runge-Kutta method to step the solution forward by a specified amount, holding the solution to an absolute (truncation) error per step no larger in absolute value than \(10^{-6}\). A statement invoking rk4 has the general form

\[
\text{soln = rk4(} \text{VectorOfDepVars, VectorOfDerivs, IndVar, TimeStep, 'FunctionName')} \\text{)}
\]

where the first two arguments are vectors containing the solution and the derivatives of the solution at the value of the independent variable provided in the third argument, the fourth argument gives the increment by which the independent variable is to be advanced, and the final argument—placed in (single or double) quotation marks—names the function pro-file defining the differential equations. An optional keyword /double will request rk4 to carry out its calculations in double precision.

The second built-in procedure for solving ODEs is lsode, an adaptive procedure that, as with rk4, advances the solution by one time step of specified size.\(^{42}\) A statement invoking lsode has the general form

\[
\text{soln = lsode(} \text{VectorOfDepVars, IndVar, TimeStep, 'FunctionName', Status, atol = AbsTol, rtol = RelTol)} \\text{)}
\]

where the first argument is a vector containing the solution at the value of the independent variable provided in the second argument, the third argument gives the increment by which the independent variable is to be advanced, and the fourth argument specifies the function pro-file defining the differential equations. The argument Status is an integer, which must be set to the value 1 on the first call to lsode, though the procedure is written to assume that value by default (so the argument can be regarded as optional). Thereafter, this quantity is manipulated by lsode itself and provides a means by which (a) one invocation of the routine can communicate its status to the next invocation and (b) errors encountered in the calculation can be communicated to the user.\(^{43}\) Finally, the two keywords give the user control over the tolerance to which each step in the calculation will be held.\(^{44}\) Specifically, atol (default \(10^{-7}\)) specifies the absolute value of the maximum difference between successive estimates of the solution at the end of the step,\(^{45}\) and rtol (default \(10^{-7}\)) specifies

---

\(^{40}\)Use of named common in IDL is discussed more fully in Sections 2.8 and 2.17.

\(^{41}\)The IDL procedure rk4 is based on the procedure of the same name in the Numerical Recipes library and is used in IDL by permission of Numerical Recipes. It is described in detail in William H. Press et al., Numerical Recipes: The Art of Scientific Computing—Second Edition (Cambridge University Press, 1992). See Appendix Z for full contact information.

\(^{42}\)The IDL procedure lsode is based on the public-domain FORTRAN program LSODE written by Alan C. Hindmarsh of the Computing and Mathematics Research Division of the Lawrence Livermore National Laboratory. See Appendix Z for full contact information.

\(^{43}\)The interpretation of the possible values of this argument is fully explained in the IDL manuals.

\(^{44}\)Note that these tolerances specify error per step; the global error may well be larger than the error per step. Beware!

\(^{45}\)Actually, atol can be a vector if a different absolute tolerance is to be specified for each dependent variable.
the absolute value of the maximum relative difference between successive estimates. Whatever the
values, the error per step in the \( i \)-th dependent variable will be kept smaller than \( rtol \cdot \text{abs}(y[i]) + atol \). Either \( atol \) or \( rtol \) can be set to zero if assessment of the error per step is to be done
purely by one means or the other.

Since \texttt{rk4} and \texttt{lsode} advance the solution by one step of the specified size, generation of an entire
solution from the initial conditions requires us to embed calls to these procedures in an explicitly
constructed loop. More specific construction of sample loops will be illustrated in the examples to
come.

In 1991, before the built-in routines \texttt{rk4} and \texttt{lsode} became part of IDL, the author, with
help from the vendors of IDL, wrote two adaptive ODE routines to solve systems of first-order
differential equations for use at Lawrence: \texttt{ludiffeq.23}, which uses a combination of second and
third order Runge-Kutta algorithms, and \texttt{ludiffeq.45}, which implements a combination of fourth
and fifth order Runge-Kutta algorithms. The algorithms used in these routines are patterned after
the algorithms then used in the MATLAB routines \texttt{ode23} and \texttt{ode45}, respectively, and are used in
and distributed with this book both by permission of The MathWorks, Inc.\textsuperscript{46}, which owns the
copyright on the original algorithms, and by permission of Research Systems, Inc.,\textsuperscript{47} whose employees
assisted in the coding of the algorithms in IDL’s language. Both \texttt{ludiffeq.23} and \texttt{ludiffeq.45}
impose a limit on the minimum size of the time step. If the solution does not converge to the
specified tolerance before the minimum time step is reached, the program will print the message,
“Beware of Singularity”, though the keyword \texttt{maxsteps} described later gives the user some control
over the appearance of this error message.

The two routines \texttt{ludiffeq.23} and \texttt{ludiffeq.45} are implemented as procedures.\textsuperscript{48} Each is
invoked with the same arguments. Specifically, the form of a call to \texttt{ludiffeq.23} is\textsuperscript{49}

\begin{verbatim}
ludiffeq.23, 'FunctionName', IndVar, DepVars, t0 = StartTime, tf = EndTime, $
init = InitConds, tol = RelativeErrorTolerance, maxsteps = MaxStepsInSol
\end{verbatim}

In this statement,

- The statement must begin with the command name (\texttt{ludiffeq.23}), the function name in
  quotation marks (either single or double), the name of a vector to store the independent
  variable, and the name of an array to store the dependent variables \textit{in this order}.
- Each of the remaining items is identified by a keyword. Thus, their order is irrelevant.
- Failure to specify a value for the tolerance results in IDL selecting default values of \( 10^{-3} \) for
  \texttt{ludiffeq.23} and \( 10^{-6} \) for \texttt{ludiffeq.45}.
- When the keyword \texttt{maxsteps} is given a value, the procedure will calculate the quantity
  \( (tf - t0)/\text{maxsteps} \) and interpret the result as a lower limit on the step size to which the built-
  in adaptive refinement will go as the procedure seeks to achieve the specified (or default)
tolerance.\textsuperscript{50} If this minimum stepsize is reached but the internal tests indicate that the desired

\textsuperscript{46}The MathWorks, Inc., is the vendor of MATLAB. See Appendix Z for full contact information.
\textsuperscript{47}IDL became commercially available in 1977, at which time the vendor was RSI (Research Systems, Inc.). In 2004,
RSI became a wholly owned subsidiary of ITT Industries and was named ITT Visual Information Solutions. In 2011,
that company changed its name to Exelis Visual Information Solutions. In 2015, Exelis was purchased by the Harris
Corporation and became Harris Geospatial. See Appendix Z for full contact information.
\textsuperscript{48}These procedures are implemented in the files \texttt{ludiffeq.23.pro} and \texttt{ludiffeq.45.pro}. These files may be placed
in IDL’s search path and will be found when invoked without specifying a path. They are also available in the
directory \$HEAD/idl and can be invoked with that path or copied from there to the default directory, in which they
will be found when invoked without specifying a path.
\textsuperscript{49}The additional (hidden) keyword \texttt{Params}, which is supplied for backward compatibility to an earlier version, is
now obsolete and is not here described. The function of this keyword is more easily accomplished through the use of
common storage, as described in Section 11.7.2.
\textsuperscript{50}The largest allowed step size is taken to be \( (tf-t0)/8 \), but no error results if the procedure leads to a larger step
size. The larger step size is simply replaced by the maximum allowed size.
11.7. SOLVING ODES NUMERICALLY WITH IDL

Table 11.9: The IDL function decay.pro.

```idl
FUNCTION decay, t, n
   ;+ ; The function DECAY defines the rate equations for a three 
   ; species radioactive decay sequence such as 
   ; n[0] --kA--> n[1] --kB--> n[2] 
   ; The parameters kA and kB must be passed to the routine in 
   ; that order in a common area named params. 
   ;- common params, kA, kB
RETURN, [ -kA*n[0], kA*n[0]-kB*n[1], kB*n[1] ]
END
```
tolerance has not been achieved, the procedure will print the message “Beware of Singularity” and terminate execution. Note two features: (1) Failure to specify a value for maxsteps results in IDL selecting the default value 20000. (2) The specified value of maxsteps should not be less than 100, since the program begins by supposing a trial step size of (tf - t0)/100, and a trial step size smaller than the smallest allowable step size will result in immediate termination of execution.

With this information as input, the user invokes ludiffeq.23 or ludiffeq.45 to generate the solution and store it in the specified pair of arrays in the IDL workspace. One of these arrays is a vector containing the values of the independent variable, usually time (e.g., the first column in Table 11.2). The second array is a two-dimensional matrix that contains values of the dependent variables (e.g., the second, third, and fourth columns of Table 11.2).

11.7.4 Chain Radioactive Decay

As a first example of the use of an IDL routine, consider the equations describing chain radioactive decay, Eq. (11.19). Before we can construct the proper pro-file to communicate the differential equations to IDL’s routine, we must establish a correspondence between the elements of a row vector—name it n—and the dependent variables. Since we have three dependent variables (the number of nuclei of species A, B, and C), n will have three elements, which IDL refers to as n[0], n[1], and n[2]. If we assign the number of nuclei of species A to the first element, the number of nuclei of species B to the second element, and the number of nuclei of species C to the third element, we have the associations A \rightarrow n[0], B \rightarrow n[1], and C \rightarrow n[2]. Rewriting the differential equations Eq. (11.19) in terms of the elements of n, we have the system of equations

\[
\frac{d}{dt} n[0] = -k_A n[0] \quad ; \quad \frac{d}{dt} n[1] = k_A n[0] - k_B n[1] \quad ; \quad \frac{d}{dt} n[2] = k_B n[1]
\]  

(11.101)

Now we are ready to construct the function pro-file describing this system of differential equations. Respecting the general structure described in Section 11.7.2, we would compose the definition in Table 11.9. Here, the first line tells IDL that the function is named decay and declares t—a scalar—to be the independent variable and n—a vector—to be the dependent variables. The common area named params provides the means to communicate the two parameters k_A and k_B from IDL’s main level. The next to last line, which begins with the mandatory word RETURN, specifies
a vector whose elements are the derivatives of the dependent variables as determined by the right hand sides of Eqs. (11.101). Finally, the last line is a mandatory statement ending the function.

With the function of the previous paragraph stored in the file decay.pro in the default directory, we would produce a solution for the differential equations it defines with the statements:

```
IDL> common params, kA, kB  
IDL> kA = 0.1 & kB = 0.1  
IDL> ic = [1000.0, 0.0, 0.0 ]  
IDL> ludiffeq_23,'decay', t, n,  
    t0=0.0, tf=50.0, init = ic  
```

Then we would plot the three dependent variables on the same axes with the statements:

```
IDL> plot, t, n[0,*], title='Three-Species Chain Decay', \  
    xtitle='Time', ytitle='Number of Atoms Present', \  
    thick=3.0, charsize=1.3  
IDL> oplot, t, n[1,*], linestyle=1, thick=3.0  
IDL> oplot, t, n[2,*], linestyle=2, thick=3.0  
```

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. 11.12 with the graph in Fig. 11.10. 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 ludiffeq.23, with the (default) tolerance of 0.001, and was accomplished with 26 points, as determined by executing the IDL command

```
IDL> print, size(t)  
```

```
1 26 4 26  
```

after the commands most recently described have been executed. To the resolution of the graphs, the two solutions appear to be pretty much the same. Evidently, the choice of the time step in our application of Euler’s method was adequate to yield a solution accurate to the resolution of the graph, which is the most restrictive statement the evidence we have quoted supports. We might alternatively have inferred the likely appropriateness of the time step \( \Delta t = 0.25 \) in Euler’s method by noting from the graphs that the solution varies appreciably over a time period on the order of a few time units. For example, the initial value of \( A \) drops by 20% from 1000 to 800 in about 2 or 3 time units. Since the time step \( \Delta t \) is small compared to that characteristic time, we would be justified in anticipating a reasonably accurate solution.

Let us also generate a solution to this same problem using rk4. The function pro-file decay.pro will serve in this capacity as well, but we must decide \textit{a priori} the instants of time at which we seek

---

\footnote{In the illustrated use of IDL, we did not have to compile the pro-file defining the system of differential equations. If, however, our first attempt to invoke any ODE solver generates compilation errors arising from problems with decay.pro, we would need to correct those errors in the source file. Then, short of exiting from IDL and reentering, we \textit{would} need to recompile decay.pro explicitly. As described more fully in Section 2.8, the IDL statement to accomplish this recompilation is .compile decay. Otherwise, a new invocation of the ODE solver will reuse the already-compiled, erroneous version of the pro-file.}

\footnote{Note that we can have more parameters in the \texttt{common} statement at IDL’s command level than are actually used in the routines called. Here we have set just the two needed by decay.pro. If we anticipated using the common area params in other contexts in the same IDL session, we might find it convenient to include some dummy parameters at the creation of that common area. The only way to remove or redefine an already defined common area is to reset the IDL session (see Section 2.17) or exit from and reenter IDL, neither of which may be desirable in some contexts. No harm is done when a common area contains unused parameters, and the “extra” ones need not be mentioned in routines that don’t need them.}
11.7. SOLVING ODES NUMERICALLY WITH IDL

Figure 11.12: Solution of chain radioactive decay via `ludiffeq` with a tolerance of 0.001. The solid, dotted and dashed lines show $A(t)$, $B(t)$, and $C(t)$, respectively.

A solution and build `rk4` into a loop that is executed for each of those times. Suppose we seek the solution over the time interval $0 \leq t \leq 50$ again, dividing that interval into 200 steps of size 0.25. We would invoke the following statements:

```
IDL> t = findgen(201)/4.0  Set times.
IDL> n = fltarr( 3, 201 ) Prepare three-column array for solution.
IDL> n[0,0] = 1000.0 Set initial value.
IDL> dt = 0.25 Set time step.
IDL> common params, kA, kB Set parameters.
IDL> kA = 0.1 & kB = 0.1
IDL> for i = 1, 200 do begin $ Loop once per point:
    IDL> der = decay(t[i-1], n[*],i-1 ) & $ Evaluate derivative at start.
    IDL> n[*],i = rk4( n[*],i-1, der, $ Calculate solution $dt$ later.
    IDL> t[i-1], dt, 'decay' ) & $
IDL> endfor
```

Here, the first statement in the loop calculates the derivatives at the current point for use in the second statement, which invokes `rk4` to step the solution from time $t_{i-1}$ to time $t_i$. 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.

Finally, let us use `lsode` to solve this problem. In essence, we execute the same statements as in the previous paragraph, replacing the loop with the statements

```
```
FUNCTION damposc, t, y
    ;+  
    ; 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. This one 
    ; parameter must be passed to the routine in a common area named params. 
    ;-
    common params, bb
    RETURN, [ y[1], -y[0] - bb*y[1] ]
END

IDL> for i = 1, 200 do begin $
IDL>  tmp = n[*,i-1] &
IDL>  n[*,i] = lsode( tmp, t[i-1], dt, 'decay' ) &
IDL>  endfor

Here, we do not need an anticipatory evaluation of the derivatives (as was required by \texttt{rk4}), but we do 
need to extract the current values of the dependent variables from the growing array (variable \texttt{tmp}) 
because \texttt{lsode} will not accept the entity \texttt{n[*,i-1]} as its first argument. Plotting statements identical 
to those that produced Fig. 11.12 will again produce a graph essentially identical to Fig. 11.12.

### 11.7.5 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 function pro-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 \( \pi \rightarrow y[0] \) and \( \sigma \rightarrow y[1] \). In these terms, the differential equations become

\[
\frac{d}{dt} y[0] = y[1] ; \quad \frac{d}{dt} y[1] = -y[0] - \beta y[1] 
\]

Then, using the symbol \( \beta \) for \( \beta \), we create the pro-file recorded in Table 11.10 defining the differential 
equations. After storing this file in the default directory with the name \texttt{damposc.pro}, we explore 
the system for two different damping constants with the statements

| IDL> ic=[1.0,0.0]  | Set initial conditions. |
| IDL> common params, bb   | Define common for parameters. |
| IDL> bb=0.0          | Set damping to zero. |
| IDL> ludiffeq_23, 'damposc', t1, x1,$ | Solve equations, putting times in \texttt{t1} and solution in \texttt{x1}. |
| IDL> init=ic, t0=0.0, tf=25.0, $  |   |
| IDL> tol=0.0001     |   |
| IDL> bb=0.25        | Set small damping. |
| IDL> ludiffeq_23, 'damposc', t2, x2,$ | Solve equations with same ic's as before but putting times in \texttt{t2} and solution in \texttt{x2}. |
| IDL> init=ic, t0=0.0, tf=25.0,$ |   |
| IDL> tol=0.0001     |   |
Finally, we produce graphs showing the position and velocity as functions of time and also the trajectory in the phase plane ($v$ versus $x$) for each value of the damping constant with the statements

IDL> !p.multi = [ 0, 2, 2 ]

IDL> plot, t1, x1[0,*], thick=2.0, title='Undamped Harmonic Oscillator', $ \$ 
IDL> subtitle='X and V as Functions of T', xtitle='Time'

IDL> oplot, t1, x1[1,*], linestyle=2, thick=2.0

IDL> plot, t2, x2[0,*], thick = 2.0, title='Damped Harmonic Oscillator', $ \$ 
IDL> subtitle='X and V as Functions of T', xtitle='Time'

IDL> oplot, t2, x2[1,*], linestyle=2, thick=2.0

IDL> plot, x1[0,*], x1[1,*], xrange=[-1,1], yrange=[-1,1], $ \$ 
IDL> subtitle='Phase Plane', thick=2.0, $ \$ 
IDL> xtitle='Position', ytitle='Velocity'

IDL> plot, x2[0,*], x2[1,*], xrange=[-1,1], yrange=[-1,1], $ \$ 
IDL> subtitle='Phase Plane', thick=2.0, $ \$ 
IDL> xtitle='Position', ytitle='Velocity'

The resulting graphs are shown in Fig. 11.13.

Alternatively, we could, of course, approach this problem using `rk4` or `lsode`. Deciding to seek a solution over the interval $0 \leq t \leq 25$ and dividing the interval into 100 segments, we set up the variables with the statements

IDL> t = findgen(101)/4.0    \ Set times.

IDL> soln = fltarr(2,101) \ Prepare two-column array for solution.

IDL> soln[0,0] = 1.0     \ Set initial value.

IDL> dt = 0.25          \ Set time step.

IDL> common params, bb \ Set parameter for no damping.

IDL> bb = 0.0

Then, we pursue the solution either with the loop

IDL> for i = 1, 100 do begin $ \$
IDL>   der = damposc( t[i-1], soln(*,i-1) ) & $ \$
IDL>   soln[*,i] = rk4( soln[*,i-1], der, t[i-1], dt, 'damposc' ) & $ \$
IDL> endfor

or with the loop

IDL> for i = 1, 100 do begin $ \$
IDL>   tmp = soln[*,i-1] & $ \$
IDL>   soln[*,i] = lsode( tmp, t[i-1], dt, 'damposc' ) & $ \$
IDL> endfor

Repeating these steps with $bb$ set to 0.25 solves the problem for a small damping. Graphs of all of these results would look just like the graphs in Fig. 11.13. Evidence of the accuracy of these solutions can be found in the preservation of the amplitude of the motion for the undamped case, in the fact that the phase-plane diagram for the undamped case traces the same path four times over without evident departure, and in the specified (default) tolerance with which `ludiffeq_23` pursues its solution. Further evidence could be garnered by comparing these solutions with those we might obtain with smaller or larger time steps.
Figure 11.13: Damped and undamped harmonic oscillators. Notice the difference between the phase plane orbits.

11.7.6 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 pair of second-order equations as a quartet of first-order equations. Thus, we view the equations as the system

$$
\frac{d\varpi}{dt} = v_x; \quad \frac{dv_x}{dt} = -\frac{\varpi}{(x^2 + \varpi^2)^b}; \quad \frac{dv_y}{dt} = v_y; \quad \frac{dv_y}{dt} = -\frac{\varpi}{(x^2 + \varpi^2)^b} \quad (11.103)
$$

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 pro-file, we set $\varpi, v_x, \varpi, v_y$ into correspondence with the elements $x[0], x[1], x[2]$, and $x[3]$ of the four-element vector $x$. Then, one possibility for the pro-file planet.pro 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 IDL and pursue solutions to the planetary problem. To achieve adequate accuracy, we here need the better algorithm and higher tolerance of
11.7. SOLVING ODES NUMERICALLY WITH IDL

Table 11.11: The IDL Ffunction planet.pro.

FUNCTION planet, t, x

; The function PLANET describes the dimensionless 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 array of dependent variables are X-position, X-velocity,
; Y-position, Y-velocity. The one parameter must be passed to
; the routine in a common area named params.

; common params, b

temp=( x[0]^2 + x[2]^2 )^b
return, [ x[1], -x[0]/temp, x[3], -x[2]/temp ]
end

ludiffeq_45. 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

IDL> common params, b
IDL> b=1.5
IDL> ic1=[ 4.0, 0.0, 0.0, 0.5 ]
IDL> ludiffeq_45, 'planet', t1, x1, init=ic1, t0=0.0, tf=60.0
IDL> ic2=[ 4.0, 0.0, 0.0, 0.3 ]
IDL> ludiffeq_45, 'planet', t2, x2, init=ic2, t0=0.0, tf=60.0

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. 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

IDL> E1 = 0.5*(x1[1,*]^2+x1[3,*]^2) - 1.0/sqrt(x1[0,*]^2+x1[2,*]^2)
IDL> L1 = x1[0,*]*x1[3,*] - x1[2,*]*x1[1,*]
IDL> E2 = 0.5*(x2[1,*]^2+x2[3,*]^2) - 1.0/sqrt(x2[0,*]^2+x2[2,*]^2)
IDL> L2 = x2[0,*]*x2[3,*] - x2[2,*]*x2[1,*]

will calculate these quantities. Then, statements like

IDL> print, [ E1, L1 ]

will display the resulting values. We find—see your own output—that E1 = −0.125000 at the
beginning and climbs to −0.125001 by the end of the interval while L1 starts at 2.00000 and falls
to 1.99999; E2 varies between −0.205001 and −0.204994 while L2 remains rock solid at 1.20000
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.

53The statements creating the graphs have been omitted but, except for the wrinkle of making circles appear as
circles (see Section 2.16.5), could be constructed with little difficulty.
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.
Open orbits, in which a satellite with positive energy moves in the gravitational field of a sun, are also possible. We utilize IDL’s for loop to produce a variety of open orbits for a planet started in different initial positions by executing the statements

IDL> b=1.5
IDL> plot, [-4, 4], [-4, 4], /nodata,
IDL> title='Open Orbits', xtitle='X-Position', ytitle='Y-Position',
IDL> charsize=1.3
IDL> for i=-4,4 do begin ic=[-4.0, 1.0, float(i), 0.0] &
IDL> ludiffeq_45, 'planet', t, x, t0=0.0, tf=30.0, init=ic &
IDL> oplot, x[0,*], x[2,*], thick=3.0, linestyle=abs(i) &
IDL> endfor

Figure 11.15 shows the resulting output.

11.7.7 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
FUNCTION stdwaves, x, f
    ;+ The function STDWAVES defines the basic equations for describing
    ; standing waves in a string. The one parameter k in the equations
    ; must be passed to the routine in a common area named params.
    ;- common params, k
RETURN, [ f[1], -k^2*f[0] ]
END

system of first-order equations is
\[
\frac{df}{dx} = g ; \quad \frac{dg}{dx} = -k^2 f ; \quad f(0) = 0 ; \quad f(\ell) = 0
\] (11.104)

Thus, setting up the correspondences \( f \mapsto f[0] \) and \( g \mapsto f[1] \), we might define this system for
IDL with the pro-file shown in Table 11.12 which we store in the default directory with the name stdwaves.pro.

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 \( \frac{df(0)}{dx} = g(0) \) 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 \texttt{ludiffeq}_23, we would invoke the statements

\[
\begin{align*}
\text{IDL} & > \text{common params, } k & \text{Set parameter.} \\
\text{IDL} & > k = 1.0 & \text{Set initial conditions.} \\
\text{IDL} & > \text{ic} = \text{[0.0, 1.0]} & \text{Invoke solver.} \\
\text{IDL} & > \text{ludiffeq}_23, \ '\text{stdwaves}', \ t, \ f, \ t0=0.0, \ tf=1.0, \ init=\text{ic} & \text{Plot solution.} \\
\text{IDL} & > \text{plot}, \ t, \ f[0,*] & \text{The result is the solid line in Fig. 11.16. Repeating this calculation while overriding the default} \\
\text{IDL} & > \text{tolerance (0.001) by stipulating a tolerance of 0.0001 for } \text{ludiffeq}_23 \text{ (do it!), we find no significant} \\
\text{IDL} & > \text{difference (to the resolution of the graph), nor do we find significant differences when the two} \\
\text{IDL} & > \text{solutions are repeated with } k = 20.0 \text{ (again, do it!), so we accept the default tolerance for subsequent} \\
\text{IDL} & > \text{calculations.} \\
\text{Clearly, however, the solution we have obtained (and in which we now have reasonable} \\
\text{confidence) fails utterly to return to the value zero at the upper end of the interval. Either the assumed} \\
\text{derivative } [g(0) = 1.0] \text{ or the assumed value } k = 1.0 \text{ is not appropriate. Because the equations are} \\
\text{linear, however, changing } g(0) \text{ will merely influence the scale of the resulting solution; a different} \\
\text{value of the derivative at the beginning will never cause the non-zero value at the end to become} \\
\text{zero. We quickly conclude that efforts to bring } f(1) \text{ to the value zero stand a chance of succeeding} \\
\text{only if we tamper with } k. \text{ Thus, we explore other values of } k \text{ with the statements}
\end{align*}
\]

\[
\begin{align*}
\text{IDL} & > \!y.\text{range}=[-1.0,1.0] & \text{IDL} \text{> } k = 1.0 \text{ & } \text{ludiffeq}_23, \ '\text{stdwaves}', \ t, \ f, \ t0=0.0, \ tf=1.0, \ init=\text{ic} \\
\text{IDL} & > \text{plot}, \ t, \ f[0,*], \ \text{thick=3.0, ticklen=1.0} & \text{IDL} \text{> oplot}, \ t, \ f[0,*], \ \text{thick=3.0, linestyle=1}
\end{align*}
\]
Figure 11.16: Shooting method to solve for standing waves in a string. The solid, dotted, dashed, and dot-dashed curves correspond to $k = 1.0, 2.0, 3.0,$ and $4.0$, respectively.

IDL> $k = 3.0 & \text{ludiffeq\_23, 'stdwaves', t, f, t0=0.0, tf=1.0, init=ic}$
IDL> oplot, t, f[0,*], thick=3.0, linestyle=2
IDL> $k = 4.0 & \text{ludiffeq\_23, 'stdwaves', t, f, t0=0.0, tf=1.0, init=ic}$
IDL> oplot, t, f[0,*], thick=3.0, linestyle=3

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

IDL> $k1 = \text{findgen(101)/5.0}$
IDL> soln = fltarr(101)
IDL> for i = 0, 100 do begin $
IDL> \quad k = k1[i] &$
IDL> \quad \text{ludiffeq\_23, 'stdwaves', t, f, t0=0.0, tf=1.0, init=ic} &$
IDL> \quad q = \text{size(t)} & \text{soln[i]} = f[ 0, q[1]-1 ] &$
IDL> \quad \text{endfor}$
IDL> plot, k1, soln, ticklen=1.0, thick=3.0

Here, we (1) establish values for $k$, (2) create an array for the values to be saved, (3) execute a loop that (a) copies the next $k$ into the common area, (b) invokes \text{ludiffeq\_23} to solve the equations, (c)
uses size to determine the number of entries in the arrays containing the solution, and (d) stores the final value (which will be in row q[i]-1), and (4) plot the resulting values.

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 \( 0 \leq k \leq 20 \).

### 11.7.8 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\(^{54} 0 \leq y < \infty \). To set up the problem, we must first regard the second-order equation as a pair of first-order equations in which we associate \( \psi_0 \) with \( \psi \) and \( \psi_1 \) with \( d\psi/dy \). The equations we must solve then become

\[
\frac{d}{dy} \psi_0 = \psi_1 ; \quad \frac{d}{dy} \psi_1 = -(2\epsilon - y^2)\psi_0
\]

and a suitable pro-file \texttt{qmshm.pro} defining these equations for IDL’s solvers is recorded in Table 11.13. With this file stored in the default directory with the name \texttt{qmshm.pro}, we are ready to seek its solutions.

\(^{54}\)We here use the symbol \( y \) for what we earlier called \( x \).
FUNCTION qmshm, y, psi

; QMSHM defines the dimensionless Schroedinger equation for the quantum
; simple harmonic oscillator, which is
; 2
; d psi 2
; ----- + (2 ee - y ) psi = 0
; 2
; dy
;
; Entries are psi and d psi/ds, and initial conditions will thus be
; [psi0, dpsi0]. The parameter ee represents a dimensionless energy
; and must be passed to the routine in a common area named params.
;
; common params, ee
RETURN, [ psi[1], -(2*ee-y^2)*psi[0] ]
END

We seek first solutions of even parity, i.e., solutions for which $\psi(0) \neq 0$ and $d\psi(0)/dy = 0$. With those values, we can address this boundary value problem as if it were an initial value problem—except that we must reject solutions that do not go to zero as $y \to \infty$. Only $\psi(0)$ and $\epsilon$ are adjustable as we seek to impose the boundary condition at infinity. Since the ODE we are solving is linear, however, we recognize right away that tampering with $\psi(0)$ will merely affect the scaling of the solution and has no power to convert a solution that doesn’t go to zero at infinity into one that does go to zero. The only parameter we need bother adjusting is $\epsilon$. Let us, therefore, standardize by setting $\psi(0) = 1.0$ and explore the dependence of the solution on $\epsilon$. At the outset, we do not know what to expect. We do, however, believe that, in a dimensionless casting, important quantities are likely to have values on the order of one. Thus, we begin a search for acceptable solutions by setting $\epsilon = 1.0$ and examining the solution in the interval $0.0 \leq y \leq 2.0$. The statements

IDL> common params, ee
IDL> ic = [ 1.0, 0.0 ] & ee = 1.0
IDL> ludiffeq_23, 'qmshm', y, psi, t0=0.0, tf=2.0, init=ic
IDL> plot, y, psi[0,*]

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

IDL> e = [ 0.3, 0.4, 0.5, 0.6, 0.7 ]
IDL> q = size(e)
IDL> for i = 0, q[1]-1 do begin $
IDL> \quad ee = e[i] &$
IDL> \quad ludiffeq_23, 'qmshm', y, psi, t0=0.0, tf=4.0, init=ic &$
IDL> \quad if i eq 0 then plot, y, psi[0,*], yrange=[-2.0,2.0], thick=3 &$
IDL> else oplot, y, psi[0,*], thick=3 &$

Table 11.13: The IDL function qmshm.pro.
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).

IDL> endfor
IDL> oplot, [0.0,4.0], [0.0, 0.0]
IDL> xyouts, 2.0, 1.7, '0.3'
IDL> xyouts, 2.8, 1.2, '0.4'
IDL> xyouts, 3.5, 0.1, '0.5'
IDL> xyouts, 2.8, -1.0, '0.6'
IDL> xyouts, 2.2, -1.6, '0.7'

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

IDL> e = [ 1.0, 1.5, 2.0, 2.5, 3.0, 3.5 ]

to set the values of \( \epsilon \) and the lines

IDL> xyouts, 2.1, -1.2, '1.0'
IDL> xyouts, 1.7, -1.7, '1.5'
IDL> xyouts, 2.7, -1.8, '2.0'
IDL> xyouts, 3.5, -0.3, '2.5'
IDL> xyouts, 2.7, -1.8, '2.0'
IDL> xyouts, 3.0, 1.2, '3.0'
IDL> xyouts, 2.5, 1.7, '3.5'
Figure 11.19: The solution to the Schrödinger equation for an even state of the quantum harmonic oscillator when \( \epsilon \) has the indicated values. As we shall confirm in the next figure, this state is actually the second excited state of the harmonic oscillator.

to place the labels appropriately, produces the graph shown in Fig. 11.19. From this graph, we conclude that \( \epsilon = 2.5 \) yields another solution that goes to zero at infinity, and we argue that another allowed energy corresponds to the value \( \epsilon = 2.5 \). We might even be tempted to speculate—correctly as it turns out—that additional allowed energies for even states will correspond to the values \( \epsilon = 4.5, 6.5, 8.5, \ldots \).

We turn next to the odd states, for which we set \( \psi(0) = 0.0 \) and \( d\psi(0)/dy = 1.0 \) before exploring solutions for various values of \( \epsilon \). Indeed, the previous paragraphs suggest that we might expect to find the lowest odd state corresponding to \( \epsilon = 1.5 \). To test that expectation, we set the initial conditions and energies with the statements

IDL> ic = [0.0, 1.0]
IDL> e = [ 1.3, 1.4, 1.5, 1.6, 1.7 ]

Then we execute the statements

IDL> q = size(e)
IDL> for i = 0, q[1]-1 do begin $
IDL> ee = e[i] &$
IDL> ludiffeq_23, 'qmshm', y, psi, t0=0.0, tf=4.0, init=ic &$
IDL> if i eq 0 then plot, y, psi[0,*], yrange=[-2.0,2.0], thick=3 $
IDL> else oplot, y, psi[0,*], thick=3 &$
IDL> endfor$
IDL> oplot, [0.0,4.0], [0.0, 0.0]

to generate the graph in Fig. 11.20 and the statements
Figure 11.20: The solution to the Schrödinger equation for an odd state of the quantum harmonic oscillator when $\epsilon$ has the indicated values.

IDL> xyouts, 2.7, 1.7, '1.3'
IDL> xyouts, 3.5, 1.5, '1.4'
IDL> xyouts, 3.5, 0.1, '1.5'
IDL> xyouts, 3.4, -1.0, '1.6'
IDL> xyouts, 2.8, -1.5, '1.7'

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.

### 11.7.9 Pro-files in the Public Library

The following pro-files are found in the directory `$HOME/idl` (and may also at some sites have been placed in the directory `$IDLHEAD/lib/idl`, which is automatically in the IDL search path) and can be used to explore the systems they describe. Further information about the file `decay.pro`, for
example, can be obtained either with the command\textsuperscript{55,56}

\texttt{IDL> doc\_library, 'decay', directory='$\text{HEAD/idl}'}

to IDL or with whatever command to your operating system lists the file \texttt{$\text{HEAD/idl/decay.pro}$} on the screen.\textsuperscript{57} The first of these commands instructs IDL to find the file \texttt{decay.pro} and display the comments between the line \texttt{;+} and the line \texttt{;-} at the beginning of the file; 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 in a common area named \texttt{params}; examine the help message and/or the file itself to determine the structure of the necessary common area.

- \texttt{2oscil.pro}, which describes a system of two equal masses between two walls, all connected by three springs. Dimensionless units are used.
- \texttt{apollo.pro}, which describes a satellite in the gravitational attraction of two fixed suns of equal mass.
- \texttt{dampharm.pro}, which describes the one-dimensional damped, unforced harmonic oscillator using dimensional variables.
- \texttt{damposc.pro}, which describes the one-dimensional, damped, unforced harmonic oscillator using dimensionless variables.
- \texttt{decay.pro}, which describes three-species radioactive decay.
- \texttt{drvnosc.pro}, which describes the one-dimensional, damped, forced harmonic oscillator using dimensional variables.
- \texttt{drvnpend.pro}, which describes the forced, damped simple pendulum in dimensionless units.
- \texttt{henon.pro}, which describes the Hénon-Heiles oscillator.
- \texttt{largamp.pro}, which describes the undamped, unforced large amplitude pendulum in dimensionless units.
- \texttt{lorenz.pro}, which describes the Lorenz attractor.
- \texttt{onedshm.pro}, which describes the one-dimensional, undamped, unforced harmonic oscillator using dimensional variables.
- \texttt{planet.pro}, which 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.
- \texttt{qmshm.pro}, which describes the dimensionless quantum harmonic oscillator.
- \texttt{rossler.pro}, which describes the Rössler attractor.
- \texttt{stdwaves.pro}, which describes the behavior of standing waves in a string.
- \texttt{twodshm.pro}, which describes the two-dimensional, undamped, unforced harmonic oscillator using dimensionless variables.
- \texttt{vandpol.pro}, which describes the Van der Pol oscillator.

\textsuperscript{55}While it is no longer allowed, older versions of IDL permitted the specification of the path as part of the file name, e.g., \texttt{doc\_library, '$\text{HEAD/idl/decay}'}.  
\textsuperscript{56}The keyword \texttt{directory} is available only in a UNIX or LINUX operating system.  
\textsuperscript{57}In UNIX, one possible command would be \texttt{more $\text{HEAD/idl/decay.pro}$; see the Local Guide}.  

11.13 Solving ODEs Numerically with Mathematica

Mathematica's command NDSolve provides a means to solve ordinary differential equations numerically. A full statement using NDSolve will have the form

\[
\text{NDSolve}\left\{ \begin{align*}
&ODE_1, ODE_2, \ldots, IC_1, IC_2, \ldots, \\
&DVar_1, DVar_2, \ldots,
\end{align*}\right\}, \{DVar_1, DVar_2, \ldots\}, \\
\{IVar, IVarMin, IVarMax \} \right\]
\]

where ODE and IC are the differential equations and initial conditions, DVar represents a dependent variable, and IVar, IVarMin, and IVarMax are the independent variable and the lower and upper limits on the range in which the solution is to be found. In this case, both the differential equations and the initial conditions must be specified; Mathematica will not introduce undetermined integration constants. Further, all parameters in the equations must be given explicit numerical values before NDSolve is executed. When invoked, NDSolve returns not a solution but an interpolation function that, once a specific numerical value has been assigned to the independent variable, can be used to generate a numerical value for the solution at the specified value of the independent variable. Thus, for example, to address the problem of chain radioactive decay that we have already solved symbolically in Section 11.5.4, we would request Mathematica to generate the interpolation function by invoking the statements

\[
\text{In}[1]:= \text{eqA} = \frac{dAA}{dt} = -kA*AA[t]; \\
\text{In}[2]:= \text{eqB} = \frac{dBB}{dt} = kA*AA[t] - kB*BB[t]; \\
\text{In}[3]:= \text{eqC} = \frac{dCC}{dt} = kB*BB[t]; \\
\text{In}[4]:= kA = 0.1; kB = 0.1; \\
\text{In}[5}=:= \text{soln} = \text{NDSolve}[\{\text{eqA}, \text{eqB}, \text{eqC}, \}
\{AA[0]==1000, BB[0]==0, CC[0]==0\}, \\
\{AA, BB, CC\}, \{t, 0, 50\}]
\]

Out[5]= \{
\{AA \to \text{InterpolatingFunction}[\text{Graphic}], BB \to \text{InterpolatingFunction}[\text{Graphic}], \\
CC \to \text{InterpolatingFunction}[\text{Graphic}]\} \}

Then, after the interpolating function has been created, the solution at any particular point can be found with statements like

\[
\text{In}[6]=:= \{AA[0.0], BB[0.0], CC[0.0]\} /. \text{soln}
\]
Out[6]= \{1000., 0., 0.\}

\[
\text{In}[7]=:= \{AA[0.1], BB[0.1], CC[0.1]\} /. \text{soln}
\]
Out[7]= \{990.05, 9.9005, 0.0496679\}

Note the form of the solution, which is presented as a list of lists, though this particular solution has only one list in the outer list. In that (one) list, the elements are the numerical values of the solution at the selected value of the independent variable—time in this case.

Should we wish a graph of the solution over some range of the independent variable, we would invoke the command Plot but use the command Evaluate to determine the values to be plotted. A graph similar to Fig. 11.9 for this decay would, for example, be displayed by the statement

\[
\text{In}[8]=:= \text{Plot}[\text{Evaluate}[\{AA[t], BB[t], CC[t]\} /. \text{soln}, \{t, 0.0, 50.0\} ], \\
\text{PlotStyle} \to \text{Thickness}[0.01], \text{AxesLabel} \to \{"t", "A,B,C"\} ]
\]

\[\text{58Because C is a protected variable in Mathematica, the dependent variables here are named AA, BB, and CC rather than A, B, and C.}\]

\[\text{59The inclusion identified here with the word ‘Graphic’ is a box that has a small graph of the function as well as an indication of the domain of the solution and the character of the output. Clicking ML on the ‘+’ sign in that graphic brings up information about other features of the solution.}\]
Additional keywords similar to those available for the command Plot can be exploited to add elements to the graph and/or override additional defaults.

The previous two paragraphs illustrate the essential strategy adopted by Mathematica to produce numerical solutions to ordinary differential equations. We here draw attention to a few additional capabilities, leaving an exhaustive enumeration and the details to the Mathematica manuals. While the defaults adopted for controlling accuracy and for protecting against infinite loops are often acceptable, the command NDSolve nonetheless possesses several options that give us the ability to override these defaults. Among these options are

- **WorkingPrecision** (default $\text{MachinePrecision}$, often 16), which specifies the number of digits to be used for internal computations.

- **AccuracyGoal** (default usually figured as half the value of WorkingPrecision), which specifies the absolute accuracy to be sought by stipulating the number of digits to be provided in the solution. AccuracyGoal should not exceed WorkingPrecision.

- **PrecisionGoal** (default usually figured as half the value of WorkingPrecision), which specifies the relative accuracy sought in the final result. PrecisionGoal should not exceed WorkingPrecision.

  *Note:* With AccuracyGoal set to $a$ and PrecisionGoal set to $p$, Mathematica strives for an absolute accuracy in a value $x$ to be less than $10^{-a} + 10^{-p}|x|$.

- **MaxSteps** (default automatic), which specifies the maximum number of steps to take in generating the solution. This option provides a means to stop the iterations even if goals on accuracy and precision have not been met.

- **MaxStepSize** (default automatic), which limits the size of the largest steps to be taken in generating the solution. This option provides a means to specify, for example, a maximum step size smaller than Mathematica determines to be adequate.

Remember, too, that the statement `Options[NDSolve]` will display all options and their current default values for the indicated command.

If we want solutions at a specific set of values of the independent variable, we could exploit the command Table. For example, the statement

```mathematica
In[9]:= tbl = Table[{t, AA[t], BB[t], CC[t]} /. soln, {t, 0, 20, 5}]
Out[9]= 
{ 
{{0, 1000., 0., 0.}},
{{5, 606.531, 303.265, 90.204}},
{{10, 367.879, 367.88, 264.241}},
{{15, 223.13, 334.695, 442.175}},
{{20, 135.335, 270.671, 593.994}}
}
```

produces a table named tbl whose $i$-th element contains four values (the time and the values of $A$, $B$, and $C$ at that time). This output can then be manipulated using any of Mathematica’s resources for working with tables.

### 11.16 Exercises

#### 11.16.1 using Symbolic Methods

11.1. Find the motion of a driven, damped oscillator satisfying the differential equation

$$m \frac{d^2 x}{dt^2} + b \frac{dx}{dt} + kx = mf \cos \omega t$$
subject to the general initial conditions
\[ x(0) = x_0 \quad ; \quad \frac{dx}{dt}(0) = v_0 \]
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
\[ m_1 \frac{d^2 x_1}{dt^2} = k(x_2 - x_1) \quad ; \quad m_2 \frac{d^2 x_2}{dt^2} = -k(x_2 - x_1) \]
Let the system be put into motion with arbitrary initial conditions
\[ x_1(0) = x_{10} \quad ; \quad x_2(0) = x_{20} \quad ; \quad \frac{dx_1}{dt}(0) = v_{10} \quad ; \quad \frac{dx_2}{dt}(0) = v_{20} \]
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 the system shown in Fig. 11.5 is extended to contain three objects coupled in a line. Take the four springs all to have the same spring constant but allow for the possibility that the middle object may have a mass different from that of the two outside objects. Your examination should include at least the following:

- Find both the natural frequencies of the normal modes of oscillation and the initial conditions that will excite the system exclusively in each of the normal modes, determining particularly the frequencies as a function of the ratio of the two masses. To display your results, draw a diagram something like a quantum energy level diagram that shows the way the three natural frequencies vary as the middle mass goes from being rather smaller than the two outer masses to being rather larger than the outer two masses. As initial conditions, it will be sufficient to release the objects from rest at individually arbitrary displacements. You are, however, quite likely to have to help with the inverse Laplace transforms, since your symbolic manipulator may be unable to deal with the inverse Laplace transform of an expression like
\[ \frac{As + Bs^3}{\alpha s^4 + \beta s^2 + \gamma} = s \left( \frac{A + Bs^2}{\alpha s^4 + \beta s^2 + \gamma} \right) \]
The denominator looks quartic in \( s \) but is better seen as quadratic in \( s^2 \). If you introduce \( z = s^2 \) temporarily, factor the resulting denominator, and then do a partial fraction expansion, you can end up with the alternative form
\[ s \left( \frac{A + Bs^2}{\alpha s^4 + \beta s^2 + \gamma} \right) = s \left( \frac{R}{s^2 + R'} + \frac{Q}{s^2 + Q'} \right) = \frac{Rs}{s^2 + R'} + \frac{Qs}{s^2 + Q'} \]
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.
11.4. The system called the double pendulum shown in Fig. 11.21 consists of a ball of mass \( m_1 \) hanging from a rigid and massless rod of length \( l_1 \) attached to the ceiling and a second ball of mass \( m_2 \) hanging from a rigid and massless rod of length \( l_2 \) attached to the first ball. The balls swing in a plane, and the configuration of the system is specified by giving two angles, the first of which, \( \theta \), gives the angle that the upper string makes with the vertical and the second of which, \( \phi \), gives the angle that the lower string makes with the vertical. The motion can be very complicated and at times will be chaotic. For small amplitudes, however, things are much more sedate. When the amplitudes of the motion of both balls are small and—to simplify a little bit—when the strings are both the same length (\( l_1 = l_2 \)), which we will symbolize with the letter \( l \)), the equations of motion turn out to be

\[
\begin{align*}
\frac{d^2 \theta}{dt^2} + \frac{m_2}{m_1 + m_2} \frac{d^2 \phi}{dt^2} + \frac{g}{l} \theta &= 0 \\
\frac{d^2 \phi}{dt^2} + \frac{d^2 \theta}{dt^2} + \frac{g}{l} \phi &= 0
\end{align*}
\]

Find the normal modes of oscillation of this system and determine the initial conditions that will cause the system to oscillate exclusively in one or the other of these modes.

11.5. Solve the differential equation

\[
m \frac{d^2 x}{dt^2} + b \frac{dx}{dt} + kx = mf \cos \omega t
\]

for the driven, damped harmonic oscillator subject to the specific initial conditions

\[
x(0) = x_0 ; \quad \frac{dx}{dt}(0) = 0
\]
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) Show that the polar components of the equations of motion in the first instance are

\[ m \left( \frac{d^2 r}{dt^2} - r \left( \frac{d\phi}{dt} \right)^2 \right) = f(r) ; \quad m \left( \frac{d^2 \phi}{dt^2} + 2 \frac{dr}{dt} \frac{d\phi}{dt} \right) = 0 \]

(b) Recognizing that

\[ \frac{d}{dt} \left( mr^2 \frac{d\phi}{dt} \right) = m \left( r^2 \frac{d^2 \phi}{dt^2} + 2r \frac{dr}{dt} \frac{d\phi}{dt} \right) \]

show that

\[ mr^2 \frac{d\phi}{dt} = \text{constant} = L \quad \Rightarrow \quad \frac{d\phi}{dt} = \frac{L}{mr^2} \]

and then that

\[ m \frac{d^2 r}{dt^2} = f(r) + \frac{L^2}{mr^3} \]

Here, \( L = m(x_0 v_{x0} - y_0 v_{y0}) \) 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,
11.16. EXERCISES

251

11.12. In an appropriate dimensionless presentation, standing waves in a string must satisfy the boundary value problem

\[ \frac{d^2 y}{dx^2} + k^2 y = 0 \quad ; \quad y(0) = y(1) = 0 \]

Suppose that the interval \( 0 \leq x \leq 1 \) is divided into \( n \) equal segments of length \( \Delta x = 1/n \), let \( x_i = i \Delta x \) (with \( i = 0, 1, 2, \ldots, n \)), and let \( y_i = y(x_i) \). Evaluate the ODE at \( x = x_i \), approximate the second derivative with the difference formula

\[ \frac{d^2 y}{dx^2} \bigg|_{x=x_i} \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{\Delta x^2} \quad ; \quad i = 1, 2, 3, \ldots, n - 1 \]

and note that \( y_0 = y_n = 0 \). Show that the values \( y_i \) for \( i = 0, 1, 2, \ldots, n \) satisfy a system of \( n + 1 \) linear algebraic equations of the form \( MY = \alpha Y \), where \( Y \) is an \( (n + 1) \)-component vector whose elements are the values of \( y_i \) and \( \alpha \) is determined from \( k^2 \) and \( \Delta x \). Then argue that the allowed values of \( k^2 \) can be determined from the eigenvalues of the matrix \( M \). That is, show that this transformation turns a boundary value problem involving a differential equation into an approximately equivalent matrix eigenvalue problem.

\[ \text{Hint: Remember (or take as given if you haven’t met them yet) that the radial and azimuthal components of the acceleration of a particle in polar coordinates are given by} \ a_r = \ddot{r} - r\dot{\phi}^2 \text{ and} \ a_{\phi} = r\ddot{\phi} + 2\dot{r}\dot{\phi}. \]
11.13. `ludiffeq_23` and `ludiffeq_45` allow an additional argument called “report” which instructs IDL to write to the screen the time value, step size, and dependent variable for each row in the data table. This option is selected using the syntax

```
ludiffeq_23, 'function name', t, x, t0=start, tf=finish, init=ic, /report
```

in the call to `ludiffeq_23` or `ludiffeq_45`. Primarily so that you can see the adaptive nature of the time step of these routines, redo the example of the radioactive decay sequence selecting the `/report` option. Solve the differential equation several times using several different tolerances and notice the effect different tolerances have on the time step. Also, compare the time step used by `ludiffeq_23` and `ludiffeq_45` in solving the same problem to the same tolerance.

11.14. Some radioisotopes exhibit a branched decay sequence such as

```
A \quad k_1 \quad B \quad k_3 \quad D
```
```
\quad k_2 \quad C \quad \quad \quad k_4
```

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 $\pi(0) = 1$ and $v(0) = 0$, produce graphs of $\pi$ versus $t$ for several values of $\beta$ on the interval $0.0 < \beta < 2.0$. Then use the knowledge that $\beta = b/m\omega = b/\sqrt{mk}$, and that $t = \omega t$, 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

```latex
\frac{d^2\pi}{dt^2} = \frac{d\pi}{dt}(1 - \pi^2) - \pi
```

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

```latex
\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

```latex
\frac{d^2\pi}{dt^2} = -\frac{\pi}{(\pi^2 + \eta^2)\psi}; \quad \frac{d^2\eta}{dt^2} = -\frac{\eta}{(\pi^2 + \eta^2)\psi}
```
11.16. **EXERCISES**

[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 looks 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 $\mathbf{E} = E_x \hat{i}$ and $\mathbf{B} = B_z \hat{k}$. In vector form, the equation of motion for this particle is

$$m \frac{d^2 \mathbf{r}}{dt^2} = q \mathbf{E} + q \frac{d \mathbf{r}}{dt} \times \mathbf{B}$$

Verify the equations of motion

$$m \frac{d^2 x}{dt^2} = qE_x + qB_z \frac{dy}{dt} \quad ; \quad m \frac{d^2 y}{dt^2} = -qB_z \frac{dx}{dt} \quad ; \quad m \frac{d^2 z}{dt^2} = 0$$

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 $t$, the equations involve a single parameter $qE_x/(m\omega^2 \ell)$, which can alternatively be written as $(E_x/B_z)/\omega t$ — the ratio of the velocity $E_x/B_z$ determined by the fields to the characteristic velocity implied by your choice of a length unit and the frequency $\omega$. Note that this exercise actually has more than one parameter, since the initial components of the velocity—probably expressed in units of $\omega \ell$—also influence the solution.

11.23. A particle having mass $m$ and carrying charge $q$ moves in the $xy$-plane while experiencing an electric field given by $\mathbf{E}(x, y) = -\alpha y \hat{j}$, where $\alpha$ is a constant. Assume that $\alpha$ and $q$ are both positive. With $\mathbf{r} = x \hat{i} + y \hat{j}$, the vector equation of motion for this particle then is

$$m \frac{d^2 \mathbf{r}}{dt^2} = -q\alpha y \hat{j}$$

(a) Show that, in component form, the equations of motion for this particle are

$$\frac{d^2 x}{dt^2} = 0 \quad ; \quad \frac{d^2 y}{dt^2} = -\frac{\alpha q}{m} y = -by$$

(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

$$\frac{dx}{dt} = a(y - x)$$

$$\frac{dy}{dt} = -xz + bx - y$$

$$\frac{dz}{dt} = xy - cz$$
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 $x$, $z$ versus $x$, and $z$ versus $y$ are interesting, the true beauty of the trajectory is best seen using a three-dimensional space curve. Be sure to examine the path from several different vantage points in the space around the path, an objective most easily accomplished if the display of the path allows rotation of the path on the screen. (The method required to produce such a plot is discussed in Section 2.16.2.)

11.27. The dynamics of the chemical reaction

\[ A + B \rightleftharpoons C + D \]

is governed by the equations

\[
\begin{align*}
\frac{dA}{dt} &= -k_f AB + k_r CD \\
\frac{dB}{dt} &= -k_f AB + k_r CD \\
\frac{dC}{dt} &= k_f AB - k_r CD \\
\frac{dD}{dt} &= k_f AB - k_r CD
\end{align*}
\]

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 an 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

\[
\begin{align*}
\frac{dx}{dt} &= -k_1 x + k_2 xy \\
\frac{dy}{dt} &= k_3 y - k_4 xy
\end{align*}
\]

where $k_1$ and $k_3$ are parameters describing the way each population would evolve in the absence of the other and $k_2$ and $k_4$ are parameters describing strength of the interaction between the two species, which we take to be proportional to the likelihood of an encounter between a member of one species and a member of the other species. Depending on the parameters and the initial populations $x(0) = x_0$, $y(0) = y_0$, the system may approach a stable equilibrium or, alternatively, one or the other of the populations may become extinct. Explore this system to determine conditions under which each of these circumstances occurs and write a paragraph or two describing your findings. Make sure your results are generated to an accuracy adequate to support your conclusions.

11.29. Examine the undamped, unforced harmonic oscillator carefully, using at least two different methods and several time steps. Monitor the accuracy of your solution by monitoring the constancy of the energy of the oscillator given (in dimensional form) by $E = \frac{1}{2}mv^2 + \frac{1}{2}kx^2$.

11.30. In the text, we illustrated the use of one or more numeric processing programs to solve the chain radioactive decay problem of Section 11.1.2 using Euler’s method. Using an available numeric processing program like IDL, MATLAB, OCTAVE, or PYTHON, repeat that development using the improved Euler method and comment on what the resulting graphs reveal about the adequacy of the several time steps used.

11.31. The type of child’s swing shown in Fig. 11.22 is hung with elastic ropes. Suppose the ropes are long enough so that the child/swing can be represented by a point mass at the end of a spring, assume the spring obeys Hooke’s law with constant $k$ and has an unstretched length $a$, and let the motion of the mass be confined to a single vertical plane. Show that, in the coordinate system illustrated,
the equations of motion are

\[ m \frac{d^2x}{dt^2} = -kx + \frac{kax}{\sqrt{x^2 + y^2}} \]

\[ m \frac{d^2y}{dt^2} = mg - ky + \frac{kay}{\sqrt{x^2 + y^2}} \]

Then, introducing \( \omega_0^2 = k/m \), \( \bar{T} = \omega_0 t \), \( \bar{x} = x/a \), and \( \bar{y} = y/a \), cast the equations in dimensionless form. After the suggested rescalings, only one parameter—\( g/\omega_0^2 \)—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

\[
\frac{dp}{dt} = f \quad \text{and} \quad \frac{dx}{dt} = v \tag{13.1}
\]

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

\[
p = \frac{mv}{\sqrt{1 - v^2/c^2}} \quad \text{or} \quad p = mv \tag{13.2}
\]
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 physical problem of predicting the trajectory reduces to the 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., \( m \), \( c \), ...) in the integrand (unless a dimensionless
casting happens to suppress them). These integrals fall into either the second or the third of our
three categories.

If, on the other hand, the force happens to depend only on \( x \), we can recast the computational
task by noting first that

\[
\frac{d^2x}{dt^2} = \frac{dv}{dt} = \frac{dv}{dx} \frac{dx}{dt} = v \frac{dv}{dx}
\]

(13.4)

Then, Newton's second law becomes

\[
m \frac{d^2x}{dt^2} = f(x) \implies mv \frac{dv}{dx} = f(x) \implies mv \, dv = f(x) \, dx
\]

(13.5)

Finally, by integrating this last expression from initial values to general values at some other time,
we find that

\[
m \int_{x_0}^x v' \, dv' = \frac{1}{2} mv^2 - \frac{1}{2} mv_0^2 = \int_{x_0}^x f(x') \, dx'
\]

(13.6)

and the task of finding the velocity (as a function of \( x \)) is reduced to the straight-forward evaluation
of an integral. Once that integral has been evaluated, we can then predict the position by exploiting
the relationship

\[
\frac{dx}{dt} = v(x) \implies dt = \frac{dx}{v(x)} \implies \int_{t_0}^{t'} dt' = t - t_0 = \int_{x_0}^{x'} \frac{dx'}{v(x')}
\]

(13.7)

and we have reduced the second step in the solution to the evaluation of an integral as well. Both
integrals deduced in this paragraph may have internal parameters, so they might fall into either the
second or the third of our three categories.

\section{13.1.2 Center of Mass}

The location \( \mathbf{r}_{cm} \) of the center of mass of an object is given by

\[
\mathbf{r}_{cm} = \frac{1}{M} \int \mathbf{r} \, dm
\]

(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

\[
\mathbf{r}_{cm} = \frac{1}{M} \int \sigma(r, \phi) \, r \, dr \, d\phi = \frac{1}{M} \int \sigma(r, \phi) \left( r \cos \phi \, \hat{i} + r \sin \phi \, \hat{j} \right) \, r \, dr \, d\phi
\]

(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 \( \sigma = M/(1/2\pi a^2) \) shown in Fig. 13.1, we would conclude that

\[
\mathbf{r}_{cm} = \frac{1}{M} \int_0^a \int_0^{\pi} \frac{M}{1/2\pi a^2} \left( r \cos \phi \, \mathbf{i} + r \sin \phi \, \mathbf{j} \right) r \, dr \, d\phi = \frac{2}{\pi a^2} \int_0^a \int_0^{\pi} r^2 \left( \cos \phi \, \mathbf{i} + \sin \phi \, \mathbf{j} \right) \, dr \, d\phi
\]

(13.10)

Recognizing as always the wisdom of casting problems to be addressed with a computer in dimensionless form, we finally introduce the dimensionless length \( \lambda = r/a \), in terms of which we then find that

\[
\frac{\mathbf{r}_{cm}}{a} = \frac{2}{\pi} \int_0^{\pi/2} \int_0^1 \lambda^2 \left( \cos \phi \, \mathbf{i} + \sin \phi \, \mathbf{j} \right) \, d\phi \, d\lambda
\]

(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 \( \sigma = M/(1/2\pi a^2) \). 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)
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

\[
I = \frac{4M}{\pi a^2} \int_0^a y^2 \sqrt{a^2 - y^2} \, dy = \frac{4}{\pi} Ma^2 \int_0^1 \lambda^2 \sqrt{1 - \lambda^2} \, d\lambda
\]

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

\[
\frac{1}{2} ml^2 \left( \frac{d\theta}{dt} \right)^2 - mgl \cos \theta = -mgl \cos \theta_0 \implies \frac{d\theta}{dt} = \pm \sqrt{\frac{2g}{l}} \sqrt{\cos \theta - \cos \theta_0}
\]  

(13.18)

(We assume that the pendulum does not have sufficient energy to swing over the top.) This last relationship then leads to the conclusion that

\[
dt = \pm \sqrt{\frac{l}{2g}} \frac{d\theta}{\sqrt{\cos \theta - \cos \theta_0}} \implies \int_0^{T/4} dt = \frac{T}{4} = \sqrt{\frac{l}{2g}} \int_0^{\theta_0} \frac{d\theta}{\sqrt{\cos \theta - \cos \theta_0}}
\]  

(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 \( dq \) so at last find that \( dq \) for a distribution of static charge, for example, we identify an element of that charge \( dq \) so at last find that 

\[
\text{dq}
\]

which 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.

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 \( r = r \cos \phi \hat{i} + r \sin \phi \hat{j} + z \hat{k} \), produced by one such ring positioned in the \( xy \) plane with its center at the origin. Let \( r' = a \cos \phi' \hat{i} + a \sin \phi' \hat{j} \) locate an element on that (single) ring. In this notation,

\[
r - r' = (a \cos \phi - a \cos \phi') \hat{i} + (a \sin \phi - a \sin \phi') \hat{j} + z \hat{k}
\]

so

\[
|r - r'| = \left[ (r \cos \phi - a \cos \phi')^2 + (r \sin \phi - a \sin \phi')^2 + z^2 \right]^{1/2}
\]

\[
= \left[ r^2 + a^2 - 2ar \cos(\phi' - \phi) + z^2 \right]^{1/2} \quad (13.31)
\]

Placing these results into the second member of Eq. (13.28) and recognizing that \( dq' = \lambda a \, d\phi' \), we at last find that

\[
V_{\text{one}}(r, \phi, z) = \frac{1}{4\pi \epsilon_0} \int_0^{2\pi} d\phi' \frac{\lambda a \, d\phi'}{\sqrt{r^2 + a^2 - 2ar \cos(\phi' - \phi) + z^2}} \quad (13.32)
\]
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

\[
V_{\text{one}}(r, \phi, z) = \frac{\lambda a}{4\pi \epsilon_0} \int_0^{2\pi} \frac{d\alpha}{\left[ r^2 + a^2 - 2ar \cos \alpha + z^2 \right]^{1/2}} \tag{13.33}
\]

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 \( 2a \). 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

\[
V_{\text{midplane}}(r) = V_{\text{one}}(r, \phi, -a) + V_{\text{one}}(r, \phi, a) = \frac{2\lambda a}{4\pi \epsilon_0} \int_0^{2\pi} \frac{d\alpha}{\left[ r^2 + 2a^2 - 2ar \cos \alpha \right]^{1/2}} \tag{13.34}
\]

Finally, introducing the variable \( s \) defined by \( r = sa \) to express the radial coordinate in dimensionless terms, we find the expression

\[
V_{\text{two}}(s) = V_{\text{midplane}}(sa) = \frac{2\lambda}{4\pi \epsilon_0 a} \int_0^{2\pi} \frac{d\alpha}{\left[ s^2 - 2s \cos \alpha + 2 \right]^{1/2}} \tag{13.35}
\]

or, even better, the expression

\[
V(s) = \frac{\sqrt{2}}{2\pi} \int_0^{2\pi} \frac{d\alpha}{\left[ s^2 - 2s \cos \alpha + 2 \right]^{1/2}} \tag{13.36}
\]

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 \( 2a \). 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 (\( \frac{1}{2} \hbar \omega \) for the ground state) is equal to the potential energy \( \frac{1}{2} k x_{\text{turn}}^2 \), the turning point of this oscillator is given by

\[
\frac{1}{2} \hbar \omega = \frac{1}{2} k x_{\text{turn}}^2 \implies x_{\text{turn}} = \sqrt{\frac{\hbar \omega}{k}} = a
\]

(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

\[
\int_{a}^{b} \phi_i(x) \phi_j(x) \, w(x) \, dx = N_{ij} \delta_{ij}
\]

(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

\[
f(x) = \sum_n a_n \phi_n(x) \quad \iff \quad a_n = \frac{1}{N_n} \int_a^b w(x) \phi_n(x) f(x) \, dx \tag{13.45}
\]

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 is\(^1\)

\[
\{1, \sin \frac{n\pi x}{l}, \cos \frac{m\pi x}{l} ; \quad n, m = 1, 2, 3, \ldots \} \tag{13.46}
\]

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

\[
a_n = \frac{1}{l} \int_{-l}^l f(x) \cos \frac{n\pi x}{l} \, dx \quad ; \quad b_n = \frac{1}{l} \int_{-l}^l f(x) \sin \frac{n\pi x}{l} \, dx \tag{13.51}
\]

### 13.4 Evaluating Integrals Symbolically with Mathematica

The *Mathematica* command for symbolic integration is `Integrate`. The first argument in this command either is or stands for the integrand, and the second argument specifies the integration variable (with an optional range to convey limits). Thus, for example, the statement

```
Integrate[ Sin[k*x], x ]
```

evaluates the *indefinite* integral \(\int \sin(kx) \, dx\) while the statement

```
Integrate[ Sin[k*x], {x, a, b} ]
```

evaluates the definite integral \(\int_a^b \sin(kx) \, dx\).
Chapter 13. Evaluating Integrals

Integrate[ Sin[k*x], {x, 0, Pi} ]

evaluates the definite integral \( \int_{0}^{\pi} \sin(kx) \, dx \). If Mathematica is unable to evaluate a particular integral, the program returns the noun form of the integral, i.e., the program simply returns the integral that was provided as input. Mathematica also evaluates improper integrals, recognizing the symbols \( \text{Infinity} \) for \( +\infty \) and \( -\text{Infinity} \) for \( -\infty \). Whether integrating by hand or with the help of a symbolic program, we must always be wary of integrals when the integrand has a singularity in the interval of integration. Indeed, Mathematica checks for these singularities automatically and prints a warning if evidence of a singularity is found.

In this section, we illustrate the use of the command Integrate to evaluate some of the integrals deduced in Section 13.1. Beyond integration per se, we also show how, in many cases, Mathematica can be used to set up the integral as well. To abbreviate the presentation of Mathematica dialogs, we make liberal use of terminating semicolons to suppress intermediate output. You are therefore urged to duplicate the dialogs in an actual session with Mathematica, removing semicolons so that intermediate output is explicitly displayed.

13.4.1 Relativistic Motion Under a Constant Force

Suppose the particle to which Eq. (13.3) applies moves relativistically, starting from rest at the origin \([x(0) = 0, v(0) = 0, p(0) = 0]\), and experiences a constant force \( f \). Then, the basic relationships from which we would determine \( x(t) \) and \( v(t) \) are

\[
\begin{align*}
 p(t) &= \int_{0}^{t} f \, dt', \quad x(t) = \int_{0}^{t} v(t') \, dt', \quad \text{and} \quad p(t) = \frac{mv(t)}{\sqrt{1 - v(t)^2/c^2}} \quad (13.52)
\end{align*}
\]

The following “conversation” with Mathematica will find the desired quantities and explore a few of their properties:

```mathematica
In[1]:= p = Integrate[f, {tp, 0, t}] Out[1]= ft
In[3]:= soln = Solve[%, v] Out[3]= \{v \rightarrow -c ft/\sqrt{c^2 m^2 + f^2 t^2}, v \rightarrow c ft/\sqrt{c^2 m^2 + f^2 t^2}\}
In[4]:= v = soln[[2,1,2]] Out[4]= \frac{c f}{\sqrt{c^2 m^2 + f^2 t^2}}
In[5]:= Limit[v, t -> Infinity] Out[5]= \frac{c f}{\sqrt{f^2}}
In[6]:= Simplify[%, f > 0] Out[6]= c
```

\(^3\)Mathematica statements are shown on the left; comments describing the statements are shown on the right.
In[7]:= vp = v /. t -> tp;  
    Substitute tp for t in v.

In[8]:= x = Integrate[ vp, { tp, 0, t} , Assumptions->{c>0,m>0,t>0,f>0} ]
    Integrate vp to find position.

Out[8]= \[ c\left(-cm + \sqrt{c^2m^2 + f^2t^2}\right) \]

Substitute \( tp \) for \( t \) in \( v \).

Find limit of \( v \) for small \( t \).

In[9]:= Series[ v, { t, 0, 4 } ];

Out[9]= \[ ft - f^3t^3 + \frac{f^4t^4}{8c^2m^3} + O[t^5] \]

Find limit of \( x \) for small \( t \).

In[10]:= vclassical = Simplify[ %, {c>0, m>0, f>0} ]

Out[11]= \[ ft - \frac{f^3t^3}{2c^2m^3} + O[t^5] \]

In[12]:= xclassical = Simplify[ %, {c>0, m>0, f>0 } ]

Out[13]= \[ \frac{ft^2}{2m} - \frac{f^3t^4}{8c^2m^3} + O[t^6] \]

Reassuringly, the velocity approaches \( c \) as \( t \to \infty \) and, with the acceleration \( a \) identified as \( f/m \), \( x \) and \( v \) for small \( t \) (classical limit) are \( \frac{at^2}{2} \) and \( at \) respectively.

13.4.2 Center of Mass

We next find the center of mass of the uniform, semicircular plate described in Section 13.1.2 as given by Eq. (13.9). We evaluate this integral with the statements

In[1]:= rr = {r*Cos[\[Phi]],
    Assign \( r \).
    r*Sin[\[Phi]],0};

In[2]:= \[\Sigma\] = m / (Pi * a^2 / 2);
    Evaluate \( \sigma \), which is constant.

In[3]:= integ = rr * \[\Sigma\] * r;
    Evaluate integrand.

In[4]:= fstint = Integrate[ integ,
    Evaluate integral on \( \phi \).
    \{ \[Phi], 0, Pi \ } ];

In[5]:= scndint = Integrate[ fstint,
    Evaluate integral on \( r \).
    \{ r, 0, a} ];

In[6]:= rcm = scndint/m
    Multiply each component by \( 1/m \).

Out[6]= \[\{ 0, \frac{4a}{3\pi}, 0 \} \]

In[7]:= N[ % ]
    Obtain floating evaluation.

Out[7]=\{ 0, 0.424413a, 0. \}

In[8]:= Quit[]

Since only the \( y \) component of \( r_{cm} \) 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 \( \text{Mathematica} \) has at lines In[4] and In[5] effected an element-by-element integration of the components of the list with a single statement.

Note also that we have evaluated the double integral as a sequence of single integrals. In the present case, the two integrals are independent of one another, and the order in which we perform the integrals is irrelevant. In some cases, however, the order may be important—as, for example, when the limits on one integration variable happen to depend on the other integration variable.
13.4.3 Moment of Inertia; Radius of Gyration

The evaluation by Mathematica of the moment of inertia of a semicircular plate as given by Eq. (13.16) is quick and straightforward. We need only the statements

\[\text{In}[1] := \text{Integrate} \left[ \lambda^2 \sqrt{1-\lambda^2}, \{ \lambda, 0, 1 \} \right]\]

\[\text{Out}[1] = \frac{\pi}{16}\]

\[\text{In}[2] := II = m \cdot a^2 \cdot \left(\frac{4}{\pi}\right) \cdot \%\]

\[\text{Out}[2] = \frac{a^2 m}{4}\]

\[\text{In}[3] := k = \text{Simplify} \left[ \frac{\text{Sqrt}[II/m]}{a}, a>0 \right]\]

\[\text{Out}[3] = \frac{a}{2}\]

\[\text{In}[4] := \text{Quit}[]\]

to evaluate the necessary integral and find both the moment of inertia \(I = ma^2/4\) and the radius of gyration \(k = a/2\).

13.4.4 Electrostatic Potential of a Finite Line Charge

For a fourth example, we evaluate the electrostatic potential of a uniformly charged finite line extending along the \(z\) axis over the interval \(-a \leq z \leq +a\). The geometry is shown in Fig. 13.4. With \(\lambda\) representing the linear charge density on the line, \(dz'\) giving the length of an element of the line, \(r\) and \(r'\) locating the observation point and an element of the source, respectively, and \(dq' = \lambda \cdot dz'\), we deduce from Eq. (13.28) that the potential established by this source is given by

\[V(x, y, z) = \frac{\lambda}{4\pi \epsilon_0} \int_{-a}^{+a} \frac{dz'}{\sqrt{r^2 + (z - z')^2}}\]  
(13.53)

Electing 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

\[\text{In}[1] := \text{Get} \left[ "$\text{HEAD/mathematica/crossdot.m"} \right]\]

\[\text{Define} \text{ludot} \text{and} \text{lucross}.
\]

\[\text{In}[2] := \text{rr} = \{ r \cdot \text{Cos}[\phi], r \cdot \text{Sin}[\phi], z \} ;\]

\[\text{Assign} \text{r in cylindrical coordinates.}\]

\[\text{In}[3] := \text{rp} = \{ 0, 0, zp \} ;\]

\[\text{Assign} \text{r'}.\]

\[\text{In}[4] := \text{sep} = \text{rr} - \text{rp} ;\]

\[\text{Evaluate} |r - r'|.\]

\[\text{In}[5] := \text{ludot}[ \text{sep}, \text{sep} ] ;\]

\[\text{Evaluate} |r - r'|^2.\]

\[\text{In}[6] := \text{Simplify}[ \% ] ;\]

\[\text{Recognize} \cos^2 \phi + \sin^2 \phi = 1.\]

\[\text{In}[7] := \text{den} = \text{Sqrt}[ \% ] ;\]

\[\text{Bind} |r - r'| \text{to} \text{den}.\]

\[\text{In}[8] := \text{cnst} = 1/(4 \cdot \text{Pi} \cdot \text{Epsilon} \cdot 0 ) ;\]

\[\text{Bind} 1/4\pi \epsilon_0 \text{to} \text{cnst}.\]

\[\text{In}[9] := \text{integ} = \{ \lambda \} * \text{cnst} / \text{den} ;\]

\[\text{Set integrand.}\]

\[\text{Out}[9] = \frac{\lambda}{4\pi \sqrt{r^2 + (z - z')^2} \epsilon_0}\]

Then, we find \(V(r, z)\), with \(r = \sqrt{x^2 + y^2}\) by executing the statement

\[\text{The variable I in Mathematica is reserved for} \sqrt{-1}, \text{so we use II for the moment of inertia.}\]

\[\text{The file $\text{HEAD/mathematica/crossdot.m}$ was described in Section 8.11 and exists in the public directory structure. (The symbol $\text{HEAD}$ is defined in the Local Guide.) It defines the two functions lucross[v1, v2] and ludot[v1, v2] for evaluating the cross and dot products of two three-component vectors when those vectors are conveyed by lists.}\]
13.4. Evaluating Integrals Symbolically with Mathematica

Figure 13.4: A line charge lying on the $z$ axis.

\[
\begin{align*}
\text{In[10] := } & \quad V = \text{Integrate[ integ, \{zp, -a, a\}, Assumptions -> \{a>0, z>0, r>0\} ]} \\
\text{Out[10] = } & \quad \frac{\lambda \log \left( \frac{a + z + \sqrt{r^2 + (a + z)^2}}{-a + z + \sqrt{r^2 + (-a + z)^2}} \right)}{4\pi\epsilon_0}
\end{align*}
\]

While correct, this result is certainly not particularly transparent. Let us test it by examining its behavior in the $xy$ plane in the two limits $a/r \ll 1$ and $a/r \gg 1$, i.e., when $r$ is large compared to the length of the line and when $r$ is small compared to the length of the line. For the first limit, we execute the statements

\[
\begin{align*}
\text{In[11] := } & \quad V_0 = V /. \ z -> 0 \\
\text{Out[11] = } & \quad \frac{\lambda \log \left( \frac{a + \sqrt{r^2 + a^2}}{-a + \sqrt{r^2 + a^2}} \right)}{4\pi\epsilon_0}
\end{align*}
\]

Let $a/r \to s$.

\[
\begin{align*}
\text{In[12] := } & \quad V_0 /. \ a -> s*r; \\
\text{In[13] := } & \quad \text{Series[ \%, \{s, 0, 2\} ];} \\
\text{In[14] := } & \quad \text{Simplify[ \%, \ r>0 \ ];} \\
\text{Out[14] = } & \quad \frac{\lambda s}{2\pi\epsilon_0} + O[s]^3
\end{align*}
\]

Convert to regular expression (and simplify).

\[
\begin{align*}
\text{In[15] := } & \quad \text{Normal[\%];} \\
\text{In[16] := } & \quad \% /. \ s -> a/r; \\
\text{In[17] := } & \quad \% /. \ \{\text{\[Lambda]} \} -> q/(2*a) \\
\text{Out[17] = } & \quad \frac{q}{4\pi\epsilon_0 r}
\end{align*}
\]

As expected, the potential at points remote from the charged line varies with distance like that of a point charge. From a point far away from the wire (compared to its length), the wire looks like a point charge.
At the other extreme, when \( r \) is small compared to the length of the line, we might think to invoke a statement like `Series[V0, {r,0,2}]`. That approach, however, is problematic because the denominator under the logarithm in \( V_0 \) is zero at \( r = 0 \) when \( a > 0 \). The expression \( V_0 \) does not have a very well behaved series expansion about the point \( r = 0 \). As an alternative, we might notice that the numerator and denominator under the logarithm in \( V_0 \) are well behaved in the vicinity of \( r = 0 \).

In fact,

\[
\text{In}[18]:= \text{Series}[-a + \sqrt{a^2 + r^2}, \{r,0,2\}, \text{Assumptions->a>0}];
\]

\[
\text{Out}[18]= \frac{r^2}{2a}
\]

\[
\text{In}[19]:= \text{T1 = Normal[Simplify[%, a>0]]}
\]

\[
\text{Out}[19]= \frac{r^2}{2a}
\]

\[
\text{In}[20]:= \text{Series}[a + \sqrt{a^2 + r^2}, \{r,0,2\}, \text{Assumptions->a>0}];
\]

\[
\text{In}[21]:= \text{T2 = Normal[Simplify[%, a>0]]}
\]

\[
\text{Out}[21]= 2a + \frac{r^2}{2a}
\]

Substituting these expansions into \( V_0 \) with the statements

\[
\text{In}[22]:= \text{SmallR = V0 /. {-a+Sqrt[a^2+r^2]->T1, a+Sqrt[a^2+r^2]->T2 }}
\]

\[
\text{Out}[22]= \frac{\lambda}{4\pi\epsilon_0} \text{Log} \left[ \frac{2a(2a + r^2/2a)}{r^2} \right]
\]

we find a much simpler, though approximate expression. Further, we can now see why a well-behaved series expansion does not exist. As \( r \) becomes ever smaller, the denominator in this logarithm approaches zero and the logarithm diverges. The numerator, however, approaches \( 4a^2 \) as \( r \to 0 \). Thus, we can simplify the expression further by ignoring the term \( r^2/2a \) in the numerator. The statement

\[
\text{In}[23]:= \text{SmallR = SmallR /. 2*a+r^2/(2*a) -> 2*a}
\]

\[
\text{Out}[23]= \frac{\lambda}{4\pi\epsilon_0} \text{Log} \left[ \frac{4a^2}{r^2} \right]
\]

effects this further simplification. Indeed, if we recognize the equivalent form,

\[
-\frac{\lambda}{2\pi\epsilon_0} \ln \frac{r}{a} + \frac{\lambda \ln 2}{2\pi\epsilon_0}
\]

we see that our result differs by a constant from the familiar logarithmic result for the potential of a uniformly charged, infinitely long straight wire when the reference point (point of zero potential) is taken at a distance \( a \) from the wire. From a point close to the wire (compared to its length) and near its center, the wire looks to be infinitely long.

### 13.4.5 The Helmholtz Coil

Next, we evaluate the magnetic field produced on the \( z \) axis by a pair of circular current loops of arbitrary separation and demonstrate the significance of the specific separation used in the Helmholtz coil. Strategically, we regard the pair as a superposition of two loops and seek first the magnetic field of a single loop. In general, the magnetic field \( B(r) \) is given by the Biot-Savart law, which is the first
member of Eq. (13.29). If the (single) loop has radius \(a\) and lies in the \(xy\) plane with its center at the origin, then, for an observation point on the \(z\) axis, the various vectors in the expression are given by \(r = sa\hat{k}\), where (to facilitate expressing things in dimensionless form) the \(z\) coordinate is written as a multiple \(s\) of the radius \(a\) of the loop, \(r' = a\cos\phi\hat{i} + a\sin\phi\hat{j}\), and \(dr' = (-a\sin\phi\hat{i} + a\cos\phi\hat{j})\,d\phi\).

Evaluation of the integrand and the integral for a single loop then proceeds with the statements

\[
\text{In[1]:= Get["$HEAD/mathematica/crossdot.m"]};
\]

\[
\text{In[2]:= r = \{0, 0, s*a\};}
\]

\[
\text{In[3]:= rp = \{a*Cos[\text{\[Phi]}], a*Sin[\text{\[Phi]}], 0\};}
\]

\[
\text{In[4]:= drp = \{-a*Sin[\text{\[Phi]}], a*Cos[\text{\[Phi]}], 0\};}
\]

\[
\text{In[5]:= sep = r - rp;}
\]

Define \(\text{lucross}\) and \(\text{ludot}\).

\[
\text{In[6]:= ludot[sep, sep];}
\]

\[
\text{In[7]:= mag2 = Simplify[%, a>0]};
\]

\[
\text{Out[7]= a^2(1 + s^2));
\]

\[
\text{In[8]:= lucross[drp, sep];}
\]

\[
\text{In[9]:= num = Simplify[%, a>0]};
\]

\[
\text{Out[9]= \{a^2sCos[\text{\[Phi]}], a^2sSin[\text{\[Phi]}], a^2\};}
\]

Evaluate \(|\text{r - r'}|^2\).

\[
\text{In[10]:= integ = num*\[Mu]0*i/(4*Pi*mag2^(3/2));}
\]

\[
\text{In[11]:= integ = Simplify[%, a>0]};
\]

\[
\text{Out[11]= \{i\[Mu]0sCos[\text{\[Phi]}]/(4a\pi(1 + s^2)^(3/2)), i\[Mu]0sSin[\text{\[Phi]}]/(4a\pi(1 + s^2)^(3/2)), i\[Mu]0/(4a\pi(1 + s^2)^(3/2))\};}
\]

\[
\text{In[12]:= B1 = Integrate[integ, \{\text{\[Phi]}, 0, 2*Pi\}];}
\]

\[
\text{In[13]:= B1z = B1[[3]];}
\]

\[
\text{In[14]:= B1c = B1z/.s->0;}
\]

\[
\text{In[15]:= B1z = B1z/B1c;}
\]

\[
\text{Out[15]= 1/(1 + s^2)^(3/2)}
\]

Finally, we bring various factors together to construct the integrand, simplify that integrand, evaluate the integral, and recast the result in a dimensionless form with the statements

\[
\text{In[16]:= Bpair = (B1z/.s->s-c) + (B1z/.s->s+c);} \quad \text{Find field of two loops, one at } s = c \text{ and the other at } s = -c.
\]

\[
\text{In[17]:= Bpairc = Bpair/.s->0;} \quad \text{Find field at center.}
\]

\[
\text{In[18]:= Bpair = Bpair/Bpairc;} \quad \text{Normalize field of two loops.}
\]

\[
\text{Out[18]= } 1/2(1 + c^2)^(3/2) \left( \frac{1}{[1 + (-c + s)^2]^{3/2}} + \frac{1}{[1 + (c + s)^2]^{3/2}} \right)
\]

\[6\text{See footnote 5.}\]
Figure 13.5: On-axis magnetic field of a pair of loops. The highest of the three graphs shows the field when the loop separation is $c = 1.0$ (separation of the two loops equal to twice the radius of each loop). The middle graph corresponds to $c = 0.5$ (the Helmholtz separation) and the lower graph corresponds to $c = 0.25$. This graph was created by Mathematica.

This expression is complicated. We explore it further in two ways. First, let us plot a few graphs of this result as a function of $s$ for representative values of $c$ with the statements

```
In[19]:= c100 = Bpair /. c -> 1.0;
In[20]:= c050 = Bpair /. c -> 0.5;
In[21]:= c025 = Bpair /. c -> 0.25;
In[22]:= Plot[ { c100, c050, c025 }, {s, 0.0, 3.0},
    PlotRange -> {0.0, 2.0}, PlotStyle -> Thickness[0.005],
    AxesLabel -> {"s", "B/Bc"} ]
```

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

```
In[23]:= tay = Simplify[ Series[ Bpair, { s, 0, 4 } ] ]
```

```
Out[23]= 1 + \frac{3(-1 + 4c^2)s^2}{2(1 + c^2)^2} + \frac{15(1 - 12c^2 + 8c^4)s^4}{8(1 + c^2)^4} + O[s]^5
```

Clearly there is a particular value of $c$—half the loop separation—at which the coefficient of the $s^2$ term is zero! For that separation, the magnetic field near the center of the loops is especially constant. We find that special value of $c$ with the statements

```
7Remember the structure of the storage of Taylor series as described in Section 8.8.7.
```
13.4. EVALUATING INTEGRALS SYMBOLICALLY WITH MATHEMATICA 273

\[ \text{In}[24]:= \text{Numerator[ Normal[tay][[2]] ]} \]
\[ \text{Out}[24]= 3(-1 + 4c^2)s^2 \]
\[ \text{In}[25]:= \text{Solve[ \% == 0, c ]} \]
\[ \text{Out}[25]= \{ \{ c \to -\frac{1}{2} \}, \{ c \to \frac{1}{2} \} \} \]

The first of these statements extracts the numerator of the second term of the Taylor series and the second finds the value of \( c \) that makes that numerator zero. The two solutions are, of course, equivalent; both reveal that the physical arrangement in which the field has no \( s^2 \) term has one loop positioned one-half of the loop radius above the midplane and the other positioned one-half of the loop radius below the midplane. That is, the loops are separated by their common radius—the Helmholtz separation.

When the loops have that special separation, we determine numerically how the field varies with \( s \) by using the statements

\[ \text{In}[26]:= \text{tay} = \text{Normal[ tay ]} \]
\[ \text{Out}[26]= 1 + 3(-1 + 4c^2)s^2 + \frac{15(1 - 12c^2 + 8c^4)s^4}{8(1 + c^2)^4} \]
\[ \text{In}[27]:= \text{tay} /. c \to 1/2 \]
\[ \text{Out}[27]= 1 - \frac{144}{125}s^4 \]
\[ \text{In}[28]:= \text{tay1} = \% // \text{N} \]
\[ \text{Out}[28]= 1. - 1.152s^4 \]

and we see again that, so long as we don’t stray too far from the center of the arrangement, the magnetic field falls away from its central value as the fourth power of the distance from the center. In particular (see next paragraph for a refinement), we might determine approximately how far we can stray from the center along the axis before the field has fallen to, say, 99% or 95% of its value at the center by executing the statements

\[ \text{In}[29]:= \text{Solve[ tay1 == 0.99, s ]} \]
\[ \text{Out}[29]= \{ \ldots , \{ s \to 0.305237 \} \ldots \} \]
\[ \text{In}[30]:= \text{Solve[ tay1==0.95, s ]} \]
\[ \text{Out}[30]= \{ \ldots , \{ s \to 0.456435 \} \ldots \} \]

Unfortunately, these solutions are only approximate, since we have used a truncated expansion of the field, an expansion that becomes increasingly incorrect as \( s \) increases.

To find more accurately how far from the center we can stray before the field has fallen to 99% or 95% of its value at the center, we need return to the exact expression of \( B_{\text{pair}} \) but restrict it to the Helmholtz case by setting \( c = 1/2 \) with the statement

\[ \text{In}[31]:= \text{Bhelm} = \text{Bpair} /. \text{c} \to 1/2; \]

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 \( \text{Bhelm} - 0.99 = 0 \) and \( \text{Bhelm} - 0.95 = 0 \). The dependence of these expression on \( s \), however, is complicated, and the command \text{Solve} either takes a very long time or fails altogether. Thus, we draw (prematurely—see Section 14.13) on Mathematica’s command \text{FindRoot}, which uses an iterative numerical procedure to refine a supplied initial guess for the root. The above results provide us with sensible initial guesses, so we invoke the statements:\footnote{In \text{FindRoot}, the first argument is the equation whose root is sought and the second is a list whose first element is}

\[ \text{In}[32]:= \text{FindRoot[ \% 99, \{ s, 0.305237 \} ]} \]
\[ \text{Out}[32]= \{ s \to 0.305237 \} \]
\[ \text{In}[33]:= \text{FindRoot[ \% 95, \{ s, 0.456435 \} ]} \]
\[ \text{Out}[33]= \{ s \to 0.456435 \} \]
In[32]:= FindRoot[ Bhelm - 0.99 == 0, { s, 0.3 } ]  
Out[32]= { s → 0.313746 }  
Find s when field is down 1%.

In[33]:= FindRoot[ Bhelm - 0.95 == 0, { s, 0.5 } ]  
Out[33]= { s → 0.488455 }  
Find s when field is down 5%.

Finally, out of curiosity, we ask about the value of the magnetic field at a point in the center of one of the coils by executing the statement

In[34]:= N[ Bhelm /. s -> 1/2 ]  
Out[34]= 0.945824

In 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 \( s^4 \) \( \left( \frac{z}{a} \right)^4 \). As the above results reveal, we would have to move 31.4% of the radius (62.8% of the distance from the center to the plane of either loop) before the field has fallen to 99% of its value at the center, and 48.8% of the radius (97.6% of the distance from the center to the plane of either loop) before the field has fallen to 95% of its value at the center. By the time we reach the plane of either loop, the field has fallen only to 94.6% of its value at the center. This optimal configuration of a pair of current loops is regularly used to produce a nearly uniform field over a large volume of space.

13.4.6 Period of a Pendulum: Correction as Amplitude Grows

The integral in Eq. (13.22) does not have a simple or familiar evaluation in closed form. *Mathematica* does, however, make it easy to examine the way in which the period departs from its limiting value as the amplitude moves away from the very small. We simply determine a Taylor expansion of the integrand with the statements

\[
\text{In[1]:= integ = } \frac{1}{\sqrt{1-k^2 \sin(\phi)^2}} \quad \text{Out[1]= } \frac{1}{\sqrt{1-k^2 \sin(\phi)^2}}
\]

\[
\text{In[2]:= intapprox = Series[ integ, \{ k, 0, 5 \} ] } \quad \text{Out[2]= } 1 + \frac{1}{2} \sin(\phi)^2 k^2 + \frac{3}{8} \sin(\phi)^4 k^4 + O[k]^6
\]

Then we insert the pre-multiplying factor \( 2/\pi \) and integrate the series on \( \phi \) with the statement

\[
\text{In[3]:= Expand[ 2*Integrate[ intapprox, \{ \phi, 0, \pi/2 \} ] / \pi ] } \quad \text{Out[3]= } 1 + \frac{1}{4} k^2 + \frac{9}{64} k^4
\]

Finally, we substitute \( \sin(\theta_0/2) \) for \( k \) with the statement
13.4. EVALUATING INTEGRALS SYMBOLICALLY WITH MATHEMATICA

\(\text{In}[4] := \% / . \ k -> \text{Sin}[\text{\[Theta]}0 / 2]\)
\(\text{Out}[4] = 1 + \frac{1}{4} \text{Sin} \left[ \frac{\theta_0}{2} \right]^2 + \frac{9}{64} \text{Sin} \left[ \frac{\theta_0}{2} \right]^4\)
and expand this result as a power series in \(\theta_0\)
\(\text{In}[5] := \text{Normal} \left[ \text{Series} \left[ \%, \{ \text{\[Theta]}0, 0, 5 \} \right] \right]\)
\(\text{Out}[5] = 1 + \theta_0^2 + \frac{11\theta_0^4}{3072}\)
Note that \(\theta_0\) must here be expressed in radians.

To see how significantly the period departs from the small-amplitude value when \(\theta_0\) is not quite zero, let us evaluate this expression for \(\theta_0 = 5^\circ, 10^\circ, 15^\circ,\) and \(20^\circ\). We use the Mathematica statements
\(\text{In}[6] := \text{deg} = \{ 5, 10, 15, 20 \};\)
\(\text{In}[7] := \text{ampl} = \text{deg} \ast (\text{Pi}/180);\)
\(\text{In}[8] := \text{Out}[4] / . \ \text{\[Theta]}0 -> \text{ampl} \ // \text{N}\)
\(\text{Out}[8] = \{ 1.00048, 1.00191, 1.0043, 1.00767 \}\)
Evidently, at amplitudes of \(5^\circ, 10^\circ, 15^\circ,\) and \(20^\circ,\) the period of the pendulum is approximately 0.05%, .2%, .4%, and .8% larger than the standard small-amplitude approximation given by the expression \(2\pi \sqrt{l/g}\).

13.4.7 Fourier Coefficients for Half-Rectified Signal

Suppose we sought to express the half-rectified sine wave
\[
    f(x) = \begin{cases} 
    0 & -l \leq x < 0 \\
    \sin \frac{\pi x}{l} & 0 \leq x < l
    \end{cases}
\]
in a Fourier series. In accordance with Eq. (13.51), the coefficients would then be given by the integrals
\[
    a_0 = \frac{1}{l} \int_0^l \sin \frac{\pi x}{l} \ dx ; \quad a_n = \frac{1}{l} \int_0^l \sin \frac{\pi x}{l} \cos \frac{n\pi x}{l} \ dx ; \quad b_n = \frac{1}{l} \int_0^l \sin \frac{\pi x}{l} \sin \frac{n\pi x}{l} \ dx
\]
where \(n = 1, 2, 3, \ldots\). We find these coefficients quickly. For \(a_0\), we submit the statement
\(\text{In}[1] := a0 = \text{Integrate} \left[ \text{Sin}[\text{Pi} \ast x/l], \{ x, 0, 1 \} \right]/1\)
\(\text{Out}[1] = \frac{2}{\pi}\)
Then for \(a_n\) when \(n\) is an integer not equal to zero, we use the statements
\(\text{In}[2] := \text{an} = \text{Integrate} \left[ \text{Sin}[\text{Pi} \ast x/l] \ast \text{Cos}[n \ast \text{Pi} \ast x/l], \{ x, 0, 1 \} \right]/1;\)
\(\text{In}[3] := \text{an} = \text{Simplify} \left[ \text{an}, \text{Element}[n, \text{Integers}] \right]\)
\(\text{Out}[3] = \frac{-1 + (-1)^n}{\pi - n^2 \pi}\)
Note, however, that this expression for \(n = 1\) is indeterminate, so we evaluate that case in isolation with the statement
\footnote{Note the distinction between 1 (el) and 1 (one). Here and in the next several statements, the character is el.}
In[4] := a1 = Integrate[ Sin[Pi*x/l]*Cos[Pi*x/l], { x, 0, l } ]/l
Out[4] = 0 

Finally, we find $b_n$ with the statement

In[5] := bn = Integrate[ Sin[Pi*x/l]*Sin[n*Pi*x/l], { x, 0, l } ]/l;
Out[6] = 0 

A close look at the integral giving $b_n$, however, reveals that this result cannot be correct for $b_1$, since the integrand in that case $[\sin^2(\pi x/l)]$ is decidedly positive and an integrand that is everywhere positive cannot yield zero for the integral.$^{10}$ The formal evaluation of the integral (see the output of line In[5] on your screen) is, in fact, indeterminate at $n = \pm 1$. Rather than trying to evaluate that indeterminacy as a limit, we simply redo the integral with the statement

In[7] := b1 = Integrate[ Sin[Pi*x/l]^2, { x, 0, l } ]/l
Out[7] := 1/2

In total, we find that

$$a_0 = \frac{2}{\pi} ; \quad a_1 = 0 ; \quad a_n = -\frac{1 + (-1)^n}{\pi(n^2 - 1)}, n > 1 ; \quad b_1 = \frac{1}{2} ; \quad b_n = 0, n > 1 \quad (13.56)$$

With these values for the Fourier coefficients, we assemble the Fourier series in accordance with Eq. (13.50) to find that

$$f(x) = \frac{1}{\pi} - \sum_{n=2}^{\infty} \frac{1 + (-1)^n}{\pi(n^2 - 1)} \cos \frac{n\pi x}{l} + \frac{1}{2} \sin \frac{\pi x}{l}$$

$$= \frac{1}{\pi} + \frac{1}{2} \sin \frac{\pi x}{l} - \frac{2}{\pi} \sum_{m=1}^{\infty} \frac{1}{4m^2 - 1} \cos \frac{2m\pi x}{l}$$

$$= \frac{1}{\pi} + \frac{1}{2} \sin \frac{\pi x}{l} - \frac{2}{\pi} \left( \frac{1}{3} \cos \frac{2\pi x}{l} + \frac{1}{15} \cos \frac{4\pi x}{l} + \frac{1}{35} \cos \frac{6\pi x}{l} + \cdots \right) \quad (13.57)$$

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

$$f_n(x) = \frac{1}{\pi} + \frac{1}{2} \sin \frac{\pi x}{l} - \frac{2}{\pi} \sum_{n=1}^{m} \frac{1}{4m^2 - 1} \cos \frac{2m\pi x}{l} \quad (13.58)$$

which approaches $f(x)$ as $n \to \infty$. Introducing the variable $\overline{x} = x/l$ (and then dropping the overbar), we then generate graphs for various values of $n$. First, we define a function from which the coefficients $a_m$ can be evaluated for various $m$ with the statement

In[8] := a[m_] := an /. n -> 2*m

$^{10}$Note also that Mathematica gave us no warning that the result displayed was potentially incorrect for $n = \pm 1$. 
13.4. Evaluating Integrals Symbolically with Mathematica

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 $n = 1, 2, 4, \text{ and } 8$, respectively. This graph was created by Mathematica.

Then, we submit the statement

\[
\text{In}[9]:= f[x_,n_]:= 1/\pi + \sin(\pi x)/2 + \sum a[m]\cos(2m\pi x), \{m,1,n\}
\]

to define a function for evaluating the truncated series as a function of $x$ for selected $n$. Finally, we plot graphs of the truncated series for $n = 1, 2, 4, \text{ and } 8$ with the statements

\[
\text{In}[10]:= \text{Plot}[f[x,1],\{x,-3.0,3.0\},\text{PlotRange}\rightarrow\{-0.5,1.5\}, \\
\text{PlotStyle}\rightarrow\text{Thickness}[0.005],\text{AxesLabel}\rightarrow\{"x","f(x,1)"\} ]
\]

\[
\text{In}[11]:= \text{Plot}[f[x,2],\{x,-3.0,3.0\},\text{PlotRange}\rightarrow\{-0.5,1.5\}, \\
\text{PlotStyle}\rightarrow\text{Thickness}[0.005],\text{AxesLabel}\rightarrow\{"x","f(x,2)"\} ]}
\]

\[
\text{In}[12]:= \text{Plot}[f[x,4],\{x,-3.0,3.0\},\text{PlotRange}\rightarrow\{-0.5,1.5\}, \\
\text{PlotStyle}\rightarrow\text{Thickness}[0.005],\text{AxesLabel}\rightarrow\{"x","f(x,4)"\} ]}
\]

\[
\text{In}[13]:= \text{Plot}[f[x,8],\{x,-3.0,3.0\},\text{PlotRange}\rightarrow\{-0.5,1.5\}, \\
\text{PlotStyle}\rightarrow\text{Thickness}[0.005],\text{AxesLabel}\rightarrow\{"x","f(x,8)"\} ]}
\]

\[
\text{In}[14]:= \text{Quit[]}
\]

Shown in Fig. 13.6, the resulting graphs clearly reveal the convergence of the sum to the half-rectified wave as more terms are included.
13.5 Algorithms for Numerical Integration

Unfortunately, very many integrals of great interest have no closed form, analytic evaluation. To address these integrals, numerical analysts have developed many formulae—often called quadrature formulae—for numerical integration. In this section, we describe several of these formulae.

13.5.1 Newton-Cotes Quadrature

One family of quadrature formulae can be deduced by starting with the recognition that the definite integral

\[ A = \int_a^b f(x) \, dx \]  

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, \( x_N = b \), and let \( f(x_0) = f(a) = f_0 \), \( f(x_1) = f_1 \), \ldots, \( f(x_i) = f_i \), \ldots, \( f(x_N) = f(b) = f_N \). To deduce the simplest quadrature formula, we approximate the area of each resulting strip by the area of a rectangle whose height is the value of \( f(x) \) at the left end of the strip [Fig. 13.8(a)]. Thus

\[ \int_a^b f(x) \, dx \approx f_0 \Delta x + f_1 \Delta x + \cdots + f_i \Delta x + \cdots + f_{N-1} \Delta x \]

which turns out to be 100% accurate only if \( f(x) \) happens to be a constant.

If, however, we approximate the area of each strip by the area of a rectangle whose height is the value of \( f(x) \) at the midpoint of the strip, we would deduce the midpoint rule

\[ \int_a^b f(x) \, dx \approx M_N = \left( \frac{1}{2} f_0 + f_1 + \cdots + f_i + \cdots + f_{N-1} \right) \Delta x \]  

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,

\[ \int_a^b f(x) \, dx \approx T_N = \frac{1}{2} (f_0 + f_1) \Delta x + \frac{1}{2} (f_1 + f_2) \Delta x + \cdots + \frac{1}{2} (f_{N-1} + f_N) \Delta x \]

\[ = \left( \frac{1}{2} f_0 + f_1 + 2 f_2 + \cdots + f_i + \cdots + f_{N-1} + \frac{1}{2} f_N \right) \Delta x \]
13.5. ALGORITHMS FOR NUMERICAL INTEGRATION

Figure 13.8: Two different approximations leading to quadrature formulae. The left figure corresponds to Eq. (13.60), the right figure to Eq. (13.62).

\[
\int_a^b f(x) \, dx \approx \frac{1}{3} \left( f_0 + 4 f_1 + 2 f_2 + 4 f_3 + \cdots + 2 f_{N-2} + 4 f_{N-1} + f_N \right) \Delta x (13.63)
\]

which 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

\[
T_1 = T_{2^0} = \frac{f(a) + f(b)}{2} (b - a) (13.64)
\]

Then, halving the step size and introducing \( x_1 = a + (b - a)/2 \), the midpoint of the interval (see Fig. 13.9, in which—for present convenience—we label the points differently than we did in Fig. 13.7), we find that

\[
T_2 = T_{2^1} = \left[ \frac{f(a)}{2} + f(x_1) + \frac{f(b)}{2} \right] \frac{b - a}{2} = \frac{1}{2} T_1 + f(x_1) \frac{(b - a)}{2} = \frac{1}{2} T_1 + f \left( a + \frac{b - a}{2} \right) \frac{b - a}{2} (13.65)
\]
Figure 13.9: A rearrangement of the labels for dividing points.

| a | x_4 | x_2 | x_5 | x_1 | x_6 | x_3 | x_7 | b |

Halving the step size again and introducing \( x_2 = a + (b - a)/4 \) and \( x_3 = a + 3(b - a)/4 \) (again, see Fig. 13.9), we find that

\[
T_4 = T_2^2 = \left[ \frac{f(a)}{2} + f(x_2) + f(x_1) + f(x_3) + \frac{f(b)}{2} \right] \frac{(b-a)}{4}
\]

\[
= \frac{1}{2} T_2 + \left[ f(x_2) + f(x_3) \right] \frac{(b-a)}{4}
\]

\[
= \frac{1}{2} T_2 + \left[ f \left( a + \frac{b-a}{4} \right) + f \left( a + 3 \frac{b-a}{4} \right) \right] \frac{(b-a)}{4}
\]  

(13.66)

Yet again, introducing

\[
x_4 = a + \frac{1}{8} (b-a) \quad x_5 = a + \frac{3}{8} (b-a)
\]

\[
x_6 = a + \frac{5}{8} (b-a) \quad x_7 = a + \frac{7}{8} (b-a)
\]  

(13.67)

and continuing one more step, we find that

\[
T_8 = T_2^2 = \frac{1}{2} T_4 + \left[ f(x_4) + f(x_5) + f(x_6) + f(x_7) \right] \frac{(b-a)}{8}
\]

\[
= \frac{1}{2} T_4 + \left[ f \left( a + \frac{b-a}{8} \right) + f \left( a + 3 \frac{b-a}{8} \right) + f \left( a + 5 \frac{b-a}{8} \right) + f \left( a + 7 \frac{b-a}{8} \right) \right] \frac{(b-a)}{8}
\]

(13.68)

and, in general, with

\[
x_{2N+j} = a + \frac{(2j+1)(b-a)}{2^{N+1}} \quad ; \quad 0 \leq j < 2^{N+1}
\]  

(13.69)

that

\[
T_{2^{N+1}} = \frac{1}{2} T_{2^N} + \left[ f \left( a + \frac{b-a}{2^{N+1}} \right) + f \left( a + 3 \frac{b-a}{2^{N+1}} \right) + \ldots 
\]

+ \left[ f \left( a + (2^{N-1} - 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

\[
T_4 = \left[ \frac{1}{2} f(a) + f(x_2) + f(x_1) + f(x_3) + \frac{1}{2} f(b) \right] \frac{b-a}{4}
\]

\[
= \left[ f(a) + 2 f(x_2) + 2 f(x_1) + 2 f(x_3) + f(b) \right] \frac{b-a}{8}
\]  

(13.71)
and that

\[ T_8 = \left[ \frac{1}{2} f(a) + f(x_4) + f(x_2) + f(x_5) \right. \]
\[ + f(x_1) + f(x_6) + f(x_3) + f(x_7) + \frac{1}{2} f(b) \] \( b-a \)

(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

\[ \frac{4T_8 - T_4}{3} = \left[ f(a) + 4f(x_4) + 2f(x_2) + 4f(x_5) + 2f(x_1) \right. \]
\[ + 4f(x_6) + 2f(x_3) + 4f(x_7) + f(b) \left( \Delta x \right) \] \( (13.74) \)

which—mira le dictu—we recognize as Simpson’s rule for evaluating the integral with 8 divisions of the interval! We conclude that

\[ S_8 = \frac{4T_8 - T_4}{3} \] \( (13.75) \)

More generally, we could also conclude that

\[ S_{2n} = \frac{4T_{2n} - T_n}{3} \] \( (13.76) \)

The extrapolation formula of Eq. (13.76) applied to two successive evaluations by the trapezoidal rule gives the result of evaluation by Simpson’s rule!

A subroutine for trapezoidal integration can thus be used in a 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$^3$. 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).

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,

$$\int_a^b f(x) \, dx - M_N = \frac{(b-a)^3}{24N^2} \left| \frac{d^2 f}{dx^2} \right|_{x=\xi}$$  \hspace{1cm} (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

$$\int_a^b f(x) \, dx - T_N = \frac{(b-a)^3}{12N^2} \left| \frac{d^2 f}{dx^2} \right|_{x=\xi}$$  \hspace{1cm} (13.78)

and

$$\int_a^b f(x) \, dx - S_N = \frac{(b-a)^5}{180N^4} \left| \frac{d^4 f}{dx^4} \right|_{x=\xi}$$  \hspace{1cm} (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.

11We shall make this criterion a bit more explicit in the next paragraphs.

12We shall see in later sections how we might decide when roundoff has started to be significant.

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

$$
\left| \int_a^b f(x) \, dx - M_N \right| \leq \frac{(b-a)^3}{24N^2} D_{\text{max}}(2) \quad (13.80)
$$
$$
\left| \int_a^b f(x) \, dx - T_N \right| \leq \frac{(b-a)^3}{12N^2} D_{\text{max}}(2) \quad (13.81)
$$
$$
\left| \int_a^b f(x) \, dx - S_N \right| \leq \frac{(b-a)^5}{180N^4} D_{\text{max}}(4) \quad (13.82)
$$

(though we must keep in mind that these approximations will frequently be extremely crude, so these upper bounds may well be very conservative). Provided that problems with computer roundoff do not begin to appear, we conclude from these results that an upper bound on the error in the midpoint and trapezoidal rules falls off like $1/N^2$ while that bound in Simpson's rule falls off like $1/N^4$. Doubling $N$ therefore reduces the error in the midpoint and trapezoidal rules by a factor of four, while doubling $N$ reduces the error in Simpson's rule by a factor of sixteen. With Simpson's rule, every doubling of $N$ should gain at least one more decimal point in accuracy, so the convergence criterion described in the first paragraph in this subsection is particularly apt when Simpson's rule is used.

13.5.4 Iterative and Adaptive Algorithms

In the previous subsections, we assumed that the user of a particular algorithm would actually view the value obtained for a succession of values of $N$ and decide personally when to stop by examining the changes that occur as $N$ is successively doubled. We can, of course, program a computer to make those decisions. One extremely common approach exploits Simpson's rule (probably via the trapezoidal rule and Romberg integration) to obtain $S_2, S_4, S_8, \ldots$, compares each new value with its predecessor and stops when the absolute value of the difference is smaller than a tolerance—either absolute or relative—prescribed in advance. As a guard against an infinite loop, these algorithms should also stop if the desired tolerance has not been achieved in some maximum number of refinements and should print a warning when the desired tolerance has not been achieved. This method is said to be iterative, because it generates a succession of results, examines each new result in turn, and repeats the process until the new result meets or exceeds the prescribed tolerance. The points at which the function is evaluated, however, are determined ahead of time and are not influenced at all by the nature of the specific integrand to which the algorithm is applied.

Another family of algorithms (which may be iterative or noniterative) aims to minimize computational labor by estimating—though the methods for doing so are often crude—the accuracy obtained with each strip as the evaluation unfolds and shrinking or enlarging that strip to achieve a particular tolerance before going on to the next strip. In these adaptive methods, the points at which the function is evaluated are adjusted in response to the particular function being integrated. Because the assessment of accuracy at a particular strip can result either in shrinking or enlarging the width of that—or the next—strip, adaptive methods focus the computational effort in regions where the function varies rapidly and give less attention to regions in which the function varies slowly.

13.5.5 Gaussian Quadrature

The approach of Gaussian quadrature to numerical integration is more complicated than the Newton-Cotes approach but significantly better in some respects. In the Gaussian approach, both the points
at which the function is to be evaluated and the weights to be applied to each value are adjusted to achieve maximum accuracy when the function is approximated by a polynomial of a given order.

The development of a formula for Gaussian quadrature is simplified if we begin by introducing a set of \( m + 1 \) points \( t_i, \ (i = 0, 1, 2, \ldots, m) \) that divide the interval \( t_0 = a \leq t \leq b = t_m \) into \( m \) segments, the \( i \)-th of which extends over the interval \( t_{i-1} \leq t \leq t_i \). The values \( t_i \) may—but need not—be equally spaced. In this notation, we write the integral of interest as a sum of integrals over each segment, i.e., we write

\[
\int_a^b g(t) \, dt = \sum_{i=1}^m \int_{t_i}^{t_{i+1}} g(t) \, dt \tag{13.83}
\]

To facilitate the discussion, however, we rescale and translate the variable in the \( i \)-th segment by introducing the variable \( x \) defined by

\[
x = \frac{2t - (t_{i+1} + t_i)}{t_{i+1} - t_i} \quad \text{or} \quad t = \frac{t_{i+1} + t_i}{2} + \frac{t_{i+1} - t_i}{2} x = t_i^{\text{mid}} + \frac{\Delta t_i}{2} x \tag{13.84}
\]

where \( t_i^{\text{mid}} \) is the coordinate at the midpoint of the \( i \)-th segment and \( \Delta t_i \) is the width of the \( i \)-th segment. With this change, \( x \) ranges from \(-1\) to \(+1\) as \( t \) ranges from \( t_i \) to \( t_{i+1} \), so the integrals of interest now assume the form

\[
\int_a^b g(t) \, dt = \sum_{i=1}^m \int_{t_i}^{t_{i+1}} g(t) \, dt = \sum_{i=1}^m \Delta t_i \int_{-1}^{1} g \left( t_i^{\text{mid}} + \frac{\Delta t_i}{2} x \right) \, dx = \sum_{i=1}^m \Delta t_i \int_{-1}^{1} f_i(x) \, dx \tag{13.85}
\]

where \( f_i(x) = g(t_i^{\text{mid}} + \Delta t_i x/2) \). In essence, then, we must evaluate an integral of the form

\[
\int_{-1}^{1} f(x) \, dx \tag{13.86}
\]

where we omit the subscript \( i \) on \( f \) for the sake of a simpler notation. If we can find a useful numerical evaluation for the integral in this standard form, then all else can be obtained by appropriate translations and rescalings.

The strategy for Gaussian integration now involves selecting the number of points—say \( N \)—at which the function is to be evaluated in the interval \(-1 < x < 1\), assuming an approximate formula of the form

\[
\int_{-1}^{1} f(x) \, dx = \sum_{k=1}^{N} w_k f(x_k) \tag{13.87}
\]

and then choosing both the weights \( w_k \) and the points of evaluation \( x_k \) to make this expression 100% accurate for a polynomial of as high an order as possible. Since we have \( 2N \) parameters to be determined, we should be able to make this expression accurate for a polynomial of order \( 2N - 1 \) with only \( N \) evaluations of the integrand.

To illustrate Gaussian integration more explicitly, let us derive a two-point formula, for which Eq. (13.87) would assume the more explicit form

\[
\int_{-1}^{1} f(x) \, dx = w_1 f(x_1) + w_2 f(x_2) \tag{13.88}
\]

We choose \( w_1, w_2, x_1 \) and \( x_2 \) so that the formula gives the correct answer for the special cases

\[
\begin{align*}
  f(x) &= 1, \quad \text{yielding that} \quad \int_{-1}^{1} dx = 2 = w_1 + w_2 \tag{13.89} \\
  f(x) &= x, \quad \text{yielding that} \quad \int_{-1}^{1} x \, dx = 0 = w_1 x_1 + w_2 x_2 \tag{13.90} \\
  f(x) &= x^2, \quad \text{yielding that} \quad \int_{-1}^{1} x^2 \, dx = \frac{2}{3} = w_1 x_1^2 + w_2 x_2^2 \tag{13.91} \\
  f(x) &= x^3, \quad \text{yielding that} \quad \int_{-1}^{1} x^3 \, dx = 0 = w_1 x_1^3 + w_2 x_2^3 \tag{13.92}
\end{align*}
\]
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

\[ w_1 x_1 = -w_2 x_2 \quad \text{and} \quad w_1 x_1^3 = -w_2 x_2^3 \]  

which, when we divide the second by the first, yields \( x_1^2 = x_2^2 \), implying that \( x_1 = -x_2 \). (We reject the plus sign so the two values will be distinct.) Then Eqs. (13.91) and (13.89) yield that

\[ 2/3 = (w_1 + w_2) x_2^2 = 2x_2^2 \quad \implies \quad x_2^2 = \frac{1}{3} \]  

from which we also conclude that \( x_1^2 = \frac{1}{3} \). Thus,

\[ x_1 = -\frac{1}{\sqrt{3}} \quad \text{and} \quad x_2 = \frac{1}{\sqrt{3}} \]  

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

\[ \int_{-1}^{1} f(x) \, dx = f \left( -\frac{1}{\sqrt{3}} \right) + f \left( \frac{1}{\sqrt{3}} \right) \]  

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.\(^\text{14}\) Returning to the original function and variables as laid out in Eq. (13.85), we finally find that Eq. (13.96) supports the expression

\[ \int_{a}^{b} g(t) \, dt = \sum_{i=1}^{m} \frac{\Delta t_i}{2} \left[ g \left( t_{i,\text{mid}} - \frac{\Delta t_i}{2} \frac{1}{\sqrt{3}} \right) + g \left( t_{i,\text{mid}} + \frac{\Delta t_i}{2} \frac{1}{\sqrt{3}} \right) \right] \]

\[ = \sum_{i=1}^{m} \frac{\Delta t_i}{2} \sum_{j=1}^{2} w_j \, g \left( t_{i,\text{mid}} + \frac{\Delta t_i}{2} \, x_j \right) \]  

The strategy invoked to develop the two-point Gaussian formula can also be applied to deduce higher order formulae. For a three-point formula, for example, we would have three points and three weights, and we would expect to be able to choose these unknowns to generate a formula that would be exact for a fifth-degree polynomial. The five-point formula

\[ \int_{-1}^{1} f(x) \, dx = 0.23692689 f(-0.90617985) + 0.47862867 f(-0.53846931) \]

\[ + 0.56888889 f(0.00000000) \]

\[ + 0.47862867 f(0.53846931) + 0.23692689 f(0.90617985) \]  

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

\[ w_1 = 0.23692689 \quad x_1 = -0.90617985 \]

\[ w_2 = 0.47862867 \quad x_2 = -0.53846931 \]

\[ w_3 = 0.56888889 \quad x_3 = +0.00000000 \]

\[ w_4 = 0.47862867 \quad x_4 = +0.53846931 \]

\[ w_5 = 0.23692689 \quad x_5 = +0.90617985 \]  

\(^\text{14}\)The advantage isn’t that great, however, because, for Simpson’s rule, the upper evaluation for one strip could also be used as the lower evaluation for the next strip. No such feature applies to the two-point—or to any—Gaussian integration formula. The advantage of the Gaussian approach increases, however, as \( N \) in Eq. (13.87) increases.
Further, returning to the original variable, we note that
\[ \int_a^b g(t) \, dt = \sum_{i=1}^m \frac{\Delta t_i}{2} \sum_{j=1}^5 w_j g \left( t_{\text{mid}}^j + \frac{\Delta t_i}{2} x_j \right) \] (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 \( (1/((1 - x_i^2)(dL_2(x_i)/dx)^2)) \). More generally, for an \( N \)-point Gaussian integration, we would discover that
\[ L_N(x_i) = 0 \quad \text{and} \quad w_i = \frac{2}{(1 - x_i^2)(dL_N(x_i)/dx)^2} \] (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)
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.\(^\text{16}\)

### 13.6 Evaluating Integrals Numerically with IDL

*Note:* All IDL program (.pro) files referred to in this chapter are available in the directory \$HEAD/idl\, 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 IDL’s default search path. If so, the files can be found by IDL 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 IDL’s elementary commands or, more simply, by invoking one of the built-in IDL routines \qsimp\ and \qromb\ or the Lawrence-provided routine \luqsimp. We mention that IDL also provides the routines \int_2d\ and \int_3d\ for evaluating two- and three-dimensional integrals and the routine \int_tabulated\ for integrating functions whose values are supplied in a table (as might arise with experimental data). We shall here illustrate only \luqsimp, but we invite you to study the IDL manuals for information on the other available routines.

#### 13.6.1 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
\[ I = \text{erf}(1) = \frac{2}{\sqrt{\pi}} \int_0^1 e^{-x^2} \, dx \] (13.103)
\(^\text{16}\)A more general integral that can be approached with the techniques of this subsection has the form \( \int_a^b f(x) w(x) \, dx \), where \( w(x) \) is a weighting function. In the case we dealt with, \( w(x) = 1 \) and, in our rescaling, the interval became the interval from \(-1\) to \( 1\). That the Legendre polynomials \( L_n(x) \) 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.
Table 13.1: Values of \( \text{erf}(1.0) \) obtained by the trapezoidal rule and a user-constructed program in IDL.

<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>0.771743</td>
<td>32</td>
<td>0.842633</td>
<td>1024</td>
<td>0.842701</td>
</tr>
<tr>
<td>2</td>
<td>0.825263</td>
<td>64</td>
<td>0.842684</td>
<td>2048</td>
<td>0.842700</td>
</tr>
<tr>
<td>4</td>
<td>0.838368</td>
<td>128</td>
<td>0.842696</td>
<td>4096</td>
<td>0.842701</td>
</tr>
<tr>
<td>8</td>
<td>0.841619</td>
<td>256</td>
<td>0.842700</td>
<td>8192</td>
<td>0.842701</td>
</tr>
<tr>
<td>16</td>
<td>0.842430</td>
<td>512</td>
<td>0.842700</td>
<td>16384</td>
<td>0.842701</td>
</tr>
</tbody>
</table>

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

```
IDL> n = 4  
IDL> x = findgen(n+1)/float(n)  
IDL> dx = 1.0/float(n)  
IDL> f = 2.0*exp(-x^2)/sqrt(!pi)  
IDL> y = 0.5*f[0]  
IDL> for i = 1,n-1 do y=y+f[i]  
IDL> y = y + 0.5*f[n]  
IDL> print, dx*y
```

To illustrate the reduction of truncation error with decreasing step size, we re-execute this procedure, starting with \( n = 1 \) and successively doubling \( n \), finding the values shown in Table 13.1. (The process comes to an end at \( n = 16384 \) because, for \( n = 32768 \), the particular coding we have adopted requires us to have more than 32767 (= \( 2^{16} - 1 \)) elements in the vectors \( x \) and \( f \) and the 16-bit integer used for the index in those arrays cannot store positive integers larger than 32767.) On the basis of this succession of numbers, however, we conclude that, to six decimal places, the value of the integral is somewhere between 0.842700 and 0.842701. The trapezoidal rule took something like 257 or 513 evaluations of the function to reach that value.

Note in this example that the result for \( n = 1 \) is quite inaccurate. As \( n \) increases, however, the result becomes more and more accurate. After a time (here at about \( n = 256 \)) the result stabilizes and further increase of \( n \) makes almost no difference in the value. 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 \( n = 16384 \). Were we able 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, from single to double 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.

---

17 Alternatively, as long as \( n \geq 2 \), the fifth, sixth, and seventh statements in this code can be replaced with the single statement `y = 0.5*f[0] + total(f[1:n-1]) + 0.5*f[n]` — a statement that will execute more quickly.

18 The fluctuations in the last digit are almost certainly a reflection of internal computer roundoff, which—in single precision arithmetic—can easily affect the sixth or seventh “significant” digit.
We illustrate with three alternative and more efficient approaches. First, note that Richardson extrapolation applied to $T_1$ and $T_2$ and then again to $T_2$ and $T_4$ yields the values
\[ S_2 = \frac{4 \cdot T_2 - T_1}{3} = \frac{4 \cdot 0.825263 - 0.771743}{3} = 0.843103 \] (13.104)
and
\[ S_4 = \frac{4 \cdot T_4 - T_2}{3} = \frac{4 \cdot 0.838368 - 0.825263}{3} = 0.842736 \] (13.105)
Continuing this process of Richardson extrapolation, we find the values $S_8 = 0.842703$, $S_{32} = 0.842701$, and $S_{64} = 0.842701$. Because $S_{32}$ and $S_{64}$ 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 33 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 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

```
IDL> n = 2
IDL> x = findgen(n+1)/float(n)
IDL> dx = 1.0/float(n)
IDL> f = 2.0*exp(-x^2)/sqrt(!pi)
IDL> w = fltarr(n+1) + 2.0
IDL> w[0] = 1.0 & w[n] = 1.0
IDL> for i=1,n-1,2 do w[i]=4.0
IDL> y = 0.0
IDL> for i = 0,n do y=y+w[i]*f[i]
IDL> print, dx*y/3.0
0.843103
```

If this process is repeated with $n = 4$, the end result is 0.842736. With $n = 32$ and $n = 64$, this process yields the values 0.842701 and 0.842701, 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 IDL statements

```
IDL> pts = [-1.0/sqrt(3.0), 1/sqrt(3.0)]
IDL> w = [1.0, 1.0]
IDL> a = 0.0 & b = 1.0
```

19Alternatively, 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 $y = w \# \text{transpose}(f)$.

20Remember that indices in IDL start at 0. Thus, values pertaining to the first segment will be identified by index 0, those pertaining to the second segment by 1, \ldots, and those pertaining to the $m$-th segment by $m - 1$. to set the evaluation points, weights, and limits. Then we would execute the statements

```
IDL> m = 4
IDL> dx = (b-a)/float(m)
IDL> x = dx * findgen(m+1) + a
IDL> xmid = fltarr(m)
IDL> for i=0,m-1 do xmid[i]=(x[i]+x[i+1])/2.0
```

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

IDL> int = 0.0
IDL> for i=0,m-1 do begin $
IDL> fcts = 0.0 &$
IDL> for j = 0,1 do fcts = fcts + w[j]*exp(-(xmid[i]+dx*pts[j]/2.0)^2) &$
IDL> int = int + fcts & endfor
IDL> int = dx*int/sqrt(!pi)
IDL> print,int
0.842699

finding that the two-point Gaussian formula with four 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 0.842442, 0.842677, 0.842699, and 0.842701, 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. We begin by invoking the statements

IDL> pts = [-0.90617985, -0.53846931, 0.0, 0.53846931, 0.90617985]
IDL> w = [0.23692689, 0.47862867, 0.56888889, 0.47862867, 0.23692689]
IDL> a = 0.0 & b = 1.0

IDL> m = 4
IDL> dx = (b-a)/float(m)
IDL> x = dx * findgen(m+1) + a
IDL> xmid = fltarr(m)
IDL> for i=0,m-1 do xmid[i]=(x[i]+x[i+1])/2.0

Set number of segments.
Set width of each segment.
Set values dividing segments.
Create array for mid points.
Evaluate midpoints.

to set the evaluation points \(x_1, \ldots, x_5\), the weights \(w_1, \ldots, w_5\), and the limits. Then, we would execute the statements

IDL> int = 0.0
IDL> for i=0,m-1 do begin $
IDL> fcts = 0.0 &$
IDL> for j = 0,4 do fcts = fcts + w[j]*exp(-(xmid[i]+dx*pts[j]/2.0)^2) &$
IDL> int = int + fcts & endfor
IDL> int = dx*int/sqrt(!pi)
IDL> print,int
0.842701

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 same result—0.842701—with \(m = 1\), i.e., when the entire interval of integration is treated as a single segment!
CHAPTER 13. EVALUATING INTEGRALS

Table 13.2: The IDL function \texttt{qausint.pro}.


def function gausint, t
;
; \texttt{GAUSINT} defines the integrand for evaluating the error function.
;
tmp1 = 2.0/sqrt( !Pi )
return, tmp1*exp(-t^2)
end

13.6.2 The Function \texttt{luqsimp}

For our purposes, the most suitable IDL routine for the numerical evaluation of one-dimensional integrals is a Lawrence-written adaptation \texttt{luqsimp} of the routine \texttt{qsimpson} written and contributed to the IDL Astronomy User’s Library by W. Landsman.\footnote{The routine \texttt{luqsimp.pro} and the routine \texttt{trapzd.pro}, which—renamed \texttt{lutrapzd.pro} in the present context—is called by \texttt{luqsimp.pro}, are used here and distributed with this book by permission of W. Landsman and of Research Systems, Inc. (RSI). Dr. Landsman wrote the original code, and employees of RSI contributed a small modification of Landsman’s \texttt{qsimpson.pro} before the Lawrence modifications were added. Dr. Landsman’s source code can be retrieved from the IDL Astronomy User’s Library by accessing the URL \url{http://idlastro.gsfc.nasa.gov/homepage.html}. Full contact information for RSI can be found in Section Z.} It is based on the routine \texttt{qsimp} described in \textit{Numerical Recipes} by William H. Press, \textit{et al.} (Cambridge University Press, Cambridge, MA, 1988) and is written in the IDL language. The \texttt{lu} modification is cast as a function to facilitate its use with all three types of integral as identified at the very beginning of this chapter. The routine \texttt{luqsimp} uses both of the strategies described in Section 13.5.2 to achieve efficiencies in execution. Before invoking \texttt{luqsimp}, the user must create a function pro-file defining the integrand as a function of a single scalar argument (the integration variable). The construction of such files is described in Section 2.8. For the integral in Eq. (13.41), for example, the function pro-file might be as shown in Table 13.2.

Having stored the pro-file in the default directory with a name identical to that of the function name and with the file extension \texttt{.pro} (here \texttt{gausint.pro}), we would evaluate the integral and display its value with the statements

\begin{verbatim}
q = luqsimp( 'function_name', a, b, eps=value, max_iter=value, tol=value );
print, q
\end{verbatim}

where \texttt{function_name}, which is mandatory, identifies the file in which the integrand is defined; \texttt{a} and \texttt{b}—also mandatory—are the lower and upper limits of integration; the optional keyword \texttt{eps} (default $10^{-6}$) specifies the desired fractional accuracy; the optional keyword \texttt{max_iter} (default 20) determines the maximum number of evaluations of the function by the formula $2^{\text{max}_{\text{iter}}}$; and the optional keyword \texttt{tol} (default 0.0) specifies the absolute accuracy. The routine \texttt{luqsimp} returns the value of the integral in the specified variable, here \texttt{q}. Finally, the \texttt{print} procedure displays the value returned by \texttt{luqsimp}.

For the example of Eq. (13.41), the integral is evaluated—once with the fractional accuracy set to $10^{-3}$ and again with the default fractional accuracy ($10^{-6}$)—and displayed by the statements

\begin{verbatim}
q = luqsimp( 'function_name', a, b, eps=1e-3, max_iter=value, tol=value );
print, q
\end{verbatim}
Table 13.3: The IDL function `moment.pro`.

function moment, lambda

; MOMENT defines the integrand for evaluating the moment of
; inertia of a semicircular plate.

; tmp = lambda^2
return, 4.0*tmp*sqrt(1.0-tmp)/!pi
end

IDL> q = luqsimp( 'gausint', 0.0, 1.0, eps = 0.001 )
0.84273603
IDL> print, q
0.84270076
IDL> q = luqsimp( 'gausint', 0.0, 1.0 )
0.84270076

Since these values agree to four digits, we conclude that both (0.8427) are accurate to four digits. The second is surely a good deal more accurate, but we would have to rely on attainment of the default value `eps` = $10^{-6}$ before asserting—correctly—that the value of the integral is 0.842701 to six digits.

IDL also has several built-in routines for evaluating one-dimensional integrals, including `qsimp` and `qromb`, which use Simpson’s rule and Romberg integration, respectively, and are based on similarly named functions described in *Numerical Recipes* by William H. Press, *et al.* (Cambridge University Press, Cambridge, MA, 1988). Because these routines do not afford the option of specifying the error as an absolute error, they will sometimes have difficulties with integrals whose value is, in fact, zero. For that reason, we have elected to focus this section on `luqsimp`. The reader is referred to the IDL manuals for information about `qsimp` and `qromb`.

### 13.6.3 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 pro-file listed in Table 13.3 and store it in a file named `moment.pro`. Then, the simple statements

IDL> I = luqsimp( 'moment', 0.0, 1.0, eps=0.001 )
0.24989874
IDL> print, I
0.24999992

evaluate the integral to one part in $10^3$ and $10^6$, respectively, and display the results. These results are clearly in agreement with one another and with those obtained by other methods in this chapter.

### 13.6.4 Quantum Probabilities

In Section 13.6.2, we have already evaluated the integral that appears in the determination of the probability that a quantum harmonic oscillator in its ground state will be found outside the classical
turning point. According to Eq. (13.41), that probability is given by the expression

\[ P(|x| > |x_{\text{turn}}|) = 1.0 - \text{erf}(1.0) = 1.0 - 0.842701 = 0.157299 \]  

We conclude that, in a bit over 15% of the measurements, the quantum oscillator will be found in the classically forbidden region!

### 13.6.5 Integrals as Functions of the Upper Limit

To evaluate an integral as a function of a parameter, \texttt{luqsimp} 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

\[ g(x) = \int_{0}^{x} f(t) \, dt \]  

We begin by constructing a function pro-file to define the integrand \( f(t) \), say \texttt{integ.pro}. Then we invoke the IDL statements

\begin{verbatim}
  x = findgen ... Set desired values for upper limit.
  q = size(x) & N = q[1] - 1 Set N to highest index in x.
  g = fltarr(N+1) Prepare array for values of integral.
  for i = 0, N do $
    g[i] = luqsimp( 'integ', 0, x[i] ) Evaluate integral for each upper limit in x,
    accepting defaults for keywords.
\end{verbatim}

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 \texttt{plot,x,g} will generate a graph of \( g \) versus \( x \).

### 13.6.6 The Error Function

To evaluate the error function as given by Eq. (13.26), we begin by constructing the pro-file giving the integrand. The function \texttt{gausint.pro} defined in Section 13.6.2 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 IDL statements

\begin{verbatim}
  x = findgen(31)/10.0 Specify values of upper limit.
  q = size(x) & N = q[1] - 1 Set N to highest index in x.
  y = fltarr(N+1) Prepare array for values of integral.
  for i = 0, N do $
    y[i] = luqsimp( 'gausint', 0, x[i] ) Evaluate integral for each upper limit in x,
      accepting defaults for keywords.
  plot, x,y, title = 'Error Function', thick=3.0, charsize=1.5, 
      xtitle= 'x', ytitle='erf(x)', ticklen=1.0
\end{verbatim}

The resulting graph is shown in Fig. 13.10.

\footnote{This sequence of statements actually is computationally inefficient. We might increase the efficiency by recognizing, for example, that \( g(x + \Delta x) = g(x) + \int_{x}^{x+\Delta x} f(t) \, dt \) and obtain integrals for larger \( x \) by adding an appropriate increment to already evaluated integrals for smaller \( x \). For our present purposes, that approach unnecessarily complicates the algorithm of evaluation.}

\footnote{The warning message declaring that the “sum did not converge after 20 steps” can be ignored. It comes because the very first integral of the 101 integrals evaluated has the value zero, and a convergence criterion based on fractional error is not compatible with an integral whose value is zero. The message can be suppressed by specifying an absolute tolerance, setting \texttt{eps=0.0} and \texttt{tol = 0.00001}, say. That step is, however, not necessary because, error message to the contrary notwithstanding, the evaluation has, in fact, converged.}
13.6.7 The Cornu Spiral

To determine the Cornu spiral, we use `qugsimp` 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 two pro-files listed in Table 13.4 to provide the two integrands. Rather than storing these definitions in the user’s default directory in the separate files `spiralc.pro` and `spirals.pro`, we elect to place both definitions in the single file `cornu.pro`.\(^{24}\)

For definiteness, we elect to explore the spiral over the range \(-5 \leq u \leq 5\), though we evaluate only

\(^{24}\)With this structure, however, automatic compilation will not occur when each function is first named, so the statement `.compile cornu` must be explicitly executed before either function can be invoked.
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 compile the functions \texttt{spiralc} and \texttt{spirals} and prepare necessary arrays with the statements

```idl
IDL> .compile cornu
IDL> u = findgen(101)/20.0
IDL> q = size(u) & N = q[1] - 1
IDL> CP = fltarr(N+1)
IDL> SP = CP & CN = CP & SN = CP
```

Then we evaluate the integrals with the single (composite) statement

```idl
IDL> for i = 0, N do begin $%
IDL> CP[i]=luqsimp( 'spiralc', 0, u[i] ) & $%
IDL> SP[i]=luqsimp( 'spirals', 0, u[i] ) & $%
IDL> CN[N - i] = -CP[i] & $%
IDL> SN[N - i] = -SP[i] & $%
IDL> end
```

Finally, we assemble the variables to be plotted and plot the graph with the statements

```idl
IDL> C = [CN, CP]
IDL> S = [SN, SP]
IDL> window, 1, xsize=400, ysize=430
IDL> !x.margin = [ 8.0, 2.0 ]
IDL> !y.margin = [ 5.0, 5.0 ]
IDL> plot, C, S, title='Cornu Spiral', ticklen=1.0, charsize=1.25, $%
IDL> xtitle='C(u)', ytitle='S(u)', thick=3.0
```

The resulting graph is shown in Fig. 13.11.\textsuperscript{25,26}

\textsuperscript{25}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.

\textsuperscript{26}See Section 2.16.5 for a description of the strategy invoked to produce equal scaling on both axes.
13.6. EVALUATING INTEGRALS NUMERICALLY WITH IDL

Table 13.5: Structure of IDL function file with parameter in the integrand.

```idl
function function_name, x
; ; Explanatory comments.
; common param, p1, p2, p3
... Statements to evaluate integrand, using p1, p2, and p3 for the parameters.
return, integrand
end
```

13.6.8 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 the integration routine `luqsimp` does not have an argument for specifying a value for a parameter to be passed to the user-defined function pro-file that returns values of the integrand to `luqsimp`. Instead, we make use of `named common`, which is a programming construct built to facilitate assigning values to variables in one routine and passing those values to another routine without the values necessarily being known to any intervening routines. Thus, we can use named common to give a value to a parameter at the main command level and pass that value to the function defining the integrand without passing it through `luqsimp`. To invoke this feature, we must do two things:

1. We add one line to the function pro-file that defines the integrand, so that the file has the format shown in Table 13.5. Here, `common` is a mandatory IDL reserved word, `param` is a user-assigned name for the common area, and `p1`, `p2`, and `p3`—however many are needed—are the variables that will store the parameters. The quantities `function_name`, `x`, and `integrand` have the same meaning as before, and the file must be stored with the name `function_name.pro`.

2. Having defined the function pro-file in this new way, we then invoke IDL to evaluate and display the desired integral by using `luqsimp` with the statements

   ```idl
   common param, p1, p2, p3
   p1 = ?? & p2 = ?? & p3 = ??
   q = luqsimp( 'function_name', a, b )
   print, q
   ```

   At command level, the `common` statement need be invoked only once, regardless of how many different values of the parameters are to be explored. Further, variables referenced in the `common` statement cannot previously have been defined.

   Note that the names of the parameters need not be the same at command level and in the function pro-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.

13.6.9 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 pro-file listed in Table 13.6. We

---

27 This construction was introduced in Section 2.8, which see.
28 Remember the IDL command `delvar` for removing variables that are no longer needed.
Table 13.6: The IDL function rings.pro.

```idl
function rings, phi
    ;
    ; RINGS defines the integrand whose integral gives the
    ; electrostatic potential in the mid-plane between two
    ; uniformly charged circular rings.
    ;
    common param, s
    tmp = sqrt( 2.0 ) / ( 2.0 * !Pi )
    y = tmp / sqrt( 2.0 - 2.0 * s * cos(phi) + s * s )
    return, y
end
```

store this file in the default directory with the name rings.pro, then invoke IDL to evaluate and display the desired integral by using `luqsimp` with the statements

```idl
common param, s
s = ⟨suitable value⟩
q = luqsimp( 'rings', a, b )
print, q
```

Since we seek to explore this integral as a function of $s$ over, say, $0.0 \leq s \leq 4.0$, we must generate a vector of values of $s$ and then, in a loop, invoke `luqsimp` once for each element in that vector. We might use the statements

```idl
IDL> common param, ss
IDL> s = findgen(41)/10.0
IDL> q = size(s) & N = q[1] - 1
IDL> V = fltarr(N+1)
IDL> for i = 0, N do begin $
    ss = s[i] & $
    V[i] = luqsimp( 'rings', 0.0, 2.0*!pi ) & end
IDL> plot, s, V, title = 'Potential in Midplane', thick=3.0, $
IDL> xtitle = 's (in units of a)', chsize=1.25, $
IDL> ytitle = 'V(s) (in units of V(0))', ticklen=1.0
```

The resulting graph is shown in Fig. 13.12.

### 13.12 Evaluating Integrals Numerically with *Mathematica*

*Mathematica* offers at least two routes for evaluating integrals numerically. For integrals that can be evaluated analytically, we can simply embed the command `Integrate` within the command `N`. In response to the simple statement

```mathematica
N[ Integrate[ f, { x, a, b } ] ]
```

where $f$ is an expression that depends on $x$. *Mathematica* first evaluates $\int_a^b f \, dx$ symbolically and then determines the numerical value from the symbolic result. This same statement will automatically invoke numerical quadrature if a symbolic result cannot be obtained (and all quantities in the
integrand have numerical values). If, however, an integral does not admit a symbolic evaluation (or we wish a numeric evaluation even when a symbolic evaluation exists), then we would instead invoke the command \texttt{NIntegrate} in a statement of the form

\[
\text{NIntegrate}[f, \{x, a, b\}]
\]

While the defaults assumed by \texttt{NIntegrate} will frequently be acceptable, this command nonetheless possesses several options that give us the ability to override defaults. Among these options are

- \texttt{WorkingPrecision} (default $\$\text{MachinePrecision}$, often 16), which specifies the number of digits to be used for internal computations.

- \texttt{AccuracyGoal} (default usually figured as half the value of \texttt{WorkingPrecision}), which specifies the absolute accuracy to be sought by stipulating the number of digits to be provided in the solution. \texttt{AccuracyGoal} should not exceed \texttt{WorkingPrecision}.

- \texttt{PrecisionGoal} (default usually figured as half the value of \texttt{WorkingPrecision}), which specifies the relative accuracy sought in the final result. \texttt{PrecisionGoal} should not exceed \texttt{WorkingPrecision}.

  \textit{Note:} With \texttt{AccuracyGoal} set to $a$ and \texttt{PrecisionGoal} set to $p$, \texttt{Mathematica} strives for an absolute accuracy in a value $x$ to be less than $10^{-a} + 10^{-p} |x|$.

- \texttt{MaxRecursion} (default automatic), which specifies the maximum number of recursive refinements to be made in seeking the desired precision. This option provides a means to stop the iterations even if goals on accuracy and precision have not been met.

- \texttt{Method} (default usually an adaptive algorithm chosen by \texttt{Mathematica} to be optimum for the integrand), which specifies the method to be used. Among the allowed values are \texttt{GaussKronrodRule}, \texttt{TrapezoidalRule}, and \texttt{MonteCarloRule}.\footnote{A full listing of allowed values can be found by searching for \texttt{NIntegrateIntegrationRules} in the Wolfram Documentation.}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{potential_in_midplane.png}
\caption{The potential in the midplane of a charged ring.}
\end{figure}
Remember, too, that the statement `Options[NIntegrate]` will display all options and their current default values for the indicated command. Fuller detail about these features of `Mathematica` can be found in the `Mathematica` manuals and in the on-line Help Browser.

With this feature of `Mathematica`, we can evaluate our standard example in many ways. We might, for example, use the statement

```
In[1]:= N[Integrate[2*Exp[-t^2]/Sqrt[Pi], {t, 0, 1}]]
Out[1]= 0.842701
```

which will evaluate the integral analytically [obtaining erf(1)], evaluate the error function of argument 1, and finally display the result with the default precision intrinsic to the command `N`. That precision can, of course, be adjusted with the second argument so that, for example, the statement

```
In[2]:= N[Integrate[2*Exp[-t^2]/Sqrt[Pi], {t, 0, 1}], 20]
Out[2]= 0.84270079294971486934
```

will yield a result with 20 digits. Alternatively, we might use the statement

```
In[3]:= NIntegrate[2*Exp[-t^2]/Sqrt[Pi], {t, 0, 1}]
Out[3]= 0.842701
```

which jumps directly to numerical integration, probably via the (adaptive) Gauss-Kronrod method. To illustrate use of some options, we might invoke the statement

```
In[4]:= NIntegrate[2*Exp[-t^2]/Sqrt[Pi], {t, 0, 1}, Method -> TrapezoidalRule]
Out[4]= 0.842701
```

Reassuringly, these results are all consistent with one another.

### 13.12.1 Quantum Probability

In the opening paragraphs of this section, we have already evaluated the integral that appears in the determination of the probability that a quantum harmonic oscillator in its ground state will be found outside the classical turning point. According to Eq. (13.41), that probability is given by the expression

\[ P(|x| > |x_{\text{turn}}|) = 1.0 - \text{erf}(1.0) = 1.0 - 0.842701 = 0.157299 \]  

(13.108)

which is (probably) correct to all of the digits shown. We conclude that in a bit over 15% of the measurements, the quantum oscillator will be found in the classically forbidden region!

### 13.12.2 The Error Function

To define the function `gaussian` and then generate successive numerical values of the function \( \text{erf}(x) \) defined by Eq. (13.26) as \( x \) varies from 0.0 to 3.0 in steps of 0.1, we might use the statements
In[1]:= gaussian[t_] := 2*Exp[-t^2]/Sqrt[Pi]
In[2]:= uplim = Range[0.0, 3.0, 0.1];
In[3]:= For[j=1, j<32, j++, q[j]=NIntegrate[gaussian[t], {t, 0, uplim[[j]]}] ]

The values of the upper limits in the list uplim and the values \(q_i\) created by these statements must then be converted into a list of the coordinates of the points with a statement like

In[4]:= pts = Table[{uplim[[i]], q[i]}, {i, 1, 31}];

before the graph of Fig. 13.13 can be created with the statement

In[5]:= ListPlot[pts, PlotJoined -> True, PlotStyle -> Thickness[0.01],
PlotLabel -> "Error Function", GridLines -> Automatic]

In this final statement, two new options have been stipulated explicitly. Setting PlotJoined to True instructs Mathematica to connect consecutive points with straight line segments; setting GridLines to Automatic requests the drawing of full-length gridlines at each major tick mark on both axes.

13.12.3 The Off-Axis Electrostatic Potential of Two Rings

Using Mathematica, we could evaluate and plot the integral in Eq. (13.36) to find the electrostatic potential at the radial coordinate \(s_a\) in the midplane between two identical, uniformly charged parallel rings of radius \(a\). The statement

In[1]:= rings[\[Phi]_, s_] := Sqrt[2]/(2*Pi*Sqrt[2-2*s*Cos[\[Phi]]+s^2])

defines the integrand, and the statements
Figure 13.14: Potential in midplane.

In[2]:= s = Range[0.0, 3.0, 0.1];
In[3]:= For[j = 1, j < 32, j++, V[j] = NIntegrate[rings[[phi], s[[j]]],
         {phi, 0, 2*Pi}]]
In[4]:= pts = Table[{s[[i]], V[i]}, {i, 1, 31}];

establish a vector of the desired parameters, evaluate the integral as a function of \( s \) for \( 0.0 \leq s \leq 3.0 \) in steps of 0.1, and create the appropriate list of lists to facilitate plotting. Finally, the statement

In[5]:= ListPlot[pts, PlotRange -> {0.0, 1.0}, PlotJoined -> True,
    PlotStyle -> Thickness[0.01], GridLines -> Automatic,
    PlotLabel -> "Midplane Potential"]

produces the graph shown in Fig. 13.14.

13.15 Exercises

13.15.1 . . . using Symbolic Methods

13.1. A particle of mass \( m \) moves non-relativistically in one dimension under the action of a constant force \( f \). Starting with Eq. (13.3) and using symbolic integration, find the position \( x \), velocity \( v \), and momentum \( p \) of this particle as functions of time if \( x(0) = x_0 \) and \( v(0) = v_0 \).

13.2. A particle of mass \( m \) moves non-relativistically in one dimension under the action of a constant force \( f \). Starting with Eqs. (13.5) and (13.6) and using symbolic integration, find the position \( x \), velocity \( v \), and momentum \( p \) of this particle as functions of time if \( x(0) = x_0 \) and \( v(0) = v_0 \).

13.3. A particle of mass \( m \) moves non-relativistically in one dimension \( x \) under the action of a force given by \( f(x) = -kx \), where \( k \) is a (spring) constant. Starting with Eqs. (13.5) and (13.6) and using
symbolic integration, find the position $x$, velocity $v$, and momentum $p$ of this particle as functions of time if $x(0) = x_0 > 0$ and $v(0) = 0$.

13.4. Suppose an object of mass $m$ moves non-relativistically in one dimension under the action of the force $f(t) = f_0 e^{-bt}$, where both $b$ and $f_0$ are positive. Let $x(0) = x_0$ and $v(0) = v_0$. Use symbolic integration to find $x(t)$ and $v(t)$ by evaluating the integrals in Eq. (13.3). Then, find and interpret both the limits of these two results as $t \to \infty$ and the Taylor expansion of these two results for small $t$.

13.5. The normalized Lorentz distribution function is given by

$$p(x) = \frac{1}{\pi} \frac{a/2}{x^2 + (a/2)^2}$$

Using symbolic integration, (a) verify that $\int_{-\infty}^{+\infty} p(x) \, dx = 1$, (b) evaluate—as best you can—the average $\bar{x}$ and variance $\sigma^2$, defined by

$$\bar{x} = \lim_{b \to \infty} \int_{-b}^{+b} x \, p(x) \, dx \quad \text{and} \quad \sigma^2 = \lim_{b \to \infty} \int_{-b}^{+b} (x - \bar{x})^2 \, p(x) \, dx$$

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, what is the value of this time for the earth-sun system? Hint: Since the gravitational potential is $-GmM/x$, conservation of energy yields

$$\frac{1}{2} m \left( \frac{dx}{dt} \right)^2 - G \frac{mM}{x} = -G \frac{mM}{x_0} \quad \Rightarrow \quad \frac{dx}{dt} = -\sqrt{2GM} \sqrt{\frac{1}{x} - \frac{1}{x_0}}$$

(The negative square root is taken because $x$, the distance to the sun, is known to be decreasing.) This expression then leads to the value

$$T_{\text{midhalf}} = \frac{1}{\sqrt{2GM}} \int_{x_0/4}^{3x_0/4} \left( \frac{1}{x} - \frac{1}{x_0} \right)^{-1/2} \, dx$$

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

$$P(r) = \frac{4}{a^2} \int_0^r e^{-2r'/a} r'^2 \, dr' = \frac{4}{a} \int_0^{r/a} e^{-2\rho} \rho^2 \, d\rho$$

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
302
CHAPTER 13. EVALUATING INTEGRALS

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, $r = z\hat{\mathbf{k}}$, for which $z > a$ by setting up and evaluating the integral

$$V(0,0,z) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(r') \, dv'}{|r - r'|}$$

symbolically to show that $V(0,0,z) = Q/4\pi\epsilon_0 z$. ($Q$ is the total charge in the distribution.) This result demonstrates formally that the potential at a point outside a spherically symmetric charge distribution can be evaluated by regarding the charge to be concentrated at the center of the distribution. Hint: Write $r' = x'\hat{\mathbf{i}} + y'\hat{\mathbf{j}} + z'\hat{\mathbf{k}}$ 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

$$I_{ij} = \int [(x_1^2 + x_2^2 + x_3^2)\delta_{ij} - x_i x_j] \, dm$$

where $x_1$, $x_2$, and $x_3$ symbolize $x$, $y$, and $z$, respectively; $\delta_{ij}$ is the Kronecker delta, which has the value 1 when $i = j$ and the value 0 otherwise; and the radius of gyration is defined in Section 13.1.3.

13.11. In quantum mechanics, the two integrals

$$x_{mn} = \int_{-\infty}^{\infty} \psi_m^*(x) x \psi_n(x) \, dx \quad \text{and} \quad p_{mn} = \int_{-\infty}^{\infty} \psi_m^*(x) \frac{\hbar}{i} \frac{\partial}{\partial x} \psi_n(x) \, dx$$

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

$$x_{mn} = \frac{16a}{\pi^2} (-1)^{(m+n+1)/2} \frac{mn}{(m^2 - n^2)^2}$$

$$p_{mn} = \frac{2i\hbar}{a} (-1)^{(m+n+1)/2} \frac{mn}{(m^2 - n^2)}$$

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

$$f(x) = \frac{x}{l} \quad ; \quad -l \leq x \leq l$$
13.15. EXERCISES

To deduce Simpson’s rule, we start by supposing three consecutive values


(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 $N = 0, 1, 2, 3, 4, 5, 6$.

*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

$$L_0(x) = 1 \quad L_5(x) = \frac{1}{8} (63x^5 - 70x^3 + 15x)$$

$$L_1(x) = x \quad L_6(x) = \frac{1}{16} (231x^6 - 315x^4 + 105x^2 - 5)$$

$$L_2(x) = \frac{1}{2} (3x^2 - 1) \quad L_7(x) = \frac{1}{16} (429x^7 - 693x^5 + 315x^3 - 35x)$$

$$L_3(x) = \frac{1}{2} (5x^3 - 3x) \quad L_8(x) = \frac{1}{16} (6435x^8 - 12012x^6 + 6930x^4 - 1260x^2 + 35)$$

$$L_4(x) = \frac{1}{2} (35x^4 - 30x^2 + 3) \quad L_9(x) = \frac{1}{16} (12155x^9 - 25740x^7 + 18018x^5 - 4620x^3 + 315x)$$

13.15.2  

using Numerical Methods


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.
13.15. Redo several of the examples in Sections 13.6.3–13.6.9 using the built-in IDL routines \texttt{quadimp} and/or \texttt{qromb}. Be sure to include at least one example in each of the three categories (integral that is a single number, integral as a function of a parameter in the limits, integral as a function of a parameter in the integrand). Write a paragraph describing any differences you observe among the several available routines.

13.17. (a) Deduce the three-point Gaussian formula, for which—paralleling Eq. (13.88)—we set

\[ \int_{-1}^{1} f(x) \, dx = w_1 f(x_1) + w_2 f(x_2) + w_3 f(x_3) \]

We then choose the six quantities \( w_i \), and \( x_i \), so that the expression gives an exact result for \( f(x) = 1, x, x^2, x^3, x^4, \) and \( x^5 \). (b) Verify that the interpolation points \( x_i \) and weights \( w_i \) are given by Eq. (13.101) where \( L_3(x) = (5x^3 - 3x)/2 \).

13.18. The (normalized) wave functions for a quantum harmonic oscillator in its first and second excited states \((n = 1, n = 2)\) are

\[ \psi_1(x) = \sqrt{2} \left( \frac{m \omega}{\pi \hbar} \right)^{1/4} y e^{-y^2/2} ; \quad \psi_2(x) = \frac{1}{\sqrt{2}} \left( \frac{m \omega}{\pi \hbar} \right)^{1/4} (2y^2 - 1) e^{-y^2/2} \]

where \( y = x/\sqrt{\hbar \omega/k} \), the energies of these states are \( 3\hbar \omega/2 \), and \( 5\hbar \omega/2 \), respectively, and the symbols have the same meanings as in Section 13.1.8. Find the probability that a harmonic oscillator in each of these states will be found outside the classical turning point.

13.19. The Maxwell-Boltzmann speed distribution yields the integral

\[ f(v) = 4\pi \left( \frac{m}{2\pi kT} \right)^{3/2} \int_0^v e^{-mv'^2/2kT} v'^2 \, dv' \]

for the fraction of the molecules having speed less than \( v \). Using numerical means, explore this integral as a function of \( v \). \textit{Hint:} Re-express the integral using \( \sqrt{2kT/m} \) as the unit of velocity.

13.20. Planck’s black body radiation law gives the expression

\[ I(\nu_2, \nu_1) = \frac{8\pi \hbar}{c^3} \int_{\nu_1}^{\nu_2} \frac{\nu^3}{e^{\hbar \nu/kT} - 1} \, d\nu \]

for the power radiated per unit area in the frequency range \( \nu_1 \leq \nu \leq \nu_2 \). Using numerical means, explore the power radiated in the visible spectrum \( 4 \times 10^{14} \text{ Hz} \leq \nu \leq 7 \times 10^{14} \text{ Hz} \) as a function of temperature. \textit{Hint:} One way to approach this exercise would be to choose a reference frequency \( \nu_0 \) arbitrarily (say \( 10^{14} \text{ Hz} \)) and recast the integral on the dimensionless variable \( s = \nu/\nu_0 \). Examination of \( I \) in units of \( 8\pi \hbar \nu_0^4/c^3 \) as a function of \( T \) in units of \( \hbar \nu_0/k \) would then be indicated.

13.21. As used in statistical data analysis, the Gaussian distribution for a variable \( t \) is usually expressed in terms of the standard deviation \( \sigma \), the distribution function being

\[ \frac{1}{\sqrt{2\pi} \sigma} e^{-t^2/(2\sigma^2)} \]

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 \( P(-\sigma, \sigma), P(-2\sigma, 2\sigma), \) and \( P(-3\sigma, 3\sigma) \) 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? \textit{Hint:} Since the gravitational potential is \(-GmM/x\), conservation of energy yields

\[
\frac{1}{2}m \left( \frac{dx}{dt} \right)^2 - G \frac{mM}{x} = -G \frac{mM}{x_0} \quad \Rightarrow \quad \frac{dx}{dt} = -\sqrt{2GM} \sqrt{\frac{1}{x} - \frac{1}{x_0}}
\]

(The negative square root is taken because \(x\), the distance to the sun, is known to be decreasing.) This expression then leads to the value

\[
T_{\text{midhalf}} = \frac{1}{\sqrt{2GM}} \int_{x_0/4}^{3x_0/4} \left( \frac{1}{x} - \frac{1}{x_0} \right)^{-1/2} dx
\]

\textit{Hint:} The evaluation will be simpler if you begin by recasting the problem in dimensionless terms, expressing lengths in units of \(x_0\) and times in units of \(\sqrt{x_0^3/(2GM)}\). To interpret the significance of this unit of time, determine the period of a circular orbit of radius \(x_0\), which will turn out to be \(2\pi \sqrt{x_0^3/(2GM)}\). For the earth around the sun, this latter time is, of course, 1 year. \textit{Optional:} See if you can develop a means to determine the time required for the first half of the journey, which unfortunately—for numerical approaches—involve 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 \textit{a priori} knowledge of the exact value. \textit{Hint:} Before evaluating the integral, introduce the dimensionless variable \(s = x/a\) and note that the result actually doesn’t depend on \(a\), so there is but one number to determine.

13.24. According to the quantum theory, the probability that the electron in the ground state of the hydrogen atom will be found between the center of the atom and some radius \(r\) is given by

\[
P(r) = \frac{4}{a^3} \int_0^r e^{-2r'/a} \frac{r'^2}{r^2} dr' = 4 \int_0^{r/a} e^{-2\rho \frac{r^2}{a^2}} d\rho
\]

where \(a\) is the Bohr radius and \(\rho = r'/a\). Using numerical integration, evaluate this integral as a function of its upper limit. Then plot and comment on a graph of \(P(r)\) versus \(r/a\).

13.25. The complete elliptic integrals of the first and second kinds are given by

\[
K(k) = \int_0^{\pi/2} \frac{d\phi}{(1 - k^2 \sin^2 \phi)^{1/2}} \quad ; \quad E(k) = \int_0^{\pi/2} (1 - k^2 \sin^2 \phi)^{1/2} d\phi
\]

Explore these integrals as functions of the \textit{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 \(\beta = \sin^{-1}(\sin(\theta/2)/k)\). Remember that, because of the choice of signs (see Section 13.1.4), this integral is valid only during the portion of the swing from \(\theta = 0\) to \(\theta = \alpha\). Obtain graphs of \(\theta\) versus \(\omega t\) over the first quarter of the pendulum’s swing for several different values of \(\alpha\).
13.27. The *n*-th order Bessel function can be defined by the integral

\[ J_n(x) = \frac{1}{\pi} \int_0^\pi \cos(n\theta - x \sin \theta) \, d\theta \]

By evaluating this integral numerically as a function of \( x \) for different values of \( n \), obtain graphs of \( J_0(x) \), \( J_1(x) \), and \( J_2(x) \) over the range \( 0 \leq x \leq 10 \).

13.28. The Bessel function \( J_1(x) \) can be defined by the integral

\[ \frac{1}{x} J_1(x) = \frac{2}{\pi} \int_0^1 \left( 1 - u^2 \right)^{1/2} \cos(xu) \, du \]

Using this definition, obtain a graph of \( J_1(x) \) versus \( x \) over the range \( 0 \leq x \leq 10 \).

13.29. A circular ring of radius \( a \) resides in the \( xy \) plane with its center at the origin and carries a charge \( Q \) uniformly distributed about its perimeter. The electrostatic potential established by this ring at an observation point whose cylindrical coordinates are \((r, \phi, z)\) is

\[ V(r, \phi, z) = \frac{Q}{4\pi\epsilon_0} \frac{1}{a} \int_0^\pi \left( 1 - \frac{2r}{a} \cos \phi' + \frac{r^2}{a^2} + \frac{z^2}{a^2} \right)^{-1/2} \, d\phi' \]

Explore this integral as a function of \( r/a \) for several values of \( z/a \).

13.30. A circular current loop of radius \( a \) lies in the \( xy \)-plane with its center at the origin and carries a current \( I' \) counterclockwise as viewed from a point on the positive \( z \) axis. The magnetic field at a point in the \( xz \) plane is given by

\[ \frac{B(x, z)}{\mu_0 I'/2\pi a} = a^2 \int_0^\pi \frac{z \cos \phi' \hat{i} + (a - x \cos \phi') \hat{k}}{[x^2 + z^2 + a^2 - 2ax \cos \phi']^{3/2}} \, d\phi' \]

Explore both components of this magnetic field numerically as functions of \( x/a \) for various values of \( z/a \), including \( z/a = 0 \) (which will require some creativity for dealing with the point \( x/a = 1 \), at which the integrand diverges at one point in the range of the integration variable).

13.31. In a dimensionless presentation, the intensity in the Fresnel diffraction pattern produced by a single slit when that slit is illuminated by a line source parallel to the slit is proportional to the quantity

\[ I \propto \left| \int_{t_b}^{t_b+\delta t} e^{i\pi r^2/2} \, dt \right|^2 \]

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

\[
 f(x) = 0 \tag{14.1}
\]

If \( f(x) \) is simple, we may be able to find a closed form, analytic solution. More often, however, \( f(x) \) is sufficiently complicated that approximate, numerical methods are needed. Further, many functions will have several roots, some of which may be physically meaningless. Thus, we must learn not only how to find the roots but also how to sort the physically meaningful roots from a possibly larger number of mathematically acceptable ones, the “extras” of which are said to be spurious. We begin this chapter by identifying several physical situations, the full addressing of which requires finding one or more roots of some function. Then we illustrate how to use symbolic algebra systems to approach those that can be addressed analytically, describe a few of many available numerical algorithms, and describe ways to find roots using a variety of numerical approaches and computational tools. Briefly at the end, we comment about the more complicated issue of finding roots of sets of simultaneous linear and non-linear equations. Until that point, our discussion will focus on functions of a single variable.

Whatever the function and whatever the approach, the first step in seeking roots should always be to learn as much as possible about the nature of the function and its roots. Further, since numerical methods in particular—most of them iterative—require a starting guess or guesses and will converge more or less rapidly and reliably depending on the quality of those guesses, a priori knowledge of the approximate location of roots is essential. Thus, we should always start by drawing a graph of \( f(x) \) in sufficient detail to reveal the approximate location of the roots of interest. Since the focus of this chapter is not graphing (and graphing has been fully addressed in earlier chapters), we shall carry out this step once in Section 14.1 as we present several sample problems rather than carrying it out repeatedly in later sections.\(^1\)

14.1 Sample Problems

In this section, we identify several physical contexts in which the essential computational problem is to find the roots of some function, and we obtain graphs of the appropriate functions for later reference.

\(^1\)The graphs could, of course, be produced in any number of ways. Except for Fig. 14.2 (which was produced with tgif), the graphs in Section 14.1 have all been produced with IDL. Those in each later section have been produced by whatever software is the subject of that section.
14.1.1 Classical Turning Points

Let \( V(x) \) be the potential energy under which an object of mass \( m \) is moving in one dimension. Turning points in the motion occur at values of \( x \) where the total energy \( E \) is entirely potential energy (kinetic energy is zero), i.e., when

\[
V(x) = E \quad \text{or} \quad V(x) - E = 0 \tag{14.2}
\]

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

\[
V(x) = \frac{x^3}{10000} + \frac{x^2}{200} - \frac{x}{500} - \frac{1}{2} \tag{14.3}
\]

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

\[
-50.00 < x_1 < -47.50 \quad ; \quad -12.50 < x_2 < -10.00 \quad ; \quad 7.50 < x_3 < 10.00 \quad (14.4)
\]

14.1.2 A Max-Min Problem: Equilibrium Points

From a different perspective, a particle moving in one dimension under the action of the force \( F = -dV/dx \) associated with the potential energy \( V(x) \) will be in equilibrium at those points \( x \) at which the force is zero, i.e., where

\[
F = 0 \quad \implies \quad \frac{dV}{dx} = 0 \tag{14.5}
\]

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

\[
\frac{dV}{dx} = \frac{3x^2}{10000} + \frac{x}{100} - \frac{1}{500} = 0 \tag{14.6}
\]

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

\[
-35.00 < x_4 < -30.00 \quad \text{and} \quad -2.5 < x_5 < 2.5 \quad (14.7)
\]

\[2\]This 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.
14.1. SAMPLE PROBLEMS

Figure 14.1: Graph of the potential energy of Eq. (14.3).

Figure 14.2: A system of two masses and three springs. The surface on which the two objects rest is assumed to be frictionless.

14.1.3 Natural Frequencies of Oscillating Systems

Suppose we seek the natural frequencies of oscillation for the system shown in Fig. 14.2, which consists of two objects, each having mass \( m \). Let these objects move in one dimension on a horizontal, frictionless surface, let them be connected to one another with a spring having constant \( k' \), and let each be connected to the nearer wall with a spring having constant \( k \). Further, let the position of each be measured from its equilibrium position. Then, Newton’s second law combined with Hooke’s law leads to the equations of motion\(^3\)

\[
m \frac{d^2x_1}{dt^2} = -kx_1 + k'(x_2 - x_1) \quad \text{and} \quad m \frac{d^2x_2}{dt^2} = -kx_2 - k'(x_2 - x_1) \quad (14.8)
\]

To cast these equations in dimensionless form, we choose a unit of length \( a \), set \( x_i/a = \xi_i \), introduce \( \omega_0 = \sqrt{k/m} \), set \( \kappa = k'/k \), and introduce \( \bar{t} = \omega_0 t \) to find that

\[
\frac{d^2\xi_1}{d\bar{t}^2} = -(1 + \kappa)\xi_1 + \kappa \xi_2 \quad \text{and} \quad \frac{d^2\xi_2}{d\bar{t}^2} = \kappa \xi_1 - (1 + \kappa)\xi_2 \quad (14.9)
\]

Next, seeking sinusoidal (or simple harmonic) solutions, we suppose that

\[
\xi_1(\bar{t}) = \xi_{10} \cos \omega \bar{t} \quad \text{and} \quad \xi_2(\bar{t}) = \xi_{20} \cos \omega \bar{t} \quad (14.10)
\]

\(^3\)These equations were also discussed in Section 11.1.6.
where the (yet to be determined) frequency $\omega$ is measured in units of $\omega_0$. Substituting these suppositions into Eq. (14.9), we conclude that the (presently unknown) amplitudes must satisfy

$$
\begin{vmatrix}
1 + \bar{\kappa} - \omega^2 & -\bar{\kappa} \\
-\bar{\kappa} & 1 + \bar{\kappa} - \omega^2
\end{vmatrix}
\begin{pmatrix}
\bar{x}_{10} \\
\bar{x}_{20}
\end{pmatrix} = 0
$$

(14.11)

The solution of this equation for the unknowns $\bar{x}_{10}$ and $\bar{x}_{20}$ will be trivial (i.e., both zero—a particularly uninteresting motion) unless the determinant of the matrix of coefficients happens itself to be zero, i.e., unless the characteristic equation

$$
D(\omega) = (1 + \bar{\kappa} - \omega^2)^2 - \bar{\kappa}^2 = \omega^4 - 2\omega^2(1 + \bar{\kappa}) + 1 + 2\bar{\kappa} = 0
$$

(14.12)

is satisfied. We arrive at a fourth-order polynomial $D(\omega)$ in $\omega$, the roots of which will give the natural frequencies of oscillation for the simple system under consideration. Since only even powers of $\omega$ appear, however, we can for the sake of a simpler solution regard the polynomial as quadratic in $\omega^2$.

Suppose we seek the dependence of the roots of this polynomial on $\bar{\kappa} = k'/k$, 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 $\bar{\kappa}$ 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 $\bar{\kappa}$, while the upper of the two roots increases steadily as $\bar{\kappa}$ 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

$$
\frac{d^2x}{dt^2} = -b \frac{dx}{dt} \quad \text{and} \quad \frac{d^2y}{dt^2} = -mg - b \frac{dy}{dt}
$$

(14.13)

Here, $g$ is the acceleration of gravity (which we take to be a positive number). These equations are to be solved subject to the initial values

$$
x(0) = 0 \quad , \quad v_x(0) = v_0 \cos \theta \quad , \quad y(0) = 0 \quad \text{and} \quad v_y(0) = v_0 \sin \theta
$$

(14.14)

The solution to this problem—see Chapter 11 on ordinary differential equations—is

$$
x(t) = \frac{mv_0 \cos \theta}{b} \left( 1 - e^{-bt/m} \right)
$$

(14.15)

$$
y(t) = -\frac{mg t}{b} + \frac{m}{b} \left( v_0 \sin \theta + \frac{mg}{b} \right) \left( 1 - e^{-bt/m} \right)
$$

(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

$$
\frac{-mg t}{b} + \frac{m}{b} \left( v_0 \sin \theta + \frac{mg}{b} \right) \left( 1 - e^{-bt/m} \right) = 0
$$

(14.17)

and then evaluate $x(t_1)$. Equivalently, if we introduce the dimensionless time $\tau = bt/m$ and the dimensionless parameter $\alpha = bv_0/(mg)$, the equation whose root we seek becomes

$$
f(\tau) = \tau - (1 + \alpha \sin \theta) \left( 1 - e^{-\tau} \right) = 0
$$

(14.18)

Once we have found the desired root, say $\tau_1$, we then determine the range $R$ by substituting $\tau_1$ into Eq. (14.15), finding that

$$
\frac{R(\theta, \alpha)}{v_0^2/g} = \frac{\cos \theta \left( 1 - e^{-\tau_1} \right)}{\alpha} = \frac{\cos \theta}{\alpha(1 + \alpha \sin \theta)} \tau_1
$$

(14.19)

The mathematical task confronting us involves finding the non-zero root of the function $f(\tau)$ defined in Eq. (14.18). More specifically, finding the angle $\theta$ at which the projectile should be fired to achieve maximum range would entail

- Choosing a value of $\alpha = bv_0/(mg)$.
- Finding the non-zero root of $f(\tau)$ for several values of $\theta$ ranging from 0 to $\pi/2$.
- Calculating and plotting values of $R(\theta, \alpha)$.
- Finding the value of $\theta$ corresponding to the peak in a graph of $R(\theta, \alpha)$ versus $\theta$.

Note that $\alpha$ increases as $b$ (the damping) increases and as $v_0$ increases but decreases as $m$ increases.

To gain some insight into the nature of the desired roots, we begin by plotting the family of graphs of $f(\tau)$ versus $\tau$ for various values of $\theta$ with fixed $\alpha$. After some exploration using techniques introduced in earlier chapters, we arrive at the several graphs shown in Fig. 14.4. Every graph exhibits a root at $\tau = 0$, corresponding to the moment of launch. Though it is hard to judge, each graph for $\theta = 0^\circ$ also exhibits a second root at $\tau = 0$, since a projectile launched at $\theta = 0^\circ$ returns to its initial altitude immediately. As the angle of launch is increased in each case, however, the second root moves to larger and larger values of $\tau$. Our task in subsequent sections will be to find numerical values for that second root for various values of $\alpha$ and $\theta$, determine the range for each, and find the particular angle at which the range is greatest for a given $\alpha$.

\(^4\)Compare Eqs. (11.3) and (11.4).
Figure 14.4: Graphs of $f(\tau)$ versus $\tau$. The individual frames in this display correspond to different values of $\alpha$; the graphs in each frame correspond to different values of $\theta$, ranging in $15^\circ$ increments from $0^\circ$ for the highest graph to $90^\circ$ 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

$$0 = \tau - (1 + \alpha \sin \theta) \left[ 1 - \left(1 - \tau + \frac{\tau^2}{2}\right) \right] = \tau + (1 + \alpha \sin \theta) \left[ -\tau + \frac{\tau^2}{2} \right]$$

$$= -\alpha \tau \sin \theta + \frac{\tau^2}{2} = 0 \quad \Rightarrow \quad \tau = 2\alpha \sin \theta \quad (14.20)$$

Then, substituting this result into Eq. (14.19), we find that

$$\frac{R(\theta, \alpha \to 0)}{v_0^2/g} = \frac{\cos \theta}{\alpha(1 + \alpha \sin \theta)}(2\alpha \sin \theta) = \frac{2 \sin \theta \cos \theta}{(1 + \alpha \sin \theta) \to 2 \sin \theta \cos \theta} \quad (14.21)$$

which is in complete agreement with the known result for the range of a projectile in the absence of air resistance.
14.1.5 Energy Levels in a Quantum Well

The quantum mechanical analysis for the energy levels of a particle of mass \( m \) in a one-dimensional quantum well characterized by the potential energy

\[
V(x) = \begin{cases} 
\infty & x < 0 \\
-V_0 & 0 \leq x < a \\
0 & a \leq x
\end{cases}
\]  

leads to the conclusion that the allowed energies \( E \) should satisfy the equation\(^5\)

\[
s \cot s = -\sqrt{c^2 - s^2}
\]

where \( c^2 = 2ma^2V_0/\hbar^2 \) 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

\[
s^2 \cot^2 s = s^2 (\sin^{-2} s - 1) = c^2 - s^2 \quad \implies \quad \sin s = \pm \frac{s}{c}
\]  

We can be confident that all solutions of Eq. (14.23) will satisfy Eq. (14.24), but we cannot be sure that all solutions of Eq. (14.24)—a potentially larger set—will satisfy Eq. (14.23). Thus, we must in the end remember to sort from all the roots of Eq. (14.24) only those that 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

\[
\sin s = \pm 0.04 s \quad \text{or} \quad \sin s \mp 0.04 s = 0
\]  

We suppose we seek roots of this function in the interval \( s \geq 0 \). The graphs in Fig. 14.5 reveal these solutions in several ways. The lower graph in Fig. 14.5(a) shows the function \( \sin s - 0.04 s \) (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 \( \sin s + 0.04 s \) (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 \( \sin s, +0.04 s, \) 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

\[
\begin{align*}
\text{root 1u} & \quad -0.5 < s < 0.5 \\
\text{root 2u} & \quad 2.5 < s < 3.5 \\
\text{root 3u} & \quad 6.0 < s < 7.0 \\
\text{root 4u} & \quad 8.5 < s < 9.5 \\
\end{align*}
\]

\[
\begin{align*}
\text{root 5u} & \quad 12.5 < s < 13.5 \\
\text{root 6u} & \quad 14.5 < s < 15.5 \\
\text{root 7u} & \quad 19.0 < s < 20.0 \\
\text{root 8u} & \quad 20.5 < s < 21.5 \\
\end{align*}
\]

and that, for the lower sign, the nine roots are bounded—again crudely—by

\[
\begin{align*}
\text{root 1l} & \quad -0.5 < s < 0.5 \\
\text{root 2l} & \quad 3.0 < s < 4.0 \\
\text{root 3l} & \quad 5.5 < s < 6.5 \\
\text{root 4l} & \quad 9.5 < s < 10.5 \\
\text{root 5l} & \quad 11.5 < s < 12.5 \\
\end{align*}
\]

\(^5\)Text books in quantum mechanics usually treat the finite-depth well for which \( V(x) = -V_0 \) in \(-a \leq x \leq a\) and \( V(x) = 0 \) outside that interval. Nevertheless, the strategy for deriving the result in Eq. (14.23) is described in almost every intermediate level text on quantum mechanics. See, for example, Section 2.6 in David J. Griffiths, *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.
Figure 14.5: Graphs to reveal solutions to Eq. (14.25). In each frame, the graph labeled ‘up’ corresponds to the upper sign in Eq. (14.25) and the graph labeled ‘low’ corresponds to the lower sign in that equation.

14.2 Symbolic Approaches

Symbolic approaches to finding roots always take advantage of specific features of the equation or system of equations for which roots are sought. The roots of at least low order polynomials can be found symbolically. Most of us learned in high school algebra, for example, that the (single) root of a linear polynomial is given by

\[ f(x) = ax + b = 0 \implies x = \frac{-b}{a} \quad (14.28) \]

and that the \textit{quadratic formula}

\[ f(x) = ax^2 + bx + c = 0 \implies x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{2c}{-b \mp \sqrt{b^2 - 4ac}} \quad (14.29) \]

gives the (two) roots of a quadratic polynomial. Similar—though more complicated—expressions exist for the roots of cubic and quartic polynomials. Roots of a polynomial of fifth or higher order cannot be found symbolically unless the polynomial happens to factor into a product of polynomials, each of which individually is of order no higher than quartic. Further, most learned in high school trigonometry that

\[ f(x) = \sin(kx) = 0 \implies x = \frac{n\pi}{k}, \quad n = 0, \pm 1, \pm 2, \ldots \quad (14.30) \]
Occasionally, we will encounter an equation that looks intractable but that can be converted into a tractable case with an appropriate variable transformation. The roots of the expression
\[ f(x) = \frac{a + x}{b + x^2} - c \] (14.31)
for example, can be found symbolically because, so long as \( b + x^2 \neq 0 \), the simple recasting
\[ \frac{a + x}{b + x^2} - c = 0 \implies a + x - c(b + x^2) = -cx^2 + x + a - cb = 0 \] (14.32)
converts the problem into one involving a quadratic polynomial. Beyond these few cases, very few roots can be found symbolically.

### 14.5 Finding Roots Symbolically with Mathematica

The Mathematica command for solving a wide variety of equations is `Solve`. It takes two mandatory arguments, the first giving the equation to be solved and the second specifying the variable for which a solution is desired. Thus, the statement

\begin{verbatim}
In[1]:= Solve[ a*x + b == 0, x ]
Out[1]= { { x \rightarrow -b/a } }
\end{verbatim}

returns the solution \( x = -b/a \) in the form of a transformation rule and the statement

\begin{verbatim}
In[2]:= Solve[ a*x^2 + b*x + c == 0, x ]
Out[2]= { { x \rightarrow \frac{-b - \sqrt{b^2 - 4ac}}{2a} }, { x \rightarrow \frac{-b + \sqrt{b^2 - 4ac}}{2a} } }
\end{verbatim}

returns a list containing two solutions, each expressed as a transformation rule. The command `Solve` actually also knows how to deal with cubic and quartic polynomials, though the result in the most general case is likely to be extremely involved and will be useful only if, once obtained, it is quickly converted to floating point form. The command can solve higher order polynomials only if it recognizes a factorization into polynomials, no one of which is of order higher than quartic.

Sometimes (e.g., for roots of polynomials of order greater than three), `Solve` may be unable to find explicit solutions. In those cases, `Solve` returns the root in the form `Root[...`, e.g.,

\begin{verbatim}
In[3]:= Solve[ a*x^6 - b*x^5 + c == 0, x ]
Out[3]= { { x \rightarrow \text{Root}[c - b #1^5 + a #1^6 & , 1] }, { x \rightarrow \text{Root}[c - b #1^5 + a #1^6 & , 2] }, ... }
\end{verbatim}

(Here, \#1 is a Mathematica-generated generic variable, `Root[f,n]` stands for the n-th root of the function f, and the ampersand identifies what precedes it as a “pure” function. \(^6\)) Basically, this response constitutes an admission by Mathematica that it can’t solve the specified equation symbolically. Fuller description of the command `Solve`, of a third (optional) argument that specifies variables to be eliminated, and of a few options that affect the behavior of the command are described in the Mathematica manuals. Further, Mathematica will display the options and their current values in response to the statement `Options[Solve].`

\(^6\)See Section 2.2.5 in *The Mathematica Book* (Fourth Edition).
14.5.1 Classical Turning Points

The turning points for the potential energy in Eq. (14.3) are readily found symbolically with the statements

\begin{verbatim}
In[4]:= V = x^3/10000 + x^2/200 - x/500 - 1/2;
In[5]:= soln = Solve[ V == 0 , x ]
Out[5]= 
   {{x → mess}, {x → mess}, {x → mess}}
\end{verbatim}

where the indicated messes, which—because of the presence of $i = \sqrt{-1}$ here and there—will appear superficially to be complex, are too long and too unfathomable to reproduce here. We can, however, convert this output directly to a (complex) floating point form with the statement

\begin{verbatim}
In[6]:= soln // N
Out[6]= 
   {{x → 9.3487041976000522706 + 0. \times 10^{-20} i},
    {-48.268267566751634131 + 0. \times 10^{-19} i},
    {-11.08043663084841393 + 0. \times 10^{-20} i}}
\end{verbatim}

thereby reducing the messes to quantities that are more easily interpreted. The real parts of the roots here obtained are certainly consistent with the bounds given in Eq. (14.4). Further, the assertion that the imaginary parts of the roots are all zero is not unreasonable, given that, by default, the precision of the command N is on the order of $10^{-15}$. Indeed, we can support this understanding of the imaginary part by evaluating the roots to more decimal places with the statements

\begin{verbatim}
In[7]:= N[ soln, 20]
Out[7]= 
   {{x → 9.3487041976000522706 + 0. \times 10^{-20} i},
    {-48.268267566751634131 + 0. \times 10^{-19} i},
    {-11.08043663084841393 + 0. \times 10^{-20} i}}
\end{verbatim}

Even at this precision, the assertion that the imaginary parts are all zero is not unreasonable, given that the precision of the command N is on the order of one in the twentieth digit when its second argument is 20.\footnote{While different machines should give the same results for most of the digits in the real parts of these roots, the values of the imaginary parts may well be quite different from machine to machine, since those values are largely a consequence of internal roundoff.}

14.5.2 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

\begin{verbatim}
In[4]:= V = x^3/10000 + x^2/200 - x/500 - 1/2;
In[5]:= soln = Solve[ V == 0 , x ]
Out[5]= 
   {{x → mess}, {x → mess}, {x → mess}}
\end{verbatim}
In[8]:= dVdx = D[ V, x ];
In[9]:= Solve[ dVdx == 0, x ]
Out[9]= 
\[ \{ \{ x \rightarrow \frac{2}{3} \left( -25 - 8\sqrt{10} \right) \}, \{ x \rightarrow \frac{2}{3} \left( -25 + 8\sqrt{10} \right) \} \} \]
In[10]:= soln = N[ % ]
Out[10]= 
\[ \{ \{ x \rightarrow -33.5321 \}, \{ x \rightarrow 0.198814 \} \} \]

finding first $dV/dx$ and then finding the values of $x$ at which $dV/dx = 0$. These results are consistent with the bounds obtained in Eq. (14.7). Finally, to assess the stability of each equilibrium, we evaluate the second derivative at each equilibrium point with the statements

In[11]:= d2Vdx2 = D[ dVdx, x ];
In[12]:= d2Vdx2 /. soln
Out[12]= 
\[ \{ -0.0101193, 0.0101193 \} \]
In[13]:= Quit[]

concluding that the first root $x = -33.5321$ locates an unstable equilibrium ($d^2V/dx^2 < 0$; local maximum in the potential energy) while the second root $x = 0.198814$ locates a stable equilibrium ($d^2V/dx^2 > 0$; local minimum in the potential energy).

### 14.5.3 Natural Frequencies

To find the natural frequencies for the coupled oscillators described in Section 14.1.3, we seek roots of the function $D(\omega)$ defined in Eq. (14.12). Since $D(\omega)$ is, in fact, quadratic in $\omega^2$, we solve first for $\omega^2$—though we must substitute a single variable $\lambda$ for $\omega^2$ before doing so—and then we take the square root of that solution with the statements

In[1]:= d = \[Omega]^4 - 2*\[Omega]^2*(1+\[Kappa]) + 1 + 2*\[Kappa];
In[2]:= d /. \[Omega] -> Sqrt[\[Lambda]]; 
In[3]:= Solve[ % == 0, \[Lambda] ]
Out[3]= 
\[ \{ \{ \lambda \rightarrow 1 \}, \{ \lambda \rightarrow 1 + 2\kappa \} \} \]
In[4]:= freqs = Sqrt[ \[Lambda] ] /. %
Out[4]= \[ 1, \sqrt{1 + 2\kappa} \]

Note—as must on physical grounds be the case—that the solutions obtained for $\lambda = \omega^2$ are both positive, so the square roots will be real. (We ignore the negative square roots, since they don’t have the physical significance of the positive square roots.) Note also that the first of the two frequencies is the same for all $\kappa$ but that the second increases steadily as $\kappa$—the strength of the coupling—increases.

To make the point at the end of the last paragraph even more explicitly, we generate a graph of the two frequencies as functions of $\kappa$ with the statement

---

8At this point, for simplicity, we drop the overbar on $\kappa$.
9Evidently, we cannot simply place the list `freqs` as the first argument to `Plot`. The first argument can be a list, but the list must be constructed within the statement rather than outside of the statement.
Figure 14.6: Natural frequencies for coupled oscillators. The horizontal coordinate is $\kappa = k'/k$ and the vertical coordinate is the frequency in units of $\omega_0 = \sqrt{k/m}$.

The resulting graph is shown in Fig. 14.6.

14.5.4 Range of Projectile in Viscous Medium

While the problem of finding the range of a projectile in a viscous medium is not effectively addressed by symbolic methods, certain limits of the behavior of that projectile are readily obtained by invoking Mathematica. We begin by defining the trajectory as given by Eqs. (14.15) and (14.16) with the statements

\begin{verbatim}
In[1]:= x = (m*v0*Cos[\[Theta]]/b) * ( 1 - Exp[-b*t/m] );
In[2]:= y = -m*g*t/b+(m/b)*(v0*Sin[\[Theta]]+m*g/b) * ( 1 - Exp[-b*t/m] );
\end{verbatim}

In many instances, the damping constant $b$ is small. Thus, we can sometimes be satisfied with a Taylor expansion of these quantities about $b = 0$, an expansion we obtain with the Mathematica statements

\begin{verbatim}
In[3]:= x = Series[x, {b, 0, 2}];
In[4]:= y = Series[y, {b, 0, 2}];
\end{verbatim}

\footnote{To facilitate visualization, some of the output in the following dialog with Mathematica has been recast from the actual form presented by Mathematica.}
14.5. FINDING ROOTS SYMBOLICALLY WITH MATHEMATICA

```
In[3]:= xb = Series[ x, { b, 0, 1 } ]
Out[3]=
t v0 Cos[θ] - t^2 v0 Cos[θ] b/2m + O[b]^2

In[4]:= yb = Simplify[ Series[ y, { b, 0, 1 } ] ]
Out[4]=
- t^2 v0 Sin[θ] + t^2 (g t - 3v0 Sin[θ]) b/6m + O[b]^2
```

We notice first—reassuringly—that the terms that are zeroth-order in \( b \), i.e., \( \frac{t v_0 \cos \theta}{2m} \) and \( \frac{t^2 v_0 \cos \theta b}{2m} \), 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 converting the expression to a polynomial, removing the root at \( t = 0 \), and solving for the roots of the Taylor expansion of \( y \) with the statements

```
In[5]:= yb = Normal[ yb ];
In[6]:= yb = Expand[ yb/t ];
In[7]:= soln = Solve[ yb == 0, t ]
Out[7]=
{ { t \rightarrow mess }, { t \rightarrow 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

```
In[8]:= Series[ t /. soln[[1]], { b, 0, 1 } ]; Find limit of first solution.
In[9]:= Simplify[ %, { m > 0, g > 0 } ]
Out[9]=
\( \frac{2 t v_0 \sin \theta}{g} - \frac{2 t^2 \sin \theta^2 b}{3g^2 m} \) + \( O[b]^2 \)

In[10]:= Series[ t /. soln[[2]], { b, 0, 1 } ]; Find limit of second solution.
In[11]:= Simplify[ %, { m > 0, g > 0 } ]
Out[11]=
\( \frac{3 m}{b} + \frac{3 t v_0 \sin \theta}{g} - \frac{2 t^2 \sin \theta^2 b}{3g^2 m} \) + \( O[b]^2 \)
```

Since the second root becomes arbitrarily large as \( b \) becomes small, we must reject it on physical grounds. Only the first root, which we can display more fully with the statement

```
In[12]:= sln1 = t /. soln[[1]];  
In[13]:= sln2 = Expand[ Simplify[ sln1 ] ]
Out[13]=
\( \frac{3 m}{2b} + \frac{3 t v_0 \sin \theta}{2g} - \frac{m \sqrt{9g^2 - 6bgv_0 \sin \theta}/m + 9b^2 v_0^2 \sin \theta^2/m^2}{2bg} \)
```

is of interest. Finally, for consistency with approximations already made, we find this root to first order in \( b \) with the statement

```
In[14]:= Series[ sln2, { b, 0, 1 } ];
In[15]:= tim = Simplify[ %, { m > 0, g > 0 } ]
Out[14]=
\( \frac{2 t v_0 \sin \theta}{g} - \frac{2 t^2 \sin \theta^2 b}{3g^2 m} \) + \( O[b]^2 \)
```

and find the range by substituting this result into the first order expression for the horizontal coordinate with the statements
CHAPTER 14. FINDING ROOTS

In[16]:= rang = Normal[ xb ] /. t -> Normal[tim];
In[17]:= rang = Series[ rang, { b, 0, 1 } ];
In[18]:= rang = Factor[ Normal[ rang ] ]
Out[18] := \( \frac{2v_0^2}{g} \sin[\theta] \cos[\theta] \left( 1 - \frac{4v_0 b}{3mg \sin[\theta]} \right) \)

In[19]:= Quit[]

(We have rearranged this expression a bit from the one presented by *Mathematica.*) Note in particular that, in the absence of viscous resistance \( (b = 0) \), this result reduces to the known result \( \left( \frac{v_0^2}{g} \right) \sin[2\theta] \).

14.6 Algorithms for Finding Roots Numerically

In this section we describe several methods for determining a root of a general function numerically.

14.6.1 The Method of Bisection

Suppose, for example, that we know that a single root of \( f(x) \) lies between \( x_{\min} \) and \( x_{\max} \). We can, of course, calculate \( f_{\min} = f(x_{\min}) \) and \( f_{\max} = f(x_{\max}) \). With this input, we can refine our knowledge of the interval in which the root lies by

1. calculating

\[ x_{\text{mid}} = \frac{1}{2}(x_{\min} + x_{\max}) \quad \text{and} \quad f_{\text{mid}} = f(x_{\text{mid}}) \]  

2. calculating \( f_{\min} f_{\text{mid}} \), which will be positive if \( f_{\min} \) and \( f_{\text{mid}} \) have the same sign (and the root then lies in the upper half of the original interval) and negative if \( f_{\min} \) and \( f_{\text{mid}} \) have opposite signs (and the root then lies in the lower half of the original interval). Note that we have here assumed the root to be a single root (or at least a root of odd multiplicity), since the criterion invoked depends on the function having opposite signs on opposite sides of the root. (For a root of even order, the function has the same sign on opposite sides of the root, and the criterion here described will not identify the half of the interval in which the root lies. Without embellishment, bisection will fail in this case.)

3. refocusing our attention on the interval \( x_{\min} < x < x_{\text{mid}} \) (i.e., replacing \( x_{\max} \) with \( x_{\text{mid}} \)) or on the interval \( x_{\text{mid}} < x < x_{\max} \) (i.e., replacing \( x_{\min} \) 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 \( x_{\max} \). With each successive iteration, the interval within which we know the root to lie is shrunk to half of its size at the start of that iteration. This method of bisection is guaranteed to converge provided only that there is a root in the interval, though it will not work if the root is a root of even order and it may be confused if the original interval happens to contain several roots. The iteration is continued until the interval has been reduced to a size that we are willing to accept as a tolerance or, though it happens rarely, if \( f_{\text{mid}} \) actually is zero at some point in the process.
14.6. ALGORITHMS FOR FINDING ROOTS NUMERICALLY

Figure 14.7: The method of bisection. The two parts show the positioning of critical points (a) when the function increases as \( x \) increases and (b) when the function decreases as \( x \) increases.

\[
\begin{align*}
(a) & \quad f & \quad x_{\text{mid}} & \quad x_{\text{max}} & \quad x_{\text{min}} \\
(b) & \quad f & \quad x_{\text{mid}} & \quad x_{\text{max}} & \quad x_{\text{min}}
\end{align*}
\]

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.

\[
\begin{align*}
f & \quad x_{n+1} & \quad f(x_n) & \quad x_n & \quad \text{tangent line at } x_n \\
f(x_n) & \quad x_{n+1} & \quad x_n & \quad \text{tangent line at } x_n
\end{align*}
\]

14.6.2 Newton’s Method

A second—one more efficient but also less stable—algorithm requires a single starting value rather than a pair of values that bracket a root. Suppose that \( x_n \) is the current estimate of the position of the desired root, and let \( f_n = f(x_n) \). In Newton’s method, we calculate the next iterate \( x_{n+1} \) by

1. calculating the derivative of the function at \( x_n \), i.e., \( df(x)/dx|_{x_n} = f'_n \).
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

\[
\frac{\Delta f}{\Delta x} = \frac{0 - f_n}{x_{n+1} - x_n} = f'_n \quad \implies \quad x_{n+1} = x_n - \frac{f_n}{f'_n}
\]

(14.34)

The relevant geometry is illustrated in Fig. 14.8. Given an initial “guess”, Newton’s method will converge more rapidly than the method of bisection, though it will have difficulties if the initial guess—or, for that matter, the root itself—is too close to a point at which the derivative of the function is zero. With a poor initial guess, the method may diverge altogether or may converge on a root remote from the one sought.

14.6.3 Other Methods

Numerous other methods for finding roots, some of them restricted to polynomials, have been developed. Because available routines in some programs sometimes use methods other than the two
described above in detail, we include here a brief outline of the main idea in each of several other methods:

- **The secant method** starts from two approximations to the root, say \( x_1 \) and \( x_2 \), calculates \( f_1 = f(x_1) \) and \( f_2 = f(x_2) \), fits a straight line through the points \((x_1, f_1)\) and \((x_2, f_2)\), finds the intersection point \( x_3 \) of that line with the \( x \) axis, and then repeats the process with the two approximations \( x_2 \) and \( x_3 \). The method is similar to Newton’s method, except that it uses the values of the function at two points to estimate the slope of the tangent line to the function \( f(x) \). It does not require explicit knowledge of the derivative of \( f(x) \).

- **Müller’s method** is mildly more sophisticated than the secant method but involves a similar idea. It starts with three estimates of the root, say \( x_1 \), \( x_2 \), and \( x_3 \), calculates \( f_1 = f(x_1) \), \( f_2 = f(x_2) \), and \( f_3 = f(x_3) \), fits a *parabola* to the points \((x_1, f_1)\), \((x_2, f_2)\), and \((x_3, f_3)\), finds the roots of that parabola using the form of the quadratic formula given second in Eq. (14.29), and finally takes as \( x_4 \)—which replaces \( x_1 \) for the next iteration—the one root that emerges when the ambiguous sign in Eq. (14.29) is chosen so that the denominator has the larger of the two possible absolute values.

- **Laguerre’s method**, which is limited to polynomials, (1) recognizes that, in terms of its roots \( x_i \), a polynomial scaled so that the coefficient of its highest order term is 1 (which has no effect on its roots) can be expressed in the form

\[
    f(x) = \prod_{i} (x - x_i) \tag{14.35}
\]

(2) finds a couple of relationships satisfied by polynomials in this form, (3) assumes with only weak justification *a priori* that the root sought is isolated from all the others (which are clustered together), and (4) deduces that the next iterate \( x_{n+1} \) for the root sought should be determined from the current guess \( x_n \) by

\[
    x_{n+1} = x_n - \frac{N}{G \pm \sqrt{(N - 1)(NH - G^2)}} \tag{14.36}
\]

where the sign in the denominator is chosen so that the denominator has the larger of the two possible absolute values, \( N \) is the order of the polynomial (equal to the number of roots), and \( G \) and \( H \) are defined by \( G = f'/f \) and \( H = (f'/f)^2 - f''/f \), each evaluated at \( x = x_n \).

### 14.6.4 Assessing Accuracy

As with all numerical operations, assessing the accuracy of the roots found by an algorithm is essential before we can have confidence in the roots. We can, of course, always substitute the root we have found into the function whose root we seek and simply notice how close to zero the value of the function at the proposed root actually is. In the end, that comparison may provide the most important test of accuracy, though whether the value \( f(x_{\text{root}}) = 0.0001 \), say, is “close enough” to zero requires a judgment that would take into account the magnitude of \( f(x) \) over the important domain of \( x \) and, even more, the magnitude of the derivative of \( f(x) \) evaluated at the root.\(^{11}\) Certainly, that criterion provides the most accessible test we might apply. It is also a criterion that is easily implemented in a computer program that monitors convergence and stops automatically when the absolute value of the function has been reduced below some specified tolerance.

We would, however, be more interested in the accuracy of the root itself, i.e., in the amount \( |x_{\text{exact}} - x_{\text{approx}}| \) by which the (unknown) exact root differs from the root on which a particular

\(^{11}\)The less steeply sloped the function at the root, the harder it is to obtain an accurate root by using the value of the function as a guide.
algorithm converges. Only the method of bisection supports a clear assessment of that difference since, at any particular step in the progress of that algorithm towards a root, we know that the root is “trapped” between two values whose separation is halved with each step in the iteration. Continuing until the difference \( |x_{i+1} - x_i| \)—sometimes called the residual—between consecutive iterates is less than some specified tolerance guarantees that we have located the root to within that tolerance. Implemented in a computer program, an algorithm using the method of bisection can monitor the separation of consecutive iterates and stop when that separation, either as an absolute value or as a fraction of its current value, is reduced below a specified absolute or relative tolerance, respectively.

Given appropriate initial bounds, the method of bisection is guaranteed to converge on a root, but the rate of convergence is slow by comparison with other methods. Assessing the accuracy of the root itself for those other algorithms is more difficult. If, however, the convergence of the algorithm in use is fairly rapid (and Newton’s method will usually satisfy this expectation), taking the (absolute) accuracy of a particular iterate to be on the order of the difference between it and the next iterate, i.e., on the order of the residual, is reasonable. For these algorithms, as with the method of bisection, a computer implementation of a method in the category of this paragraph can also monitor the residual and stop when it has been reduced below a specified (absolute or relative) tolerance. For these methods, however, convergence is not guaranteed, and computer implementations of these methods should include an alternate criterion for stopping the iteration. Such an implementation might, for example, limit the number of iterations and—to keep the user fully informed—display a warning message if the iteration is stopped because this limit is exceeded rather than because the residual has been made sufficiently small. Without some such fail-safe stopping criterion, these methods are in danger of iterating forever.

Beyond monitoring the value of the function or the residual, seeking a particular root by more than one method may give some insight into the accuracy of that root. Different methods have different strengths, weaknesses, and quirks. When they agree to some number of digits, we can have more confidence in the result than we would have if we had obtained it by only one method. When they disagree, we have at least a hint that the function at hand possesses some pathology that we should perhaps attempt to understand before accepting the root we have found.

### 14.7 Finding Roots Numerically with IDL

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

- **fz_roots**, which uses Laguerre’s method to find the roots of a polynomial, whose coefficients may be complex. It is invoked with a statement like

  \[
  q = \text{fz\_roots}( c, /\text{double}, \text{eps}=value) \]

  where \( c \), which is mandatory, is a (possibly complex) vector containing the coefficients of the terms in the polynomial (constant first), and \( q \) is a complex vector storing the returned roots. If included, the optional keyword \text{double} will select a double-precision calculation; the optional keyword \text{eps} (default \( 2 \times 10^{-6} \)) specifies the desired fractional accuracy.

- **fx_root**, which uses Müller’s method to find real and complex roots of nonlinear functions of a single variable. It is invoked with a statement like

  \[
  q = \text{fx\_root}( x, 'func', /\text{double}, \text{itmax}=value, \text{tol}=value) \]
where \(x\), which is mandatory, is a three-element vector containing initial guesses, \(func\)—also mandatory—is a string giving the name of a function pro-file that returns the value of the function whose roots are sought, and \(q\) stores the returned root. If included, the optional keyword \(\text{double}\) will select a double-precision calculation; the optional keyword \(\text{itmax}\) (default 100) specifies the maximum number of iterations and the optional keyword \(\text{tol}\) (default \(10^{-4}\)) specifies the convergence criterion by stipulating the largest acceptable difference between consecutive iterates.

- \(\text{newton}\), which uses Newton's method to determine a root of a nonlinear function of a single variable. It is invoked with a statement like

\[
q = \text{newton}( x, func', /\text{double}, \text{itmax}=\text{value}, \text{tolf}=\text{value}, \text{tolx}=\text{value})
\]

where \(x\), which is mandatory, provides the initial guess for the desired root, \(func\)—also mandatory—is a string giving the name of a function pro-file that returns the value of the function whose roots are sought, and \(q\) stores the returned root. If included, the optional keyword \(\text{double}\) will select a double-precision calculation; the optional keyword \(\text{itmax}\) (default 200) specifies the maximum number of iterations; the optional keyword \(\text{tolf}\) (default \(10^{-4}\)) specifies how close to zero the value of the function at the root is to be before stopping; and \(\text{tolx}\) (default \(10^{-7}\)) stipulates the largest acceptable difference between consecutive iterates. Other optional keywords are described in the IDL manuals.

In addition, Section 14.A contains a listing of the Lawrence-written program \(\text{lubisect.pro}\), which defines an IDL function that uses the method of bisection. It is called with a statement of the form

\[
q = \text{lubisect}( func', xlb, xub, \text{tol}=\text{value}, \text{itmax}=\text{value})
\]

where \(func\), which is mandatory, is a string giving the name of a function pro-file that returns the value of the function whose roots are sought, \(xlb\) and \(xub\)—also mandatory—give the lower and upper bounds on an interval containing the desired root, and \(q\) stores the returned root. If included, the optional keyword \(\text{tol}\) (default \(10^{-6}\)) stipulates the largest acceptable difference between consecutive iterates; the optional keyword \(\text{itmax}\) (default 20) specifies the maximum number of refinements before the search will terminate in case convergence as specified by \(\text{tol}\) fails. 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 \(xlb\).

### 14.7.1 Classical Turning Points

We might, for example, find the turning points for the potential energy in Eq. (14.3) with the IDL statements

```
IDL> A = [ -0.5, -1.0/500.0, 1.0/200.0, 1.0/10000.0]
IDL> roots = fz_roots( A )
IDL> print, roots
(-48.2683, 0.00000) (-11.0804, 0.00000) ( 9.34870, 0.00000)
```

Even though they have been presented as complex numbers, these roots have turned out to be real (zero imaginary part). For later manipulations, we might actually prefer to have the roots in a three-component real vector rather than a three-component complex vector. IDL’s command \(\text{float}\), which converts other data types to floating point representation, is built so that the conversion of a complex value to a floating value is equivalent to extracting the real part of the complex value. Thus, the simple statements
14.7. FINDING ROOTS NUMERICALLY WITH IDL

IDL> roots = float(roots)
IDL> print, roots
   -48.2683  -11.0804   9.34870

extract the real parts alone, store them in a simple real vector, and display that vector.

Other methods for finding the required roots all require that the function be defined in a pro-file. For this example, we might construct the file

\[
\text{function turnpts, x} \\
\text{return, x^3/10000.0+x^2/200.0-x/500.0 - 0.5} \\
\text{end}
\]

which we store in the default directory with the name \text{turnpts.pro}. Then, we would find the roots of the defined polynomial with the IDL statements

IDL> trial = [-60.0, -50.0, -40.0]  \quad \text{Guess for lowest root.}
IDL> root = fx_root( trial, 'turnpts' ) \quad \text{Invoke fx_root.}
IDL> print, root
   -48.2683

IDL> trial = [-50.0, -40.0, -30.0]  \quad \text{Guess for middle root.}
IDL> ifx_root( trial, 'turnpts' ) \quad \text{Invoke fx_root and display root.}
IDL> print, root
   -11.0804

IDL> trial = [10.0, 20.0, 30.0] \quad \text{Guess for highest root.}
IDL> print, fx_root( trial, 'turnpts' ) \quad \text{Invoke fx_root and display root.}
IDL> print, root
   9.34870

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 pro-file can alternatively be used in an invocation of \text{newton} with the statement

IDL> print, newton( -50.0, 'turnpts' )
   -48.2683

or in an invocation of \text{lubisect} with the statement

IDL> print, lubisect( 'turnpts', -50.00, -47.50 )
   -48.2683

With different starting values, the routines \text{newton} and \text{lubisect} would, of course, converge on the remaining roots.

14.7.2 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 \text{fz_roots}. We seek the roots as functions of \(\rho\), which we shall represent by the variable \text{kvals} in our statements to IDL. 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 \(\rho\) over the interval \(0.0 \leq \rho \leq 4.0\), 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 \(\rho\). Let us therefore begin by creating two arrays, the first of which—a \text{column} vector with 41 elements—contains the values of \(\rho\) 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 \(\rho\). These ends are achieved with the statements
CHAPTER 14. FINDING ROOTS

Table 14.1: The IDL function projectile.pro.

```idl
function projectile, tau
  common params, alpha, theta ; Provide for passing parameters
  tmp = 1.0 + alpha * sin(theta) ; Calculate coefficient
  return, tau - tmp * (1 - exp(-tau)) ; Return value of f(tau)
end
```

IDL> kvals = transpose( findgen(41)/10.0 )
IDL> ones = transpose( 1.0 + fltarr(41) )
IDL> coeffs = [ 1.0 + 2.0 * kvals, -2.0 * (1.0 + kvals), ones ]

which set the values of \( \kappa \), create a 41-element column vector of ones, and then create the \( 3 \times 41 \) array of coefficients. Then, we find the roots and concatenate them into a single array with the statements

IDL> roots = fz_roots( coeffs[*,0] )
IDL> for i = 1, 40 do roots = [ [roots], [fz_roots( coeffs[*,i] )] ]

The result is a \( 2 \) column \( \times 41 \) row array of complex values, the first column containing the first root for each value of \( \kappa \) and the second column containing the second root. As we suspect (and as the statement print, roots will confirm), these roots are all real; their imaginary parts are all zero. Thus, we extract the only physically significant part—the real part—with the statement

IDL> roots = float( roots )

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 roots and the other contained in the second column—with the statements

IDL> roots = sqrt( roots )
IDL> plot, kvals, roots[0,*], ticklen=1.0, thick=3.0, yrange=[0.0,3.0]
IDL> oplot, kvals, roots[1,*], thick=3.0

The resulting graph is shown in Fig. 14.9.

14.7.3 Range of Projectile

As laid out in Section 14.1.4, finding the range of a projectile moving in a viscous medium begins with finding the time at which a projectile launched at some angle \( \theta \) in a medium characterized by a (dimensionless) viscous damping coefficient \( \alpha = \frac{b v_0}{m g} \) returns to its initial altitude. In other words, we must find the one non-zero root of the function \( f(\tau) \) defined in Eq. (14.18). Central to the entire calculation is the function pro-file listed in Table 14.1, which we store in the default directory in a file named projectile.pro. Indirectly, we have already used this function, since, for example, in creating Fig. 14.4, we set various preliminary variables with the statements
14.7. FINDING ROOTS NUMERICALLY WITH IDL

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.

IDL> common params, alpha, theta
IDL> thetavals = !pi * findgen(7)/12.0
IDL> !p.ticklen=1.0 & !p.thick = 3.0
IDL> !p.multi=[0,2,2] & !p.charsize=0.8
IDL> tau=findgen(201)/100.0

Provide for passing parameters.
Set \( \theta = 0^\circ, 15^\circ, \ldots, 90^\circ \) in radians.
Set ticklength; set line thickness.
Set \( 2 \times 2 \) display; set character size.
Set values of \( \tau \).

Then, we produced the upper left frame with the statements

IDL> alpha = 0.1
IDL> theta = thetavals[0]
IDL> plot, tau, projectile(tau),
IDL> xrange=[0.0,0.25],
IDL> yrange=[-0.005,0.005],
IDL> title='!4a!3=0.1', xtitle='!4s!3',
IDL> ytitle='f(!4s!3)
IDL> for i=1,6 do begin
IDL> theta = thetavals[i] &
IDL> oplot, tau, projectile(tau) &
IDL> endfor

Choose \( \alpha \).
Plot graph for \( \theta = 0^\circ \).
Overplot graphs for \( \theta = 15^\circ, \ldots, 90^\circ \).

The remaining frames were produced with similar statements.

For this example, we choose IDL’s built-in routine \texttt{newton}.\footnote{The function of interest is not a polynomial, so \texttt{fx_roots} is not an option. The routines \texttt{fx_root} and \texttt{lubisect} would work, but they require three and two initial guesses, respectively. Since we want to find many roots, our task will be simplest if we are obliged to provide only one initial guess.} Starting with a fresh IDL 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

```idl
IDL> for i=1,6 do begin
IDL> theta = thetavals[i] &
IDL> oplot, tau, projectile(tau) &
IDL> endfor
```
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

```idl
IDL> range = cos(theta) * (1-exp(-root)) / alpha
IDL> print, range
0.679178
```

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 \(\theta\). Continuing in the same session with IDL, we prepare for a calculational loop with the statements

```idl
IDL> thetadeg = findgen(91)  
Set values of \(\theta\) in degrees.
IDL> thetarad = !pi*thetadeg/180.0  
Convert \(\theta\) to radians.
IDL> roots = fltarr(91)  
Prepare array for roots.
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

```idl
IDL> alpha = 0.1 & tau0 = 0.2  
Set \(\alpha\), initial guess for root.
IDL> for i = 0, 90 do begin  
Find root for each \(\theta\).
IDL>    theta = thetarad[i] &
IDL>    roots[i] = newton( tau0, 'projectile',
IDL>        /double, tolf=1e-8, tolx=1e-10 ) &
IDL> endfor
IDL> range1 = cos(thetarad)*(1-exp(-roots))/alpha  
Evaluate \(R\) for each root.
IDL> plot, thetadeg, range1, thick=3.0, ticklen=1.0, 
IDL> xtitle='\4h!3 (deg)', ytitle='R(\4h!3)' 
Plot \(R\) versus \(\theta\).
```

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 \(range2, range4,\) and \(range8,\) respectively, and overplotting each on the axes created above, we finally produce Fig. 14.10. As \(\alpha\) increases (\(b\) increases, \(v_0\) increases, \(m\) decreases, or some combination), the maximum range (measured in units of \(v_0^2/g\)) decreases and the angle of fire to achieve that range becomes shallower and shallower.

Even more specifically, we can examine the values in the four variables \(range\) to find the maximum range and the approximate firing angle to achieve it in each case. For example, the statements

```idl
IDL> print, max(range1)
0.913793
IDL> print, range1[43], range1[44], range1[45]
0.913564 0.913793 0.912932
```

\(^{13}\)A bit of exploration was needed before choosing the tolerances specified in the call to \texttt{newton}. As might be expected from the appearance Fig. 14.4, particular difficulties with accuracy appeared for smaller values of \(\alpha\) and especially for small values of \(\theta\), where the zero and non-zero roots are quite close together.
14.7. FINDING ROOTS NUMERICALLY WITH IDL

Figure 14.10: Range versus $\theta$ for $\alpha = 0.1$ (highest graph), $\alpha = 0.2$, $\alpha = 0.4$, and $\alpha = 0.8$ (lowest graph).

Table 14.2: The IDL functions qmupper.pro and qmlower.pro.

```
function qmupper, s ; Name function and variable
    return, sin(s) - 0.04*s ; return value with upper sign
end

function qmlower, s ; Name function and variable
    return, sin(s) + 0.04*s ; Return value with lower sign
end
```

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 $44^\circ$. Fitting a parabola to the three points and finding its maximum yields the angle $43.77^\circ$. Similar statements applied to the other three values of $\alpha$ yield that the maximum range and firing angle are $0.841218$ and about $43^\circ$ ($42.50^\circ$) for $\alpha = 0.2$, $0.725949$ and about $40^\circ$ ($40.35^\circ$) for $\alpha = 0.4$, and $0.569568$ and about $37^\circ$ ($37.00^\circ$) for $\alpha = 0.8$, respectively.

14.7.4 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 pro-file qmlevels.pro containing the lines listed in Table 14.2. Since we have put two routines in a single file, we must begin our search for the actual roots by explicitly compiling qmlevels. Thus, we prepare for the task at hand with the statements

\[^{14}\text{See an exercise towards the end of the chapter(s) on specific symbol manipulating programs.}\]
IDL> .compile qmlevels
Compile function pro-files.
IDL> upper = fltarr(8)
Prepare vector for roots with upper sign.
IDL> lower = fltarr(9)
Prepare vector for roots with lower sign.

Then, taking the limits as given in Eqs. (14.26) and (14.27), we find the roots in each category with the IDL statements

IDL> upper[0] = lubisect( 'qmupper', -0.5, 0.5 )
IDL> upper[1] = lubisect( 'qmupper', 2.5, 3.5 )
IDL> ...
IDL> upper[7] = lubisect( 'qmupper', 20.5, 21.5 )
IDL> lower[0] = lubisect( 'qmlower', -0.5, 0.5 )
IDL> lower[1] = lubisect( 'qmlower', 3.0, 4.0 )
IDL> ...
IDL> lower[8] = lubisect( 'qmlower', 23.5, 24.5 )

and display them with the statements

IDL> print, upper
0.00000  3.02048  6.54821  9.05418  13.1188  15.0614
  19.7611  20.9943
IDL> print, lower
0.00000  3.27289  6.03920  9.82884  12.0628  16.4248
  18.0433  23.1778  23.8645

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 IDL are

IDL> print, upper*\cos(upper)/\sin(upper) + \sqrt(25.0^2-upper^2)
  NaN -0.000127792  48.2543  6.48499e-05  42.5627  7.62939e-06
  30.6267  2.47955e-05
IDL> print, lower*\cos(lower)/\sin(lower) + \sqrt(25.0^2-lower^2)
  NaN  49.5696 -9.15527e-05  45.9736 -1.90735e-06  37.6949
  -5.72205e-05  18.7393  1.81198e-05

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[2], upper[4], upper[6], lower[1], lower[3], lower[5], and lower[7]. We also reject upper[0] and lower[0]; 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 IDL statements

IDL> s = lower[1:8]
Discard first root in lower.
IDL> for i = 1, 7, 2 do s[i-1]=upper[i]
Replace “bad” roots in s with “good” ones in upper.
IDL> print,s
  20.9943  23.8645

Finally, remembering that we are interested not so much in the values of \( s \) but in the associated energies, we exploit the equation \( E/V_0 = -(1-s^2/c^2) \) to find the allowed energies with the statements

\[^{15}\text{The value ‘NaN’ for the first root in each category arises because of a division by zero in evaluating the function.}\]
14.13 Finding Roots Numerically with Mathematica

When Mathematica is able to find an analytic solution for the roots of an equation of interest, numerical values of the resulting solution can be found simply by embedding the command Solve in the command N in a statement of the form

\[ N[ \text{Solve}[ \text{Eqn}, \text{Var} ] ] \]
Here, \( \text{Eqn} \) stands for the equation to be solved and \( \text{Var} \) represents the variable to be found. If, however, \( \text{Mathematica} \) is unable to find an analytic solution (or we want to go directly to a numerical approach anyway), we must resort to numerical methods. \( \text{Mathematica} \)'s workhorse when the function whose roots are sought is a polynomial is the command \text{NSolve}. The syntax for \text{NSolve} is essentially identical to that of the command \text{Solve}. In its simplest form, we merely invoke the function in a statement like

\[
\text{NSolve}[ \text{Eqn}, \text{Var} ]
\]

though we might also exploit a further argument in a statement of the form

\[
\text{NSolve}[ \text{Eqn}, \text{Var}, \text{NumberOfDigits} ]
\]

to stipulate a precision different from the default. More specifically, the statement

\[
\text{In}[1]:= \text{NSolve}[ 3x^3-2x+x-4==0, x ]
\]
\[
\text{Out}[1]= \{\{x \to -0.600668 - 0.86549i\}, \{x \to -0.600668 + 0.86549i\}, \{x \to 1.20134\}\}
\]

returns numerical values for the three roots of the specified cubic polynomial, including both real and complex roots. Despite its nominal restriction to polynomials, however, \text{NSolve} is able to find roots for some non-polynomial functions. For example, the statement

\[
\text{In}[2]:= \text{NSolve}[ \text{Cos}[x]==0, x ]
\]
\[
\text{Out}[2]= \{\{x \to \text{ConditionalExpression}[1.(-1.5708 + 6.28319C[1]), C[1] \in \text{Integers}]\},
\{x \to \text{ConditionalExpression}[1.(1.5708 + 6.28319C[1]), C[1] \in \text{Integers}]\}\}
\]

reveals that \( \text{Mathematica} \) recognizes the equation \( \cos(x) = 0 \) has many roots and successfully if obscurely reports them all. To obtain any particular roots, we would execute statements like

\[
\text{In}[3]:= \text{Out}[2] \ /\ /. \ C[1] \rightarrow 0
\]
\[
\text{Out}[3]= \{\{x \to -1.5708\}, \{x \to 1.5708\}\}
\]
\[
\text{In}[4]:= \text{Out}[2] \ /\ /. \ C[1] \rightarrow 3
\]
\[
\text{Out}[4]= \{\{x \to 17.2788\}, \{x \to 20.4204\}\}
\]

\( C[1] \) can, of course, be given either positive or negative integer values. Apparently, when \text{NSolve} is presented with a non-polynomial, it reverts to \text{Solve} followed by \text{N}, and the effort will be successful only if the given equation admits an analytic solution.\(^\text{16}\)

When the function is not a polynomial, when it does not possess an analytic solution, or when we seek a numerically determined solution anyway, we must use the command \text{FindRoot}, which has arguments specifying not only the equation to be solved and the variable sought but also one or two starting values for an iterative search. The general syntax is

\[
\text{FindRoot}[ \text{Eqn}, \{ \text{Var}, \text{Start1} \} ]
\]

in which case \text{FindRoot} uses Newton's method, or

\[
\text{FindRoot}[ \text{Eqn}, \{ \text{Var}, \text{Start1}, \text{Start2} \} ]
\]

\(^{16}\text{Some earlier versions of Mathemathica found only two of the many roots of } \cos(x) = 0, \text{ the command making its own judgment about the interval in which to search and offering no means for us to specify some other interval. For the particular equation } \cos(x) = 0, \text{ those versions of Mathemathica also displayed the warning } "\text{Solve::ifun: Inverse functions are being used by Solve, so some solutions may not be found."}
in which case \texttt{FindRoot} uses the secant method.\footnote{Remember that the two starting values given to the secant method are not necessarily lower and upper bounds for the root that will be found. If they do happen to bound a root, however, that is the root that will be found.} \texttt{FindRoot} admits several options, the most significant of which are \texttt{MaxIterations} (default 100) to specify the maximum number of iterations to use before giving up and \texttt{AccuracyGoal} (default normally half of the value of \texttt{WorkingPrecision})\footnote{The default for \texttt{WorkingPrecision} is \$\texttt{MachinePrecision}, often 16.} to stipulate the accuracy with which the function will be zero at the returned root. Thus, for example, we might seek a particular root of \cos(x) with the statement

\begin{verbatim}
In[3]:= FindRoot[ Cos[x] == 0, {x, 4.0} ]
Out[3]= \{ \{x \to 4.71239\} \}
\end{verbatim}

or the statement

\begin{verbatim}
In[4]:= FindRoot[ Cos[x] == 0, {x, 4.0, 5.0 } ]
Out[4]= \{ \{x \to 4.71239\} \}
\end{verbatim}

a value we might recognize as \(3\pi/2\). Properly chosen, the additional stipulations in the second argument permit us to select a particular root from the many that we know \textit{a priori} exist.

Out of curiosity, let us test \texttt{NSolve} with a polynomial that has a multiple root. We execute the statements

\begin{verbatim}
In[5]:= p1 = Expand[ (x-1)^2*(x+3)*(x-4) ];
In[6]:= NSolve[ p1==0, x ]
Out[6]= \{ \{x \to -3\}, \{x \to 1\},
\{x \to 1\}, \{x \to 4\} \}
\end{verbatim}

Clearly, \texttt{NSolve} has returned all four roots of the given polynomial even though one of the three distinct roots is a double root.

### 14.13.1 Classical Turning Points

To find the turning points for the potential energy in Eq. (14.3), we simply define the polynomial and invoke \texttt{NSolve} with the statements

\begin{verbatim}
In[7]:= V = x^3/10000+x^2/200-x/500-1/2;
In[8]:= NSolve[ V == 0, x ]
Out[8]= \{ \{x \to -48.2683\}, \{x \to -11.0804\}, \{x \to 9.3487\} \}
In[9]:= Quit[]
\end{verbatim}

These results are certainly consistent with the bounds identified in Section 14.1.1.

### 14.13.2 Range of Projectile

As laid out in Section 14.1.4, finding the range of a projectile moving in a viscous medium begins with finding the time at which a projectile launched at some angle \(\theta\) in a medium characterized by a (dimensionless) viscous damping coefficient \(\alpha = bv_0/(mg)\) 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). To find that root and then the range for a particular choice of \(\alpha\) and \(\theta\), we begin with the statements

\begin{verbatim}
In[7]:= V = x^3/10000+x^2/200-x/500-1/2;
In[8]:= NSolve[ V == 0, x ]
Out[8]= \{ \{x \to -48.2683\}, \{x \to -11.0804\}, \{x \to 9.3487\} \}
In[9]:= Quit[]
\end{verbatim}
In[1]:= \text{rang} = \text{Cos[Theta]}*(1-\text{Exp[-Tau]})/\text{Alpha}

\text{Out[1]=} \frac{(1-e^{-\tau})\text{Cos}[\theta]}{\alpha}

In[2]:= fct = \text{\[Tau]} - (1+\text{\[Alpha]}*\text{Sin[Theta]})*(1-\text{Exp[-\[Tau]})

\text{Out[2]=} \tau - (1-e^{-\tau})(1 + \alpha \sin \theta)

which define the range given by Eq. (14.19) and the auxiliary function given by Eq. (14.18). Then, we proceed with the statements

\text{In[3]:= } \!\text{\[Alpha]} = 0.4; \!\text{\[Theta]} = \pi \frac{30.0}{180.0}; \!\text{Choose } \alpha, \theta.
\text{In[4]:= } \!\text{FindRoot[ fct == 0, \{ \[Tau], 0.2, 0.5 \} ] } \!\text{Find root.}
\text{Out[4]=} \{ \tau \rightarrow .376438 \}

\text{In[5]:= } \!\text{rang} /. \% \!\text{Evaluate range.}
\text{Out[5]=} 0.679177

This result is expressed in units of $v_0^2/g$.

To find the angle of fire to attain maximum range, we need to repeat the sample calculation of the previous paragraph for values of $\theta$ ranging from 0° to 90°, choosing, say, increments of 1° between consecutive values of $\theta$. To find the range for each of the selected values of $\theta$, we invoke the single loop

\text{In[6]:= } \!\text{For[ i=0, i<91, i++,}
\text{\quad \\[Theta] = \pi \frac{i}{180.0};}
\text{\quad tmp = FindRoot[ fct == 0, \{ \[Tau], 0.005, 0.8 \} ];}
\text{\quad rng[i+1] = rang /. tmp ]}

where we have accepted the default value 100 of Max\text{\textit{Iterations}}. (Remember that $\alpha$ was set to 0.4 in the previous paragraph.) The starting values are set with reference to Fig. 14.4 to prevent Find\text{\textit{Root}} from finding the root $\tau = 0$ (which exists for all values of $\theta$). Unfortunately, that strategy doesn't work for $\theta = 0$, at which value the only root is $\tau = 0$. Even though no warning is presented, Find\text{\textit{Root}} returns an incorrect value for the root at $\theta = 0$, which we must correct with the statement

\text{In[7]:= } \!\text{rng[1] = 0.0;}

after the above loop has terminated. Finally, we produce the graph in Fig. 14.12 with the statements

\text{In[8]:= } \!\text{pts = Table[ \{i, rng[i+1]\}, \{i, 0, 90\};}
\text{In[9]:= } \!\text{ListPlot[ pts, PlotJoined -> True,}
\text{\quad AxesLabel -> \{"theta", "R(theta)"\},}
\text{\quad PlotStyle -> Thickness[0.005] ]}

Further, by noting the three values,

\text{In[10]:= } \!\text{\{ rng[40], rng[41], rng[42] \}}
\text{Out[10]=} \{.725174, .725949, .725889 \}
\text{In[11]:= } \!\text{Quit[]}

we conclude that the maximum range for $\alpha = 0.4$ occurs near\textsuperscript{19} $\theta = 40^\circ$ (or, with parabolic interpolation to find the maximum, $\theta = 40.43^\circ$), rather lower than the 45° angle at which maximum range occurs when $\alpha = 0$ (no viscous damping).

\textsuperscript{19}Remember that the index of the entries in rng is one larger than the angle in degrees to which that entry corresponds.
14.13. FINDING ROOTS NUMERICALLY WITH MATHEMATICA

Figure 14.12: $R(\theta)$ versus $\theta$ for $\alpha = 0.4$.

Had we started with some other value of $\alpha$, of course, statements identical to those in the previous paragraph would produce a graph corresponding to that other value of the viscous damping.

14.13.3 Energy Levels in Quantum Well

Finally, we turn to addressing the energy levels for the finite-depth quantum well discussed in Section 14.1.5, seeking solutions to the equations in Eq. (14.25). In preparation, we first create lists of the lower and upper bounds given in Eqs. (14.26) and (14.27). To that end we execute the statements

\begin{verbatim}
In[1]:= ulb = { -0.5, 2.5, 6.0, 8.5, 12.5, 14.5, 19.0, 20.5 }; 
In[2]:= uub = { 0.5, 3.5, 7.0, 9.5, 13.5, 15.5, 20.0, 21.5 }; 
In[3]:= llb = { -0.5, 3.0, 5.5, 9.5, 11.5, 16.0, 17.5, 22.5, 23.5 }; 
In[4]:= lub = { 0.5, 4.0, 6.5, 10.5, 12.5, 17.0, 18.5, 23.5, 24.5 }; 
\end{verbatim}

Then, we find the roots in each category with the simple loops

\begin{verbatim}
In[5]:= For[ i = 1, i<9, i++,  
    rootu[i] = FindRoot[ Sin[s]-0.04*s, { s, ulb[[i]], uub[[i]]} ] ] 
In[6]:= For[ i = 1, i<10, i++,  
    rootl[i] = FindRoot[ Sin[s]+0.04*s, { s, llb[[i]], lub[[i]]} ] ] 
\end{verbatim}

Next, remembering that the values in rootu[...] and rootl[...] have the form \{s \rightarrow ...\}, we concatenate the elements into lists and display them with the statements

\begin{verbatim}
\end{verbatim}
In[7]:= rtu = s /. Table[ rootu[i], { i, 1, 8 } ]
In[8]:= rtl = s /. Table[ rootl[i], { i, 1, 9 } ]

Remember, however, that physically acceptable solutions must satisfy not Eq. (14.25), which we have solved, but Eq. (14.23). To determine which of these roots are physically meaningful, we substitute each into Eq. (14.23), remembering that \( c = 25 \) for the case we have treated. The appropriate statements to Mathematica are

\[
\text{In[9]:= } f[x_] := x \cdot \cot(x) + \sqrt{25^2 - x^2}
\]

\[
\text{In[10]:= } \text{Map}[f, \text{rtu}]
\]

\[
\]

\[
\text{In[11]:= } \text{Map}[f, \text{rtl}]
\]

\[
\]

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 \( \text{rtu}[[1]], \text{rtu}[[3]], \text{rtu}[[5]], \text{rtu}[[7]], \text{rtl}[[1]], \text{rtl}[[2]], \text{rtl}[[4]], \text{rtl}[[6]] \), and \( \text{rtl}[[8]] \). Thus, we assemble a single list containing the eight physically meaningful roots with the Mathematica statements

\[
\text{In[12]:= } s = \text{Rest}[\text{rtl}];
\]

\[
\text{In[13]:= } \text{For} [ i=1, i<8, i=i+2, \text{ s[[i]] = rtu[[i+1]] } ]
\]

\[
\text{In[14]:= } s
\]

\[
\]

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 statement

\[
\text{In[15]:= } \text{energy} = -(1-s^2/25.0^2)
\]

\[
\text{Out[15]= } \{-0.985403, -0.941645, -0.868835, -0.76718, -0.637047, -0.479105, -0.294784, -0.0887774\}
\]

Then, the statement

\[
\text{In[16]:= } \text{pts} = \text{Table} [ \{0.2, \text{energy}[[i]]\}, \{0.8, \text{energy}[[i]]\} \}, \{i,8\} ]
\]

\[
\text{Out[16]= } \{\{0.2, -0.985403\}, \{0.8, -0.985403\}\}...\}
\]

creates a list, each of whose elements is itself a list of the two points at opposite ends of a single line in an energy level diagram for the system we are studying. Finally, we produce the energy level diagram itself in steps. First, we draw the axes and the lowest energy level with the statement

\[21\]Remember that values close to zero often should be exactly zero but fail in that regard because of internal roundoff. These values may differ from operating system to operating system.
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 \( V_0 \). Among other things, the grid lines show the the energies of the bottom \((-1)\) and the top \((0)\) of the well.

\[
\begin{align*}
\text{E/V0} & \quad 0.5 \\
& \quad 0.25 \\
& \quad 0 \\
& \quad -0.25 \\
& \quad -0.5 \\
& \quad -0.75 \\
& \quad -1 \\
& \quad -1.25
\end{align*}
\]

Then, we create several graphic objects, each containing one of the energy levels, with the statement\(^{22}\)

\[
\text{In}[18]:= \text{For[ i=2, i<9, i++}, \\
\text{lv[i] = Graphics[ { Thickness[0.005], Line[ pts[[i]] ] } ] 
}\]

Finally, we concatenate the graphics objects into a single list and create the display with the statement

\[
\text{In}[19]:= \text{Show[ Table[ lv[i], \{i, 8\} ] ]}
\]

The resulting diagram is shown in Fig. 14.13.

### 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

\(^{22}\)The list presented as an argument to the command Graphics may contain graphics primitives, e.g., Line, and graphics directives, e.g., Thickness. The order is important, since the directive affects only those primitives that follow it in the list.
many contexts, however, the task of finding (one or more) roots will require solving a system of \( n \) simultaneous equations for \( n \) unknowns.\footnote{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.} 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.\footnote{For example, S. D. Conte, \textit{Elementary Numerical Analysis} (McGraw-Hill, New York, 1965), Fornab S. Acton, \textit{Numerical Methods that Work} (Harper and Row, New York, 1970), or William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling, \textit{Numerical Recipes} (Cambridge University Press, Cambridge, 1992), Second Edition, Chapters 2 and 9.}

### 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.,

\[
\begin{pmatrix}
  a_{11} & a_{12} & a_{13} \\
  a_{21} & a_{22} & a_{23} \\
  a_{31} & a_{32} & a_{33}
\end{pmatrix}
\begin{pmatrix}
  x_1 \\
  x_2 \\
  x_3
\end{pmatrix}
= \begin{pmatrix}
  b_1 \\
  b_2 \\
  b_3
\end{pmatrix}
\]  

where, with \( i \) and \( j \) independently assuming the values 1, 2, and 3, the quantities \( x_i \) are the \textit{unknowns}, the quantities \( b_i \) are the \textit{inhomogenieties}, and the quantities \( a_{ij} \) are the \textit{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 inhomogenieties, we would write the equations compactly in the form

\[
A \mathbf{x} = \mathbf{b}
\]

and we would write their solution formally as

\[
\mathbf{x} = A^{-1} \mathbf{b}
\]

Less compactly but more usefully (at least occasionally), we might remember \textit{Cramer’s rule} and write the solution in terms of determinants in the form

\[
\begin{align*}
x_1 &= \frac{\begin{vmatrix}
  b_1 & a_{12} & a_{13} \\
  b_2 & a_{22} & a_{23} \\
  b_3 & a_{32} & a_{33}
\end{vmatrix}}{\begin{vmatrix}
  a_{11} & a_{12} & a_{13} \\
  a_{21} & a_{22} & a_{23} \\
  a_{31} & a_{32} & a_{33}
\end{vmatrix}}, \\
x_2 &= \frac{\begin{vmatrix}
  a_{11} & b_1 & a_{13} \\
  a_{21} & b_2 & a_{23} \\
  a_{31} & b_3 & a_{33}
\end{vmatrix}}{\begin{vmatrix}
  a_{11} & a_{12} & a_{13} \\
  a_{21} & a_{22} & a_{23} \\
  a_{31} & a_{32} & a_{33}
\end{vmatrix}}, \\
x_3 &= \frac{\begin{vmatrix}
  a_{11} & a_{12} & b_1 \\
  a_{21} & a_{22} & b_2 \\
  a_{31} & a_{32} & b_3
\end{vmatrix}}{\begin{vmatrix}
  a_{11} & a_{12} & a_{13} \\
  a_{21} & a_{22} & a_{23} \\
  a_{31} & a_{32} & a_{33}
\end{vmatrix}}
\end{align*}
\]  

In this rule, which can be readily extended to \( n \) equations, the denominator in the expression for the \( i \)-th unknown is the determinant of the coefficient matrix and the numerator is the determinant of the matrix created from the coefficient matrix by replacing its \( i \)-th column with the column of inhomogenieties. Cramer’s rule provides a direct, symbolic, and exact solution to the system of \( n \) simultaneous linear equations. Note, however, that the rule gives problems if the determinant of the coefficient matrix happens to be zero or, equivalently, if the inverse \( A^{-1} \) of the coefficient matrix fails to exist. When \( |A| = 0 \), the equations are said to be \textit{singular} and will either have no solution...
the remaining equations and reorder either the equations (partial pivoting) or the equations and the

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.

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.\(^{27}\) That is, we can write

\[
A = LU \quad \text{where} \quad L = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}
\] (14.45)

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

\[
Ax = b \quad \Rightarrow \quad LUx = b \quad \Rightarrow \quad Ly = b \quad \text{where} \quad y = Ux
\] (14.46)

Since the first equation in the (matrix) equation \( Ly = b \) tells us \( y_1 \) directly, the second tells us \( y_2 \) directly once \( y_1 \) is known, and the third tells us \( y_3 \) directly once \( y_1 \) and \( y_2 \) are known, the equation \( Ly = b \) is readily solved for \( y \). Then, however, once \( y \) is known, a similar process that starts with \( x_3 \), then moves to \( x_2 \) and finally to \( x_1 \) directly solves the equation \( Ux = y \) for the original unknowns in \( x \).

The methods outlined in the previous paragraphs will all work for (almost) any system of linear equations. Sometimes, however, the coefficient matrix may have special properties that can be exploited to simplify the algorithm or—important for large systems—reduce the requirements on memory for storage of matrices and intermediate results. The coefficient matrix may be symmetric \((a_{ij} = a_{ji})\) or it may be sparse (only a small fraction of its elements differing from zero). In the latter category, the matrix may be tridiagonal (non-zero elements only on the main diagonal, on the diagonal immediately above the main diagonal, and on the diagonal immediately below the main diagonal). Special savings in storage can be achieved if the matrix happens to be both symmetric and tridiagonal, since in that case the only elements that need be stored are those on the main diagonal and those on the diagonal immediately above the main diagonal, \(2n-1 \) elements rather than \(n^2 \) elements.

The methods outlined in the previous paragraphs are also all direct methods, i.e., each leads directly to the desired solution in a finite number of steps. Except for roundoff error, each would yield an exact solution to the equations at hand. When the coefficient matrix is sparse, an iterative approach may be computationally more efficient. Such approaches entail finding a means by which an initial guess can be repeatedly refined until some criterion of convergence has been met. Among the most common such procedures applies to Laplace’s equation. As judiciously as possible, we “guess” a solution at a regular grid of points laid over the domain of the problem. Then, to carry out the first iteration, we step systematically through that grid, replacing the value at each grid point with the average of the values at its nearest neighbors. Then, we repeat the process with the results of the first iteration as input, generating the second iteration, and continuing until—say—no value changes by more than some specified tolerance. The only drawback to an iterative method is that we have added the uncertainties associated with convergence to those potentially generated by computer roundoff. In terms of computational labor, however, the sacrifice is often worth the gains.

Finally, we merely mention the more sophisticated approach—singular value decomposition—we must adopt if the system of equations confronting us happens to be nearly singular, in which case

---

the presence of computer roundoff introduces instabilities in the simpler methods. Each of the software packages described in this book makes available a spectrum of commands to solve simultaneous linear equations. We here merely enumerate those commands and indicate the general strategy each implements. We leave the descriptive details to the vendor-supplied documentation, and we leave illustrative applications to the exercises at the end of the chapter.

MATHEMATICA: Within Mathematica, systems of linear equations can be solved analytically either with the special command LinearSolve, which accepts the matrix of coefficients and the vector of inhomogenieties as arguments, or with the command Solve, whose arguments then are lists of the equations and the unknowns. Numerical solutions can be found by nesting either LinearSolve or Solve in the command N. While NSolve cannot handle systems of equations, FindRoot can address systems of equations if we simply make its first argument a list of the equations to be solved, supply as its second argument a list containing the first variable and a starting guess for its value in the solution, supply as its third argument a list containing the second variable and a starting guess for its value in the solution, etc.

IDL: For numerical solution of systems of linear equations, the most commonly used IDL commands are cramer, which uses Cramer's rule; ludc and lusol, which effect the LU decomposition of the coefficient matrix and the solution of the associated system; trisol, which is especially adapted to tridiagonal systems; and svdc and svsol, which effect the singular value decomposition of the coefficient matrix and the solution of the associated system. All of these routines use direct methods.

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

\[ f_1(x_1, x_2) = 0 \quad ; \quad f_2(x_1, x_2) = 0 \] (14.47)

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

\[ f_1(x_1, x_2) = \sin \left( \frac{x_1^2}{20} \right) - \cos \left( \frac{x_2}{5} \right) \quad ; \quad f_2(x_1, x_2) = x_1 \tanh(x_2) - 5.0 \] (14.48)

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

\[ \textit{For a full description of this technique, see William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling, } \textit{Numerical Recipes} \textit{(Cambridge University Press, Cambridge, 1992), Second Edition, Section 2.6.} \]
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

\[ f_1(x_1, x_2, x_3) = 0 \quad f_2(x_1, x_2, x_3) = 0 \quad f_3(x_1, x_2, x_3) = 0 \]  

(14.49)

Suppose, further, that we have examined the equations and determined that there does indeed exist a root in the immediate vicinity of the point \((x_1^{(0)}, x_2^{(0)}, x_3^{(0)})\). We might then suppose that the actual root differs from this guess by a small amount, say,

\[ x_1 = x_1^{(0)} + \delta x_1 \quad x_2 = x_2^{(0)} + \delta x_2 \quad x_3 = x_3^{(0)} + \delta x_3 \]  

(14.50)
and demand that the “corrections” satisfy the equations

\[0 = f_1\left(x_1^{(0)} + \delta x_1, x_2^{(0)} + \delta x_2, x_3^{(0)} + \delta x_3\right)\]

\[0 = f_2\left(x_1^{(0)} + \delta x_1, x_2^{(0)} + \delta x_2, x_3^{(0)} + \delta x_3\right)\]  \hspace{1cm} (14.51)

\[0 = f_3\left(x_1^{(0)} + \delta x_1, x_2^{(0)} + \delta x_2, x_3^{(0)} + \delta x_3\right)\]

These equations are, of course, not really any more tractable as they stand than were the original equations. Because the corrections are all presumed small, however, we should be able to approximate these equations by expanding each in a three dimensional Taylor series. Keeping only the linear terms, we find—at least approximately—that

\[0 = f_1 + \frac{\partial f_1}{\partial x_1} \delta x_1 + \frac{\partial f_1}{\partial x_2} \delta x_2 + \frac{\partial f_1}{\partial x_3} \delta x_3\]

\[0 = f_2 + \frac{\partial f_2}{\partial x_1} \delta x_1 + \frac{\partial f_2}{\partial x_2} \delta x_2 + \frac{\partial f_2}{\partial x_3} \delta x_3\]  \hspace{1cm} (14.52)

\[0 = f_3 + \frac{\partial f_3}{\partial x_1} \delta x_1 + \frac{\partial f_3}{\partial x_2} \delta x_2 + \frac{\partial f_3}{\partial x_3} \delta x_3\]

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.

**MATHEMATICA:** *Mathematica*’s command `Solve` is capable of solving some sets of nonlinear equations analytically. In that case, the first argument to `Solve` becomes a list containing the equations to be solved and the second argument becomes a list containing the variables for which solutions are sought. When `Solve` can find solutions to a system of nonlinear equations, nesting `Solve` in the command `N` will find numerical solutions. Whether `Solve` is successful or not, numerical solutions to systems of nonlinear equations can be found with `FindRoot`. We simply make its first argument a list of the equations to be solved. Then, for each unknown, we add one further argument in the form of a list containing (first) the name of the variable and (second) an initial guess for the value of that variable in the solution.

**IDL:** For numerical solution of systems of nonlinear equations, the most commonly used IDL commands are `newton`, which uses an embellishment of Newton’s method that has improved convergence, and `broyden`, which implements a multidimensional secant method described for one variable in Section 14.6.3.

### 14.17 Exercises

#### 14.17.1... using Symbolic Methods

14.1. Derive both forms of the quadratic formula as given in Eq. (14.29) for the roots of the polynomial \(ax^2 + bx + c\). Hint: Start by completing the square, i.e., by adding and subtracting the right amount so that the polynomial can be expressed in the form \(a(x - \alpha)^2 + \beta\).
14.2. Find the value of \( x \) at which \( f(x) = ax^2 + bx + c \) has an extremum and determine a criterion involving the coefficients (or some of them) for deciding whether the extremum is a maximum or a minimum. Assume that \( a, b, \) and \( c \) are real.

14.3. Find the points at which the function \( f(x) = ax^3 + bx^2 + cx + d \) has (local) extrema and find a criterion involving the coefficients (or some of them) that will assure that the function has three real roots. Assume that \( a, b, c, \) and \( d \) are real.

14.4. In some quantum calculations, the need to solve the equation \( x(x + 1) = l(l + 1) \) for \( x \) arises. Find those roots, noting particularly that, since the equation is quadratic, there are two roots. The obvious root \( x = l \) is not the only one.

14.5. Each of the three blades of a lawn mower has radius \( a \). As shown in Fig. 14.15, the center blade is invariably mounted somewhat in front of the two outside ones so that the areas cut by each blade can overlap without risking collision of the blades with one another. What must be the minimum offset \( x \) of the center of the middle blade from the line joining the centers of the two outer blades so that their cutting paths will overlap by an amount \( y \) without collision of the blades?

14.6. For the projectile discussed in Section 14.1.4, generate a family of graphs showing \( \tau(\theta) \) as a function of \( \theta \) for selected fixed values of \( \alpha \).

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 \( \alpha = bv_0/(mg) \), 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

\[
\begin{align*}
md_{x_1}^2 t^2 &= -kx_1 + k(x_2 - x_1) \\
m'd_{x_2}^2 t^2 &= -k(x_2 - x_1) + k(x_3 - x_2) \\
m'd_{x_3}^2 t^2 &= -k(x_3 - x_2) - kx_3
\end{align*}
\]

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 ... 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

\[
\begin{align*}
L_0(x) &= 1 \\
L_1(x) &= x \\
L_2(x) &= \frac{1}{2}(3x^2 - 1) \\
L_3(x) &= \frac{1}{2}(5x^4 - 30x^2 + 3) \\
L_4(x) &= \frac{1}{4}(63x^5 - 315x^3 + 105x) \\
L_5(x) &= \frac{1}{8}(231x^6 - 315x^4 + 105x^2 - 5)
\end{align*}
\]
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, 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

\[
\cot \alpha = \pm \tanh \alpha
\]

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, The Physics of Musical Instruments, Second Edition (Springer-Verlag, New York, 1991), pp. 57ff.]

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 \( c \), which is determined by the depth and the width of the well, increases. At base, changing \( c \) changes the slopes of the straight lines in Fig. 14.5. As \( c \) increases and the well becomes deeper, the lines become more nearly horizontal and the number of energy levels increases. Seek ultimately to generate a graph that shows the energy of each allowed level on the vertical axis as a function of the parameter \( c \) along the horizontal axis.

14.18. The intensity \( I(x) \) in the diffraction pattern produced by a single slit is given by

\[
\frac{I(x)}{I_0} = \frac{\sin^2 x}{x^2}
\]

where \( I_0 \) is the intensity in the center and \( x \) is related to the position of the observation point away from the central maximum. The zeroes in this pattern are easy to locate (they occur at \( x = n\pi \), \( n = 0, \pm 1, \pm 2, \ldots \)). Careful location of the maxima, however, is more complicated. They don’t occur where \( \sin^2 x = 1 \) because of the influence of the denominator that steadily increases as \( x \) increases. Locate the positions of the first half dozen maxima in this pattern, which—basically—is a request to find the roots of the derivative of the function (though note that not all roots correspond to maxima). Use at least three different methods and at least two different computational tools, and compare the results. Do your results confirm that the roots approach odd multiples of \( \frac{1}{2}\pi \) as they become large? Optional: You might also find it interesting to approximate the function with a power series expansion for \( \sin x \), keeping quite a few terms but converting the root finding problem into that of finding the roots of a polynomial. Then, use methods for finding roots of polynomials and see if you can come to understand how accuracy depends on how many terms you keep and which root you seek.

14.19. Using at least three different methods and at least two different computational tools, find the first half dozen roots of the zeroth-order Bessel function \( J_0(x) \). Note that these roots are related to the radii of circular nodes in some of the vibrations of a circular membrane. The values of
these roots tabulated in Abramowitz and Stegun\textsuperscript{29} 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

\[ V(x) = \frac{-V_0 a^2 (a^2 + x^2)}{8a^4 + x^4} \]

where \( x = x/a \). Using at least three different methods and at least two different computational tools, find the coordinates \( x \) of all turning points when the total energy \( E \) of the particle is \( E = -0.2V_0 \) and also when the total energy is \( E = -0.1V_0 \). *Optional:* Obtain graphs of the position of each turning point as a function of particle energy over the allowed range of energies for bound states.

14.21. Generate a graph showing the turning points of the potential energy given by Eq. (14.3) as a function of the energy of the particle.

14.22. Suppose a straight railroad track of length 1 mile (5280 ft) is held absolutely immovable at its two ends. On a hot summer day, the track expands in length by 1 ft. If the track bows upward from the earth in a circular arc, how high above the earth will the track rise at its midpoint? *Hint:* The geometry of this exercise is shown in the accompanying figure, where \( l \) is the original length of the track, \( d \) is the rise at the center, and \( a \) is the length, \( R \) the radius, and \( \theta \) half the subtended angle of the circular arc. Thus, \( R\theta = a/2, \sin \theta = l/(2R) \), and \( R - d = R \cos \theta \). The task is to find \( \theta \) and \( R \) from the first two of these equations and then use the third to find \( d \). The only quantities known \textit{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 \textit{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

\[ \frac{I(\theta)}{I_0} = \left[ \frac{2J_1(\pi d \sin \theta / \lambda)}{\pi d \sin \theta / \lambda} \right]^2 \]

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.

14.28. The Lennard-Jones potential energy \( V_{LJ} \), which is given in terms of the coordinate \( r \) by the expression

\[
V_{LJ} = 4\epsilon \left( \frac{\sigma^{12}}{r^{12}} - \frac{\sigma^{6}}{r^{6}} \right)
\]

or

\[
\frac{V_{LJ}}{\epsilon} = 4 \left[ \frac{1}{r^{12}} - \frac{1}{r^{6}} \right]
\]

where \( \epsilon \) and \( \sigma \) are constants and \( \bar{r} = r/\sigma \), plays a prominent role in some theories of chemical bonding. Obtain a graph of this potential energy and then obtain graphs of the lower turning point and the upper turning point as functions of the energy of the system over the range of energies for which the particle experiencing the potential energy is bound in the associated potential well.

14.17.4 Finding More than One Unknown

14.33. In a global positioning system, the raw data from which the position is determined consists of distances from various reference points together with knowledge of the location of those reference points. In two dimensions, for example, we might try to locate our position \((x, y)\) in a plane from knowledge that we are a distance \( r_1 \) from the point \((x_1, y_1)\) and a distance \( r_2 \) from the point \((x_2, y_2)\). 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 even if points 1 and 2 are, in fact, 100 miles apart. Depending on the circumstances, there may be no points, one point, or—most often—two. Develop an algorithm for finding your location in two dimensions when you know your distance from each of two reference points whose coordinates are known, implement your algorithm in a program using whatever computational tool seems appropriate, solve two or three test problems that you invent, and—in particular—try to describe and defend the conditions under which two, one, or no solutions exist. Optional: Extend your entire consideration into three dimensions, which will require knowledge of distance from each of (at least) three known locations. Hint: Quite a bit of information is available on the website www.trimble.com gps. There is also an article in the January, 1994, issue of Physics Today (“Where I Stand” by Daniel Kleppner, page 9).

14.34. Given the three points \((x_i, y_i)\), \(i = 1, 2, 3\), (a) find symbolic expressions for the coefficients \(a\), \(b\) and \(c\) of the parabola \(y = ax^2 + bx + c\) that passes through these three points and then (b) find a symbolic expression for the value of \(x\) at which the extremum point of the parabola occurs. Finally, (c) determine numerically the angle at which the maximum range of a projectile occurs if the ranges at \(\theta = 30^\circ\), \(40^\circ\), and \(45^\circ\) 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 \(R_{eff} = V/I\) 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 $\texttt{HEAD/data/freefall.dat}$ contains 31 lines, the \(i\)-th of which contains one numerical value—the value of the position \(x_i\) in cm of a particle at time \(t_i = (i - 1)/60\) s. You have reason to believe that the set of data \((t_i, x_i)\) with \(i = 1, 2, 3, \ldots, n\) is described by the parabolic relationship

\[x = at^2 + bt + c\]

The method of least squares identifies the optimum values of the coefficients \(a\), \(b\), and \(c\) as the particular values that minimize the residual

\[R(a, b, c) = \sum_{i=1}^{n} \left( x_i - (at_i^2 + bt_i + c) \right)^2\]
Using whatever language you choose, write a program that reads the positions from the file into a vector, generates a second vector containing the corresponding times and then enters a loop in which it asks for entry of trial values of \( a \), \( b \), and \( c \), calculates and displays \( R(a, b, c) \), and returns to ask for a new set of trial values. Using this program, conduct a manual search for the values of \( a \), \( b \), and \( c \) which make \( R(a, b, c) \) as small as possible. Hint: Think carefully about the initial guesses for \( a \), \( b \), and \( c \). A graph of \( x \) versus \( t \) may be useful.

**14.37.** The file $\text{HEAD/data/freefall.dat}$ contains 31 lines, the \( i \)-th of which contains one numerical value—the value of the position \( x_i \) in cm of a particle at time \( t_i = (i - 1)/60 \text{ s} \). You have reason to believe that the set of data \((t_i, x_i)\) with \( i = 1, 2, 3, \ldots, n \) is described by the parabolic relationship

\[
x = at^2 + bt + c
\]

The method of least squares identifies the optimum values of the coefficients \( a \), \( b \), and \( c \) as the particular values that minimize the residual

\[
R(a, b, c) = \sum_{i=1}^{n} \left(x_i - (at_i^2 + bt_i + c)\right)^2
\]

i.e., as the particular values satisfying the three equations

\[
\frac{\partial R}{\partial a} = 0 ; \quad \frac{\partial R}{\partial b} = 0 ; \quad \frac{\partial R}{\partial c} = 0
\]

These equations will turn out to be linear in \( a \), \( b \), and \( c \), with coefficients and inhomogenieties 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 inhomogeneties 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

\[
m \frac{d^2x}{dt^2} = F(t) , \quad x(0) = 0 , \quad x(t_f) = a
\]

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

\[
x_0 = 0 , \quad x_n = a
\]

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.,

\[
V_{12} = \frac{1}{4} \left( V_{11} + V_7 + V_{13} + V_{17} \right) ; \quad V_{13} = \frac{1}{4} \left( V_{14} + V_{10} + 0.0 + V_{20} \right)
\]

(a) Write out the twenty-five equations implied by this requirement, (b) cast them in the matrix form

\[
\begin{pmatrix}
? & ? & ? & \ldots & ? \\
? & ? & ? & \ldots & ? \\
? & ? & ? & \ldots & ? \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
? & ? & ? & \ldots & ?
\end{pmatrix}
\begin{pmatrix}
V_1 \\
V_2 \\
V_3 \\
\vdots \\
V_{25}
\end{pmatrix}
= \begin{pmatrix}
? \\
? \\
? \\
\vdots \\
?
\end{pmatrix}
\]
figuring out the value represented by each question mark, and (c) solve the system for the unknowns. Then (d) structure the boundary values and solution in a $7 \times 7$ matrix and create a surface plot of that solution.

14.40. In Ruchardt’s experiment, a steel ball bounces up and down in a vertical tube that ends in a gallon (or larger) jug. Ultimately, the ball falls into the jug, but it may bounce up and down many times before doing so. The file $\$\text{HEAD/data/ruchardt.dat}$ contains 170 lines, the $i$-th of which contains one numerical value—the position $x_i$ in cm of the steel ball at time $t_i = 0.05(i - 1)$ s. A quick graph of the data (you should make it) suggests that the motion might be described by an exponentially decaying cosine curve on which is superimposed a linear sinking of the “equilibrium” position, i.e., by a function of the form

$$x(t) = Ae^{-bt} \cos(\omega t + \phi) + Ct + D$$

where the parameters $b$, $A$, $\omega$, $\phi$, $C$, and $D$ are to be determined. Develop a means to find optimal values of these six parameters by seeking values that minimize the quantity

$$R = \sum_{i=1}^{N} \left( x_i - Ae^{-bt_i} \cos(\omega t_i + \phi) + Ct_i + D \right)^2$$

Warning: This exercise is not for the faint-hearted. Some useful background will be found in Chapter 8 of Data Reduction and Error Analysis for the Physical Sciences (Second Edition) by Philip R. Bevington and D. Keith Robinson (McGraw-Hill, New York, 1992).
14.A Listing of IDL Procedure lubisect.pro

function lubisect, func, xlb, xub, tol=tol, itmax=itmax

if not keyword_set(tol) then tol=1.0e-6
if not keyword_set(itmax) then itmax=20

fl = call_function( func, xlb )
fu = call_function( func, xub )

if fl*fu gt 0.0 then begin
  print, 'The given interval does not contain a root.'
  return, xlb
endif

i = 0
repeat begin
  xm = 0.5*(xlb + xub)
  fm = call_function( func, xm )
  if fl*fm ge 0 then xlb = xm else xub = xm
  i = i + 1
endrep until (xub-xlb le tol) or (fm eq 0.0) or (i ge itmax)

return, 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 AEs, (or, in some cases, by a system of ODEs), which is then solved for the dependent variable at each node.

In finite element methods (FEMs), however, the nodes forming the grid may be non-uniformly spaced. Collections of these nodes form geometric shapes—lines in one dimension; triangles or quadrilaterals or ... in two dimensions; tetrahedrons or hexahedrons (sometimes called bricks) or ... in three dimensions—that divide the region of interest into subregions, i.e., into the (finite) elements that give the method its name. Interpolation or shape functions are defined to facilitate approximating the dependent variable at points within each element from presumed known values of the dependent variable at the nodes or vertices associated with that element. The differential equation is then replaced with an equivalent integral statement, which is in turn converted into a system of algebraic equations by substituting the shape functions into this integral form, integrating, assembling the results from all elements, and imposing the constraints dictated by the boundary and initial conditions. Thus, as with the finite difference approach, the original partial differential equation(s) is (are) replaced by a system of algebraic equations, which is then solved for the dependent variable at each node.

Finite element methods have significant advantages over finite difference methods. For example, since finite element methods allow a nonuniform mesh, portions of the region in which a solution is sought can be treated with a fine mesh while, at the same time, other portions of that region can be represented with a coarse mesh, so the computational effort can be focused where it is most needed. Furthermore, especially if the boundaries are irregular, the boundary conditions are more easily incorporated with finite element methods than with finite difference methods. While we gain considerably in generality by invoking a method of this second type, we pay a heavy price: finite element methods are difficult to describe and even more difficult to code.

In Section 15.1, we deduce and provide physical contexts for a number of important PDEs. Then, in subsequent sections, we introduce both finite difference and finite element methods, illustrating each with explicit solutions to representative problems. For the sake of the quickest exposition of the essential ideas, we start by addressing problems with one independent variable before moving on to problems having two or more independent variables and requiring coordinate systems other than Cartesian. Because the methods are elaborate and their coding in computer languages is involved and lengthy, this chapter will undoubtedly seem more mathematical and abstract than physical. However obscured it may appear to be by the detailed discussion of method and implementation, the desire to address physical situations does assuredly lie underneath the entire exposition.
15.1. Sample Problems

We begin by laying out several of the most common PDEs of mathematical physics, placing each in at least some of the physical contexts in which it appears. As we shall discover in Section 15.1.5, the most common equations fall into one of three categories, represented respectively by the (classical) wave equation, the diffusion equation, and the Laplace equation.

15.1.1 Motion of a String: The Wave Equation

Consider a string that extends along the (horizontal) $x$ axis when it is in its equilibrium position, and let a point of the string when the string is in that position be located at $x$. Suppose that the string is moving in a plane (not the most complicated possible motion) such that, at a general time $t$, the point nominally located at $x$ is displaced transversely (i.e., perpendicular to the equilibrium orientation of the string)\(^1\) by an amount $u(x,t)$ and longitudinally (i.e., parallel to that equilibrium orientation) by an amount $v(x,t)$. The geometry is shown in Fig. 15.1. The string is, of course, under tension $\tau(x,t)$, which in general will vary from point to point and from time to time but, for a perfectly flexible string (which we explicitly assume), will always be directed tangent to the string. A force diagram for an isolated element—the element located between $x$ and $x + \Delta x$ when the string is in equilibrium—of the string might look like Fig. 15.2. For the transverse motion of the string, Newton’s second law $F = ma$ then requires that

$$\rho(x) \Delta x \frac{\partial^2 u}{\partial t^2} = \tau(x + \Delta x, t) \sin(\theta(x + \Delta x, t)) - \tau(x, t) \sin(\theta(x, t)) - \rho(x) \Delta x g$$

(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)

\(^1\)We take this transverse direction to be vertical, and include gravity among the forces acting on the string.
Figure 15.3: Geometry for determining $\theta(x, t)$ from $u(x, t)$ and $v(x, t)$.

Here, $\rho(x) \Delta x$ is the mass of the element, and $\rho(x) g \Delta x$—the gravitational force on the element when the string is near the surface of the earth—illustrates a possible external force on the string. Similarly, for the longitudinal motion of the string, Newton’s second law requires that

$$\rho(x) \Delta x \frac{\partial^2 v}{\partial t^2} = \tau(x + \Delta x, t) \cos \theta(x + \Delta x, t) - \tau(x, t) \cos \theta(x, t)$$

or that

$$\rho(x) \frac{\partial^2 v}{\partial t^2} = \frac{\partial}{\partial x} \left( \tau(x, t) \cos \theta(x, t) \right)$$

for the horizontal motion.

This set of equations still contains too many unknowns. We must eliminate $\theta(x, t)$, because we really want to find only $u(x, t)$ and $v(x, t)$. Figure 15.3 supports the conclusion that

$$\sin \theta = \frac{\Delta u}{\Delta s} = \frac{\Delta u}{\sqrt{\Delta u^2 + (\Delta x + \Delta v)^2}}$$

$$= \sqrt{\left( \frac{\partial u}{\partial x} \right)^2 + \left[1 + \frac{\partial v}{\partial x} \right]^2}^{1/2}$$

and that

$$\cos \theta = \frac{\Delta x + \Delta v}{\Delta s} = \frac{1 + \frac{\partial v}{\partial x}}{\sqrt{\left( \frac{\partial u}{\partial x} \right)^2 + \left[1 + \frac{\partial v}{\partial x} \right]^2}^{1/2}}$$

Thus, the equations describing the motion of the string are

$$\rho(x) \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x} \left[ \tau(x, t) \frac{\partial u}{\partial x} \right] - \rho g$$

$$= \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left[1 + \frac{\partial v}{\partial x} \right]^2 \right]^{1/2}$$

$$\rho(x) \frac{\partial^2 v}{\partial t^2} = \frac{\partial}{\partial x} \left[ \tau(x, t) \left(1 + \frac{\partial v}{\partial x} \right) \right]$$

$$= \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left[1 + \frac{\partial v}{\partial x} \right]^2 \right]^{1/2}$$

(15.7)

(15.8)
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

\[ u(x, 0) \quad , \quad v(x, 0) \quad , \quad \frac{\partial u}{\partial x}(x, 0) \quad , \quad \frac{\partial v}{\partial x}(x, 0) \quad (15.9) \]

but also appropriate boundary conditions on the string at its two ends and a connection between \( \tau \) on the one hand and \( u \) and \( v \) on the other. This system is decidedly non-linear and very difficult to solve.

Approximations are almost always necessary to turn the problem we would really like to solve into one that we can solve. If, for example, (1) the amplitude of the motion is small so that

\[ \Delta u \ll \Delta x \quad \Rightarrow \quad \frac{\partial u}{\partial x} \ll 1 \quad (15.10) \]

(2) \( \tau \) is sufficiently large that, in small amplitude motion, \( \tau \) remains essentially constant, and (3) the motion is transverse so that \( v = 0 \) everywhere and always, then the equation for \( v \) is automatically satisfied and the equation for \( u \) decouples from that for \( v \) and becomes the (inhomogeneous) wave equation,

\[ \rho \frac{\partial^2 u}{\partial t^2} = \tau \frac{\partial^2 u}{\partial x^2} - \rho g \quad (15.11) \]

which is correct under the given circumstances even if \( \rho \) varies with \( x \).

We have derived this result for motion of a string in one dimension. Two special cases are worth noting:

- If the situation is static so there is no time dependence (\( \partial u/\partial t = 0 \)), then the shape of a string hanging under its own weight is given by the solution to the equation

\[ \tau \frac{d^2 u}{dx^2} = \rho g \quad (15.12) \]

which reduces in this case to an ODE.

- If there is no outside force and \( \rho \) is constant, then Eq. (15.11) becomes

\[ \frac{\rho \partial^2 u}{\partial t^2} = \tau \frac{\partial^2 u}{\partial x^2} \quad \Rightarrow \quad \frac{\partial^2 u}{\partial t^2} = \frac{\tau}{\rho} \frac{\partial^2 u}{\partial x^2} \quad \Rightarrow \quad \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \quad (15.13) \]

Note that the constant \( c = \sqrt{\tau/\rho} \) has the units of velocity.\(^2\) It turns out—though we don’t know this as yet—that \( c \) is the speed of propagation of the wave conveyed by the solution to this equation.

In both cases, of course, appropriate boundary and initial conditions must be specified before the equation has a unique solution.

While its deduction is more complicated, Eq. (15.13) has a natural extension to two and three dimensions, in which case we would consider, for example, the displacement of a two-dimensional membrane or the pressure in a gas in a three-dimensional enclosure. The function we seek would then be \( u = u(x, y, z, t) = u(r, t) \) and the second derivative with respect to one spatial dimension becomes the Laplacian. The equation in that case is

\[ \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \nabla^2 u \quad \text{or} \quad \left( \nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) = 0 \quad (15.14) \]

\(^2\)Since \( [\tau] = \text{N} = \text{kg} \cdot \text{m/s}^2 \) and \( [\rho] = \text{kg/m} \), \( [\tau/\rho] = (\text{kg} \cdot \text{m/s}^2)/\text{(kg/m)} = \text{m}^2/\text{s}^2 \) (Here, the symbol \([\ldots]\) stands for the units of \ldots)
Here

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$  \hspace{1cm} (15.15)$$

$$= \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2}{\partial \phi^2} + \frac{\partial^2}{\partial z^2}$$  \hspace{1cm} (15.16)$$

$$= \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2}{\partial \phi^2}$$  \hspace{1cm} (15.17)$$

where \((x, y, z), (r, \phi, z), \text{ and } (r, \theta, \phi)\) are the physicists’ standard Cartesian, cylindrical, and spherical coordinates. We thus arrive at the (classical) wave equation, one of the three prototype equations in mathematical physics.

**15.1.2 Heat Flow: The Diffusion Equation**

To motivate a second of the important equations, we seek the temperature \(u(x, t)\) in the one-dimensional rod shown in Fig. 15.4. Let the rod be insulated along its sides and characterized by a thermal conductivity \(K\), which gives the rate of heat flow per unit area in the direction of increasing \(x\) by the expression

\[
\text{rate of heat flow in rod towards positive } x \text{ through cross section at right angles to rod} = -KA \frac{\partial u}{\partial x}
\]

where \(A\) is the area of a cross section of the rod. [In effect, Eq. (15.18) defines \(K\).] Here, the derivative \(\partial u/\partial x\) is the gradient of the temperature, and the minus sign appears because heat flows in the *positive* direction when the gradient is *negative* and in the *negative* direction when the gradient is *positive*, i.e., heat flows from regions of higher temperature to regions of lower temperature; the explicit minus sign then assures that \(K\) will be a positive quantity. With this definition of a material property, which may depend on position in the rod, we conclude that, as heat flows in the rod,

\[
\text{heat flow in time } \Delta t \text{ into shaded element across surface at } x = -KA \left. \frac{\partial u}{\partial x} \right|_x \Delta t
\]

and

\[
\text{heat flow in time } \Delta t \text{ into shaded element across surface at } x + \Delta x = +KA \left. \frac{\partial u}{\partial x} \right|_{x+\Delta x} \Delta t
\]

so that, in the end,

\[
\text{net heat conducted into shaded element in time } \Delta t = \left[ \left( KA \frac{\partial u}{\partial x} \right)_{x+\Delta x} - \left( KA \frac{\partial u}{\partial x} \right)_x \right] \Delta t
\]
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

\[
\text{heat necessary to increase temperature of shaded element by amount } \Delta u = (A \Delta x \rho c) \Delta u \quad (15.22)
\]

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

\[
\rho c \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left( K \frac{\partial u}{\partial x} \right) \quad (15.23)
\]

or, if \( K \) happens to be constant, that

\[
\frac{\rho c}{K} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad \Rightarrow \quad \frac{\partial^2 u}{\partial x^2} = \frac{1}{\alpha^2} \frac{\partial u}{\partial t} \quad (15.24)
\]

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

\[
\frac{\partial}{\partial x} \left( K \frac{\partial u}{\partial x} \right) = 0 \quad \text{or (} K \text{ constant)} \quad \frac{\partial^2 u}{\partial x^2} = 0 \quad (15.25)
\]

with, of course, appropriate boundary values at the ends of the rod.

While its deduction is more complicated, Eq. (15.25) has a natural extension to two and three
dimensions, in which case we would consider, for example, the evolution of the temperature distribution
in a plate or in a three-dimensional object when the initial temperature throughout the object
is given and either Dirichlet or Neumann or mixed boundary conditions are specified at all points
on its boundary. We would then seek the function \( u(x,y,z,t) = u(r,t) \) giving the temperature at
all points in space and time, and the fundamental equation becomes

\[
\nabla \cdot \left( K \nabla u \right) = \rho c \frac{\partial u}{\partial t} \quad \text{or (} K \text{ constant)} \quad \nabla^2 u = \frac{1}{\alpha^2} \frac{\partial u}{\partial t} \quad (15.26)
\]

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

\[
\nabla^2 u = 0 \quad (15.27)
\]

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 \( E \) support the argument

\[
\nabla \times E = 0 \quad \Rightarrow \quad E = -\nabla V \quad \Rightarrow \quad \nabla \cdot (\epsilon \nabla V) = -\rho \quad (15.28)
\]
leading to an equation that determines the electrostatic potential $V$ from the charge density $\rho$ and the dielectric permittivity $\epsilon$. If $\epsilon$ is constant, the equation becomes the Poisson equation,

$$\nabla^2 V = -\frac{\rho}{\epsilon}$$ (15.29)

If, in addition, $\rho = 0$, the equation becomes the Laplace equation

$$\nabla^2 V = 0$$ (15.30)

for the electrostatic potential.

For another example, suppose we seek a sinusoidal solution to the wave equation, i.e., suppose we seek a solution to Eq. (15.14) of the form

$$u(x, y, z, t) = \psi(x, y, z) \cos \omega t$$

Then, if $u$ satisfies the wave equation and we set $k^2 = \omega^2/c^2$, $\psi$ will satisfy the equation

$$\nabla^2 \psi + k^2 \psi = 0$$

an equation called the Helmholtz equation. In one dimension, the Helmholtz equation becomes

$$\frac{d^2 \psi}{dx^2} + k^2 \psi = 0$$ (15.31)

which describes standing waves in a string and many other physical situations.

We need not limit our enumeration to equations of importance in classical physics. For example, the time-dependent Schrödinger equation,

$$-\frac{\hbar^2}{2m} \nabla^2 \psi + V \psi = i\hbar \frac{\partial \psi}{\partial t}$$ (15.32)

for the quantum wave function $\psi$ of a particle of mass $m$ in a potential energy $V$ is in some sense a diffusion equation (second order in the space derivatives, first order in the time derivative), though the term $V \psi$ and the imaginary unit $i = \sqrt{-1}$ create significant differences between the two equations.

Beyond Maxwell’s equations and the Schrödinger equation, second-order PDEs can be found in the Dirac equation (relativistic quantum mechanics), the equations of fluid dynamics (the Navier-Stokes’ equations; see Section 15.1.6), the equations of magnetohydrodynamics (MHD, which combine Maxwell’s equations and the Navier-Stokes’ equations), and in many other contexts in classical and contemporary physics.

### 15.1.5 Classification of Second-Order PDEs

In general terms, we have arrived in all cases at second-order, linear, often homogeneous PDEs, equations that we might collectively denote by the symbolism

$$\mathcal{L}\{u\} = 0$$ (15.33)

where the operator $\mathcal{L}$ symbolizes the (linear) differential operator that defines the equation. The most general example in two variables would have the form

$$A(x, y) \frac{\partial^2 u}{\partial x^2} + 2B(x, y) \frac{\partial^2 u}{\partial x \partial y} + C(x, y) \frac{\partial^2 u}{\partial y^2} + D(x, y) \frac{\partial u}{\partial x} + E(x, y) \frac{\partial u}{\partial y} + F(x, y) u = 0$$ (15.34)

where $A, B, C, D, E,$ and $F$ may depend on $x$ and $y$ but do not depend on $u$ or its derivatives. Many, many important equations (though, significantly, not all) fall into this category. The solutions
of those equations that are in this category exhibit the very important property of superposition, namely
\[ L\{u_1\} = 0, L\{u_2\} = 0 \implies L\{au_1 + bu_2\} = aL\{u_1\} + bL\{u_2\} = 0 \] (15.35)
i.e., any linear combination of two solutions is itself a solution of the differential equation.

As it turns out, second-order, linear PDEs in two variables fall into three distinct categories, depending on the algebraic sign of the quantity \( AC - B^2 \). If, for example, \( AC - B^2 > 0 \), the equation is said to be elliptic; if \( AC - B^2 = 0 \), the equation is parabolic; and if \( AC - B^2 < 0 \), the equation is hyperbolic. When the coefficients actually depend on \( x \) and/or \( y \), an equation may fall into one category for some portions of the region of interest and another category for other portions of that region. Each of the prototype equations we have identified above, however, falls cleanly into one of the three categories. For the wave equation, we note that
\[ \frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = 0 \implies A = 1, B = 0, C = \frac{1}{c^2} \implies AC - B^2 = -\frac{1}{c^2} < 0 \] (15.36)
and conclude that the wave equation is hyperbolic. For the diffusion equation,
\[ \frac{\partial^2 u}{\partial x^2} - \frac{1}{\alpha^2} \frac{\partial u}{\partial t} = 0 \implies A = 1, B = C = 0 \implies AC - B^2 = 0 \] (15.37)
and the diffusion equation is parabolic. Finally, for the Laplace equation,
\[ \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \implies A = C = 1, B = 0 \implies AC - B^2 = 1 \] (15.38)
and the Laplace equation is elliptic. In focusing on these three equations, we will therefore be exhibiting techniques that would be applicable to all possible second-order, linear, homogeneous PDEs in two variables.

Note, incidentally, that the variable transformation \( t \rightarrow iy \) will convert the wave equation (hyperbolic) into the Laplace equation (elliptic) and remember that this transformation also converts trigonometric functions into hyperbolic functions. Within the framework of complex variable theory, we might therefore expect that solutions to the wave equation and solutions to the Laplace equation could be related to one another (though we shall not pursue that connection in this book).

15.1.6 Equations of Fluid Dynamics: Navier-Stokes’ Equations

Let us next work out some of the fundamental equations of fluid dynamics. The state of a fluid at time \( t \) is described by a number of functions, the most important of which are the (vector) velocity field \( \mathbf{v}(\mathbf{r}, t) \), the (scalar) pressure \( p(\mathbf{r}, t) \), and the (scalar) density \( \rho(\mathbf{r}, t) \). In general, all of these quantities are functions of position \( \mathbf{r} \) within the fluid and of time \( t \). Further, \( p \) and \( \rho \) are related by the equation of state of the fluid involved.

Consider a volume—see Fig. 15.5—in the shape of a rectangular parallelepiped with its faces parallel to the coordinate planes. Let the lower left back corner be at \((x, y, z)\) and the upper right front corner be at \((x + \Delta x, y + \Delta y, z + \Delta z)\). We regard the volume as fixed in space, and the fluid as flowing through the volume in a way described by the velocity field already introduced. Further, the fluid in this volume experiences (internal) forces from its contact with the rest of the fluid at the surface of the volume and may also experience (external) forces from things like a nearby earth.\(^3\) The basic equations for fluid motion reflect the conservation of mass (the continuity equation), Newton’s second law, and conservation of linear and angular momentum. We deduce here only the first two of these fundamental relationships.

\(^3\)We assume the fluid is not electrically charged, and we ignore internal forces arising from the gravitational attraction of one portion of the fluid for another. Thus, the only internal forces on one element of the fluid will arise from its direct contact with contiguous elements.
15.1.6.1 Conservation of Mass: The Equation of Continuity

In the time interval from \( t \) to \( t + \Delta t \), the change in the mass of the fluid in the volume can be calculated in two different ways. First, we focus on the density, concluding that

\[
\text{mass added during interval } t \text{ to } t + \Delta t = \left[ \rho(t + \Delta t) - \rho(t) \right] \Delta x \Delta y \Delta z \approx \frac{\partial \rho}{\partial t} \Delta t \Delta x \Delta y \Delta z \quad (15.39)
\]

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 in time \( \Delta t \) over these two surfaces by the fluid flow is given by

\[
\text{net mass transported into volume through sides parallel to } yz \text{ plane in time } t \text{ to } t + \Delta t = \rho(x) v_x(x) \Delta t \Delta y \Delta z - \rho(x + \Delta x) v_x(x + \Delta x) \Delta t \Delta y \Delta z = - \left( \frac{\partial (\rho v_x)}{\partial x} \right) \Delta t \Delta x \Delta y \Delta z \quad (15.40)
\]

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

\[
\text{net mass transported into volume through all sides in time } t \text{ to } t + \Delta t = - \nabla \cdot (\rho v) \Delta t \Delta x \Delta y \Delta z = - \nabla \cdot (\rho v) \Delta t \Delta x \Delta y \Delta z
\]

(15.41)
Figure 15.6: Stresses on the front (dotted) surface of an element in a fluid. (Be sure you visualize the intersection of the three vectors to lie on that front surface.) The signs are defined to give the stresses exerted on an element by the next element in the direction in which the coordinate increases.

The values obtained in Eqs. (15.39) and (15.41) for the mass added in the chosen volume in the time $\Delta t$ must, of course, be the same, and we arrive at the equation of continuity,

$$\nabla \cdot (\rho \mathbf{v}) + \frac{\partial \rho}{\partial t} = 0$$

(15.42)

which expresses the conservation of mass, i.e., the conviction that—however the fluid flows—any change in the mass in a fixed volume can come only by the flow of matter through the surface bounding that volume, i.e., that mass can be neither created nor destroyed in the volume. If the fluid is incompressible ($\rho$ constant), then the equation of continuity reduces to $\nabla \cdot \mathbf{v} = 0$.

15.1.6.2 Newton’s Second Law

The volume on which we are focusing our attention experiences forces from the surrounding fluid at its surface. Both normal forces (perpendicular to the surface and related to pressure) and shear forces (tangent to the surface and related to viscosity) may be experienced. All are expressed in terms of stress, normal stress for the first and shear stress for the second, where a stress in general is defined as a force per unit area. For stress, we use the notation $T_{rs}(x, y, z, t)$, where the first subscript conveys the direction of the normal to the surface involved and the second subscript conveys the direction of the stress, i.e.,

$$T_{rs} = \begin{pmatrix}
\text{the component in the } s \text{ direction of the force on a surface of unit area oriented perpendicular to the } r \text{ direction}
\end{pmatrix}$$

(15.43)

Even more specifically, $T_{xx}$ stands for the force in the (positive) $x$ direction (second subscript) on a surface of unit area oriented perpendicular to the $x$ axis (first subscript) and would be a normal stress. Similarly, the symbol $T_{xy}$ stands for the force in the (positive) $y$ direction (second subscript).

---

4The equation, of course, is quite similar to the parallel equation in electrodynamics expressing the conservation of charge.
on a surface of unit area oriented perpendicular to the \( x \) axis (first subscript) and would be a shear stress. The three stresses \( T_{xx}, T_{xy}, \) and \( T_{xz} \) are shown in Fig. 15.6. As a vector, the total stress on a surface of unit area oriented perpendicular to the \( x \) axis would be given by

\[
\mathbf{T}_x = T_{xx} \hat{i} + T_{xy} \hat{j} + T_{xz} \hat{k}
\]  

(15.44)

Note particularly that the stresses shown are the stresses on the indicated side of the illustrated element arising from its contact with the adjacent portion of the fluid in the direction in which the \( x \) coordinate increases. The stresses that the illustrated element exerts on its neighbor in the positive \( x \) direction will, via Newton’s third law, be equal in magnitude but opposite in direction to those shown. In particular, then, the net force arising from internal stresses on the front and back surfaces of the illustrated element would be given by

\[
\mathbf{F}_x = \left( T_{xx}(x + \Delta x, y, z) - T_{xx}(x, y, z) \right) \hat{i} + \left( T_{xy}(x + \Delta x, y, z) - T_{xy}(x, y, z) \right) \hat{j} + \left( T_{xz}(x + \Delta x, y, z) - T_{xz}(x, y, z) \right) \hat{k} \Delta y \Delta z
\]

(15.45)

Here, we can evaluate the tensions at \( y,z \) because variation of \( y \) and \( z \) over the front and back surfaces will contribute only to second order in \( \Delta y \) and \( \Delta z \)—and we are ignoring terms beyond first order. Further, the derivatives can all be evaluated at argument \((x, y, z)\), since the variation between the front and back surfaces will contribute only to second order in \( \Delta x \). For clarity in the final result, those arguments have been omitted.\(^5\)

Only the subscripts and the variable with respect to which stresses are differentiated must be adjusted to yield expressions for the forces arising from internal stresses on the other two pairs of surfaces on the element in question. We find that

\[
\mathbf{F}_y = \left( \frac{\partial T_{yx}}{\partial y} \hat{i} + \frac{\partial T_{yy}}{\partial y} \hat{j} + \frac{\partial T_{yz}}{\partial y} \hat{k} \right) \Delta x \Delta y \Delta z
\]

(15.46)

and

\[
\mathbf{F}_z = \left( \frac{\partial T_{zx}}{\partial z} \hat{i} + \frac{\partial T_{zy}}{\partial z} \hat{j} + \frac{\partial T_{zz}}{\partial z} \hat{k} \right) \Delta x \Delta y \Delta z
\]

(15.47)

Finally, bringing together all of the \( x \) components of the forces arising from internal stress on all surfaces of the element on which our attention is focused and adding a possible external force\(^6\) \( \rho f_x \Delta x \Delta y \Delta z \), we conclude that the \( x \) component of the net force on the element in question would be given by

\[
F_x = \left( \rho f_x + \frac{\partial T_{xx}}{\partial x} + \frac{\partial T_{yx}}{\partial y} + \frac{\partial T_{zx}}{\partial z} \right) \Delta x \Delta y \Delta z
\]

(15.48)

According to Newton’s second law, this force must, of course, also be given as the product of the mass of the element times the \( x \) component of its acceleration, i.e., by \( (\rho \Delta x \Delta y \Delta z) a_x \). Thus, we have for the \( x \) component of the equation of motion the result

\[
\rho a_x = \rho f_x + \frac{\partial T_{xx}}{\partial x} + \frac{\partial T_{yx}}{\partial y} + \frac{\partial T_{zx}}{\partial z}
\]

(15.49)

Similar equations could be deduced for the \( y \) and \( z \) components of the equation of motion for the fluid.

\(^5\)It is extremely easy to become very confused about these signs. Think about them several times.

\(^6\)For convenience, we define \( f_x \) to be a force per unit mass.
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

$$
\Delta v = v(x + v_x \Delta t, y + v_y \Delta t, z + v_z \Delta t, t + \Delta t) - v(x, y, z, t)
$$

which implies that the acceleration we really want is given by

$$
a = \frac{\Delta v}{\Delta t} = v_x \frac{\partial v}{\partial x} + v_y \frac{\partial v}{\partial y} + v_z \frac{\partial v}{\partial z} + \frac{\partial v}{\partial t} = \frac{\partial v}{\partial t} + (v \cdot \nabla)v
$$

(15.51)

A derivative that is sometimes said to “follow the fluid” and is sometimes called the substantial derivative. Thus, the translation of the $x$ component of Newton’s second law into the vocabulary used to describe the state of a fluid is

$$
\rho \left( \frac{\partial}{\partial t} + v \cdot \nabla \right) v_x = \rho f_x + \frac{\partial T_{xx}}{\partial x} + \frac{\partial T_{yx}}{\partial y} + \frac{\partial T_{zx}}{\partial z}
$$

(15.52)

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 ($\rho$, $v$, and the stresses). We need some further relationships, particularly relationships linking stresses to fluid velocities (and involving new fluid parameters like viscosity). For example, in some simple cases, we can sometimes suppose that the shear stress in one direction (say $x$) on a surface perpendicular to another direction (say $y$) is related to the rate at which the component of the velocity in the first direction changes with displacement in the second direction, i.e., $T_{yx} = \mu \frac{\partial v_x}{\partial y}$, where $\mu$ is the viscosity of the fluid (which in effect is defined by this relationship).\(^8\)

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 $T$ 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 $T_{xy} = T_{yz} = T_{zx} = \ldots = 0$. In that case, Eq. (15.52) reduces to

$$
\rho \left( \frac{\partial}{\partial t} + v \cdot \nabla \right) v_x = \rho f_x - \frac{\partial p}{\partial x}
$$

(15.53)

Combined in vector notation with the other two components, the basic equation for non-viscous flow then is that

$$
\rho \left( \frac{\partial}{\partial t} + v \cdot \nabla \right) v = \rho f - \nabla p
$$

(15.54)

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.

\(^7\)To 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.

\(^8\)A fluid described by this relationship is said to be a Newtonian fluid. The relationship is an approximation, but—fortunately—there are several fluids to which it—and its generalizations to 2D and 3D flow—seem to be accurately applicable.
15.1.6.4 A Second Special Case: Sound Waves

Suppose the flow of interest is 1D, so \( v_y \) and \( v_z \) are both zero and \( v_x \) depends only on \( x \) and \( t \). Further, suppose that external forces are absent and that the flow is inviscid, i.e., that the viscosity is zero and hence that shear stresses are zero \( (T_{yx} = 0 \) and \( T_{zx} = 0) \). Then Eq. (15.42) expressing conservation of mass and Eq. (15.52) expressing Newton’s second law become

\[
\frac{\partial}{\partial x} (\rho v_x) = -\frac{\partial \rho}{\partial x}; \quad \rho \left( \frac{\partial v_x}{\partial t} + v_x \frac{\partial v_x}{\partial x} \right) = -\frac{\partial p}{\partial x} \quad (15.55)
\]

(We have recognized that \( T_{xx} = -p \).) Further, suppose that the velocity of the fluid is small, so we can ignore the non-linear term and reduce these equations to

\[
\frac{\partial}{\partial x} (\rho v_x) = -\frac{\partial \rho}{\partial x}; \quad \rho \frac{\partial v_x}{\partial t} = -\frac{\partial p}{\partial x} \quad (15.56)
\]

Next, let us suppose that \( p = p(\rho) \), i.e., that the pressure in the fluid and its density are functionally related. Then,

\[
\frac{\partial p}{\partial x} = \frac{dp}{d\rho} \frac{\partial \rho}{\partial x} = K \frac{\partial \rho}{\partial x} \quad (15.57)
\]

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

\[
\frac{\partial}{\partial x} (\rho v_x) = -\frac{\partial \rho}{\partial t}; \quad \rho \frac{\partial v_x}{\partial t} = -K \frac{\partial \rho}{\partial x} \quad (15.58)
\]

Finally, we suppose that the pressure departs only slightly from its nominal equilibrium value \( \rho_0 \) and that the velocity in equilibrium is zero (the fluid is quiescent), i.e., we set

\[
\rho = \rho_0 + \delta \rho (x, t) \quad ; \quad v_x = \delta v_x (x, t) \quad (15.59)
\]

Then, ignoring all but the linear terms in small quantities, we find that the equations now are

\[
\rho_0 \frac{\partial \delta v_x}{\partial x} = -\frac{\partial \delta \rho}{\partial t}; \quad \rho_0 \frac{\partial \delta v_x}{\partial t} = -K_0 \frac{\partial \delta \rho}{\partial x} \quad (15.60)
\]

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

\[
\rho_0 \frac{\partial^2 \delta v_x}{\partial t \partial x} = \frac{\partial^2 \delta \rho}{\partial t^2}; \quad \rho_0 \frac{\partial^2 \delta v_x}{\partial x \partial t} = -K_0 \frac{\partial^2 \delta \rho}{\partial x^2} \quad (15.61)
\]

Since the mixed second-partial derivatives are equal, we then conclude that

\[
\frac{\partial^2 \delta \rho}{\partial t^2} = K_0 \frac{\partial^2 \delta \rho}{\partial x^2} \quad (15.62)
\]

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 \( p = nRT/V = MRT/\langle V_{m_i} \rangle \), 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 \( pV^\gamma = p_0 V_0^\gamma \) and \( V = M/\rho \), we find that

\[
p \left( \frac{M}{\rho} \right)^\gamma = p_0 V_0^\gamma \quad \Rightarrow \quad p = \frac{p_0 V_0^\gamma}{M^\gamma / \rho} \quad \Rightarrow \quad K = \frac{dp}{d\rho} = \gamma \frac{p_0 V_0^\gamma}{M^\gamma / \rho} \quad (15.63)
\]
or, evaluating at \( \rho = \rho_0 \) and noting the relationship \( V_0 = M/\rho_0 \), that

\[
K_0 = \frac{dp}{d\rho_0} = \gamma \frac{p_0 V_0^2 \rho_0^{-1}}{M^2 \rho_0^2} = \gamma \frac{p_0 M \gamma \rho_0^{-1}}{M^2 \rho_0} = \gamma \frac{p_0}{\rho_0} = \gamma \frac{RT_0}{m_n}
\]

(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 \( R = 8.31 \text{ N m/(mol K)} \) 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 \^\circ\text{C} = 294 \text{ K}) that

\[
v = \sqrt{\gamma \frac{RT_0}{m_n}} = \sqrt{1.4 \frac{(8.31)(294)}{0.0288}} \text{ m/s} = 344 \text{ m/s}
\]

(15.65)

A further interesting connection appears if we remember from the equipartition theorem of kinetic theory that the rms speed \( v_{rms} \) of molecules of mass \( m \) in a gas at absolute temperature \( T_0 \) satisfies \( \frac{1}{2}mv_{rms}^2 = \frac{3}{2}kT_0 \) or, on multiplying by Avogadro’s number, \( \frac{1}{2}m_n v_{rms}^2 = \frac{3}{2}RT_0 \). We conclude that

\[
v_{rms} = \sqrt{\frac{3RT_0}{m_n}} \implies v = \sqrt{\frac{\gamma RT_0}{m_n}} = \sqrt{\frac{\gamma}{3} \frac{3RT_0}{m_n}} = \sqrt{\frac{\gamma}{3} v_{rms}}
\]

(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 \( x \)), \( \alpha' = d\alpha/dx \), 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

$$- \frac{\partial}{\partial x} \left( \alpha_x(x, y) \frac{\partial \varphi}{\partial x} \right) - \frac{\partial}{\partial y} \left( \alpha_y(x, y) \frac{\partial \varphi}{\partial y} \right) + \beta(x, y) \varphi = f(x, y)$$  \hspace{1cm} (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

$$\varphi = p \quad \text{(on } \Gamma_1)$$  \hspace{1cm} (15.71)

on the portion $\Gamma_1$ of the boundary and the Neumann conditions

$$\left( \alpha_x \frac{\partial \varphi}{\partial x} \hat{i} + \alpha_y \frac{\partial \varphi}{\partial y} \hat{j} \right) \cdot \hat{n} = q \quad \text{(on } \Gamma_2)$$  \hspace{1cm} (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

$$-\alpha_i \left. \frac{d^2 \varphi(x)}{dx^2} \right|_{x_i} - \alpha'_i \left. \frac{d \varphi(x)}{dx} \right|_{x_i} + \beta_i \varphi_i = f_i$$  \hspace{1cm} (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

$$\left. \frac{d \varphi(x)}{dx} \right|_{x_i} \approx \frac{\varphi_{i+1} - \varphi_i}{\Delta x} \quad \text{or} \quad \left. \frac{d \varphi(x)}{dx} \right|_{x_i} \approx \frac{\varphi_i - \varphi_{i-1}}{\Delta x} \quad \text{or} \quad \left. \frac{d \varphi(x)}{dx} \right|_{x_i} \approx \frac{\varphi_{i+1} - 2 \varphi_i + \varphi_{i-1}}{2 \Delta x}$$  \hspace{1cm} (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. 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

$$\left. \frac{d^2 \varphi(x)}{dx^2} \right|_{x_i} \approx \left. \frac{d \varphi(x)}{dx} \right|_{x_i + \Delta x/2} - \left. \frac{d \varphi(x)}{dx} \right|_{x_i - \Delta x/2} = \frac{\varphi_{i+1} - \varphi_i}{\Delta x} - \frac{\varphi_i - \varphi_{i-1}}{\Delta x} = \frac{\varphi_{i+1} - 2 \varphi_i + \varphi_{i-1}}{\Delta x^2}$$  \hspace{1cm} (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.\(^{13}\) 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

\[
-\alpha_i \left( \frac{\varphi_{i+1} - 2\varphi_i + \varphi_{i-1}}{\Delta x^2} \right) - \alpha_i' \left( \frac{\varphi_{i+1} - \varphi_{i-1}}{2\Delta x} \right) + \beta_i \varphi_i = f_i \tag{15.76}
\]

or, on multiplying by \( \Delta x^2 \) and collecting terms with the same index on \( \varphi \), that

\[
\left( -\alpha_i + \frac{\alpha_i' \Delta x}{2} \right) \varphi_{i-1} + \left( 2\alpha_i + \beta_i \Delta x^2 \right) \varphi_i + \left( -\alpha_i - \frac{\alpha_i' \Delta x}{2} \right) \varphi_{i+1} = f_i \Delta x^2 \tag{15.77}
\]

Since \( i \) can assume any value in the interval \( 0 \leq i \leq N \), we thus have \( N + 1 \) equations, just the right number to determine the \( N + 1 \) unknowns \( \varphi_i \).

The conclusion of the last sentence of the previous paragraph, however, is premature. Unfortunately, when evaluated at \( i = 0 \) or \( i = N \), Eq. (15.77) makes reference to \( \varphi_{-1} \) or \( \varphi_{N+1} \), each of which is outside the domain of the problem! Thus, we really have \( N + 3 \) unknowns. The additional information we need lies in the boundary conditions, though the precise way in which these conditions resolve this problem depends on the type of boundary condition. For the 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

\[
\varphi_0 = p \tag{15.78}
\]

For the mixed boundary condition of Eq. (15.69), we use a central difference approximation to the derivative in the boundary condition, finding that

\[
\alpha_N \left( \frac{\varphi_{N+1} - \varphi_{N-1}}{2\Delta x} \right) + \gamma \varphi_N = q \implies \varphi_{N+1} = \varphi_{N-1} + \frac{2 \Delta x}{\alpha_N} (q - \gamma \varphi_N) \tag{15.79}
\]

where \( \alpha_N = \alpha(x_N) = \alpha(L) \). Then, we write out the last equation \((i = N)\) and substitute from Eq. (15.79) to eliminate \( \varphi_{N+1} \), finding ultimately that

\[
-2 \alpha_N \varphi_{N-1} + \left[ 2 \left( \alpha_N + \gamma \Delta x \right) + \left( \beta_N + \frac{\gamma \alpha_N'}{\alpha_N} \right) \Delta x^2 \right] \varphi_N = \frac{\alpha_N'}{\alpha_N} q \Delta x^2 + 2q \Delta x \tag{15.80}
\]

With these resolutions of the first and last equations, we arrive at the now fully defined and complete set

\[
\varphi_0 = p
\]

\[
\left( -\alpha_i + \frac{\alpha_i' \Delta x}{2} \right) \varphi_{i-1} + \left( 2\alpha_i + \beta_i \Delta x^2 \right) \varphi_i + \left( -\alpha_i - \frac{\alpha_i' \Delta x}{2} \right) \varphi_{i+1} = f_i \Delta x^2, \quad 1 \leq i \leq N - 1
\]

\[
-2 \alpha_N \varphi_{N-1} + \left[ 2 \left( \alpha_N + \gamma \Delta x \right) + \left( \beta_N + \frac{\gamma \alpha_N'}{\alpha_N} \right) \Delta x^2 \right] \varphi_N = \frac{\alpha_N'}{\alpha_N} q \Delta x^2 + 2q \Delta x \tag{15.81}
\]

of \( N + 1 \) equations for the \( N + 1 \) unknowns \( \varphi_0, \varphi_1, \ldots, \varphi_N \). Remember that, for definiteness, we have chosen to use 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

\(^{13}\)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 \( \phi_0, \phi_1, \ldots, \phi_N \). 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

\[
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
? & ? & ? & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & ? & ? & ? & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & ? & ? & ? & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & ? & ? & ? & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & ? & ? & ? & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & ? & ? & ? & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & ? & ? & ? & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & ? & ? & \?
\end{pmatrix}
\begin{pmatrix}
\phi_0 \\
\phi_1 \\
\phi_2 \\
\phi_3 \\
\phi_4 \\
\phi_5 \\
\phi_6 \\
\phi_7 \\
\phi_8 \\
\phi_9 \\
\phi_{10}
\end{pmatrix}
= 
\begin{pmatrix}
p \\
? \\
? \\
? \\
? \\
? \\
? \\
? \\
? \\
? \\
?
\end{pmatrix}
\tag{15.82}
\]

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 \( \phi_i \)'s is known at the outset.

Two cases are important. In the first case, the equations are inhomogeneous and the determinant of the coefficient matrix is not zero; and the equations have a unique solution that will approximate the solution to the original boundary value problem. In the second case, the equations are homogeneous, and we hope that the determinant of that matrix zero; and our problem involves seeking the eigenvalues and eigenvectors of the coefficient matrix; we expect to find several solutions, one corresponding to each eigenvalue of the coefficient matrix.\(^{12}\)

Finite difference approaches to PDEs are, of course, subject to error. In general, these errors fall into three distinct and independent categories. In the first instance, we have replaced a continuous variable by a discrete set of values of that variable, i.e., we have discretized the independent variable. In so doing, we have rendered the solution vulnerable to discretization error, an error that will be reduced as the separation of the discrete values is made smaller. From that perspective alone, we would want to make that separation as small as possible, the smaller the better. Unfortunately, as we make the separation smaller, we also increase the computational labor of generating the solution and hence the time required to complete the solution. What’s worse, increasing the computational labor also increases the likelihood that errors will be generated by computer roundoff, which arises because computers retain floating point numbers only to finite precision. We thus must seek a compromise: We want to use a grid with small intervals so as to reduce discretization error but we can’t make the grid too small because roundoff errors may then become significant. Fortunately, in most cases, we will be able to find a grid that is simultaneously fine enough to render discretization error of little consequence and large enough to keep roundoff errors at bay. The standard assessment of these issues involves solving the problem twice, once with a coarser grid and then again with a finer grid. Until roundoff becomes a problem, we can (almost) always safely assume that the solution on the finer grid is more accurate than that on the coarser grid, and a comparison of the two solutions will provide some indication of the accuracy of the solution.

Sometimes, we will use a direct algorithm to solve the discretized equations that convey the solution, and the third type of error will not be present. When an iterative algorithm is used,

\(^{12}\)Two other cases exist but are of no significance. The equations might be inhomogeneous and the determinant of the coefficient matrix zero, in which case no solution exists, or the equations might be homogeneous and the determinant of the coefficient matrix non-zero, in which case only the trivial solution (all \( \phi \)'s zero) exists.
however, we need be concerned not only about discretization and roundoff errors; we must also assess convergence error. In those cases, we are using a method that may or may not converge to the solution of the approximate equations. We must therefore worry not only about the extent to which the solution to the equations we are solving actually represents the solution to the original problem (discretization and roundoff errors) but also about the extent to which the solution we ultimately deduce is actually a solution to the approximating equations (convergence error). Again, we assess the accuracy of the solution to the approximating equations by comparing successive iterates that we hope are converging on that solution. That hope will be more or less justified depending on the rapidity of the apparent convergence.

We can carry our analysis no further without selecting a particular programming language in which to implement explicit coding. That task is undertaken in the next section(s).

15.3 Using IDL to Solve 1D PDEs via an FDM

15.3.1 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 $\alpha, \alpha', \beta,$ 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 IDL statements are\(^{14}\)

```idl
read, prompt='Enter number of segments (N): ', N & N = fix(N)
read, prompt='Enter alpha: ', alpha & alpha = float(alpha)
read, prompt='Enter beta: ', beta & beta = float(beta)
read, prompt='Enter f: ', f & f = float(f)
read, prompt='Enter L: ', L & L = float(L)
read, prompt='Enter p: ', p & p = float(p)
read, prompt='Enter gamma: ', gamma & gamma = float(gamma)
read, prompt='Enter q: ', q & q = float(q)
```

Next, with the statements

```idl
dx = L/N & dx2 = dx^2
x = dx*findgen(N+1)
```

we calculate the size of each segment $dx$, the square of that size $dx^2$, and the values of $x$ at which (equally spaced) nodes will be placed in the interval $0 \leq x \leq L$. Then, recognizing that the solver trisolv we will ultimately invoke requires vectors containing the inhomogeneity, the diagonal

\(^{13}\)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—or computed—for all of those elements.

\(^{14}\)The fix and float functions are invoked to guarantee that entered quantities will be stored in the proper format (integer or real) for the subsequent calculations, regardless of how they happen to be entered.
elements of the coefficient matrix, the superdiagonal elements of that matrix, and the subdiagonal
elements of that matrix,\textsuperscript{15} we invoke the statements

\begin{verbatim}
inomo = fltarr(N+1) + 1.0
inao[0] = p & inomo[N] = inomo[N] + 2.0*q*dx

diag = fltarr(N+1) + 1.0
diag[N] = diag[N] + 2.0*gamma*dx

supdiag = fltarr(N+1) + 1.0
supdiag[0]=0

subdiag = fltarr(N+1) + 1.0
subdiag[N] = -2.0*alpha

phi = trisol( subdiag, diag, supdiag, inomo )
\end{verbatim}

to create those vectors, which reflect Eqs. (15.81) and (15.82). (Note that supdiag[N]
and subdiag[0] will be ignored by trisol.) Finally, we invoke the single statement

\begin{verbatim}
 phi = trisol( subdiag, diag, supdiag, inomo )
\end{verbatim}

to solve the system of equations defined by the several vectors.

A more fully commented command file containing these statements is named fmd1d.pro, is
listed in Appendix 15.A.1, and can be copied from the directory \$HEAD/id1.\textsuperscript{16}

\subsection{15.3.2 An Example: Simple Harmonic Motion}

We now illustrate the application of fmd1d.pro to a simplification of Eq. (15.67), specifically the equation

\[ \frac{m d^2 \varphi}{dt^2} + k \varphi = 0 \quad ; \quad 0 \leq t \leq T \]

(15.83)

for the simple harmonic motion of a mass \( m \) attached to a spring of stiffness \( k \). To match this
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.\textsuperscript{17} 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 \( t = 0 \text{ s} \) and the velocity at \( t = T \). To effect these conditions, we need to assign the values

\[ p = 0.0 \text{ m} \quad ; \quad \gamma = 0.0 \text{ kg/s} \quad ; \quad q = -4.0 \text{ kg\cdot m/s} \]

\textsuperscript{15}IDL literature describes the equation in the form \( A^T \phi = f \), the notation of the transpose being necessary to reflect
IDL’s conventions regarding which index in a two-dimensional matrix designates the row and which the column. From
IDL’s perspective, the matrix we have constructed, in which the first index designates the row and the second index
the column, is \( A^T \). By using this description, the IDL manual makes it possible for us to interpret the equation with
the conventional assignments of row and column indices.

\textsuperscript{16}See the Local Guide for the meaning of \$HEAD\ at your site. At some sites, the file may also be located in the IDL
directory structure such that IDL can find it when it is identified only by its name without a prepended path.

\textsuperscript{17}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.
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 $t = 10$ 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

IDL> @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 $\varphi$. We will have no further interaction with the program. Presently, the IDL 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 IDL’s main prompt. In particular, we could plot the solution by invoking the single statement

IDL> plot, x, phi, thick=3, title='N=20', ticklen=1.0

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

IDL> x020=x & phi020=phi
IDL> @fdm1d
Enter number of segments (N): 50
. (rest same as for N=20)
IDL> x050=x & phi050=phi
IDL> @fdm1d
Enter number of segments (N): 100
. (rest same as for N=20)
IDL> x100=x & phi100=phi

Then, to plot all solutions in a single display, we execute the statements

IDL> !p.multi = [0, 2, 2]
IDL> plot, x020, phi020, thick=3, title='N=20', ticklen=1.0
IDL> plot, x050, phi050, thick=3, title='N=50', ticklen=1.0
IDL> plot, x100, phi100, thick=3, title='N=100', ticklen=1.0
IDL> oplot, x020, phi020, thick=3
IDL> oplot, x050, phi050, thick=3
IDL> oplot, x100, phi100, thick=3
Figure 15.7: Harmonic motion via finite difference analysis. This graph was produced with IDL.

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

```idl
IDL> diff = fltarr(51)
IDL> for i=0,50 do diff[i] = phi100[2*i] - phi050[i]
IDL> print, max(diff), min(diff)
0.0113518  -0.0147179
IDL> print, max(phi100), min(phi100)
1.59632     -1.59715
```

which suggest that the solution for \( N = 50 \) is accurate to an absolute error of about ±0.015 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 \(-0.00357211 \) to \(0.00276732\), 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 initial velocity and check the final velocity with the statements

```idl
IDL> print, 'Initial velocity =', $
```


IDL> (phi100[1]-phi100[0])/(x100[1]-x100[0]), ' m/s'
Initial velocity = -1.38197 m/s
IDL> print, 'Final velocity =', $
IDL> (phi100[100]-phi100[99])/(x100[100]-x100[99]), ' m/s'
Final velocity = 0.958663 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 $-1.38458$ m/s and $0.979258$ m/s, respectively; had we used $N = 500$, we would have found those velocities to be $-1.384574$ and $0.991685$ m/s. Apparently, with increasing $N$, the calculated final velocity converges on the prescribed boundary value.

### 15.9 Finite Element Methods (FEMs) in One Dimension

Finite element methods provide an alternative to finite difference methods for approximating the solution of a boundary value problem—ordinary or partial differential equation plus boundary conditions—in one or more dimensions. The total process involves

1. **Discretizing the domain (preprocessing).** The region of interest is subdivided into a number of small elements by appropriately chosen nodes, at each of which an approximation to the dependent variable $\phi$ will be sought. The discretization of the domain through the identification of suitable nodes and elements is known as preprocessing.

2. **Selecting the interpolation or shape functions.** The interpolation or shape functions for approximating the dependent variable within an element are selected. Because of its simplicity, linear interpolation is commonly used. Higher-order polynomials are more accurate but they also result in a more complicated formulation.

3. **Formulating the equations for a single element.** Equations for each element—the elemental equations—are formulated using the Ritz variational method, the Galerkin method, or some alternative and less common method.

4. **Assembling the system of equations.** The full system of equations is then obtained by (1) assembling the elemental equations into a set applying to the entire region and (2) imposing continuity conditions at the interfaces between elements.

5. **Incorporating the boundary conditions.** Constraints imposed at the boundaries of the region of interest are brought to bear to resolve unknowns left undetermined by the previous steps.

6. **Solving the system of equations.** The (probably large) system of simultaneous, linear, algebraic equations deduced at the previous step is solved for the dependent variable at each node. Such numerical methods as Gauss-Jordan elimination, Gaussian elimination with backsubstitution, and LU (lower-upper) decomposition can all be invoked.

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.
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.

Table 15.1: The connectivity matrix for the example in Section 15.9.1.

<table>
<thead>
<tr>
<th>Element</th>
<th>Global number of local node 1</th>
<th>Global number of local node 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>M − 1</td>
<td>N − 2</td>
<td>N − 1</td>
</tr>
<tr>
<td>M</td>
<td>N − 1</td>
<td>N</td>
</tr>
</tbody>
</table>

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 \(x_N = L\).\(^{18}\) 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 \(x_1^{(e)}\) and \(x_2^{(e)}\), with \(x_1^{(e)} < x_2^{(e)}\). In the present one-dimensional context, the locally and globally numbered coordinates of the nodes are related by

\[
x_1^{(e)} = x_e, \quad x_2^{(e)} = x_{e+1}; \quad e = 1, 2, 3, \ldots, M
\]  

(15.84)

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 \(x_1^{(2)}\) 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, \(l(e) = x_{e+1}^{(e)} - x_e^{(e)} = x_2^{(e)} - x_1^{(e)}\) and, in general, will vary from element to element.

\(^{18}\)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 \(i = 0\).
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.\(^\text{19} \) Thus, in the \( e \)-th element, we approximate the unknown function by

\[
\tilde{\varphi}^{(e)} = a^{(e)} + b^{(e)} x
\]  

(15.85)

where \( a^{(e)} \) and \( b^{(e)} \) are constants to be determined. For subsequent convenience, however, it is preferable to express this linear relationship in terms of the values \( \tilde{\varphi}_1^{(e)} \) and \( \tilde{\varphi}_2^{(e)} \) at the end points of the element rather than in terms of the slope \( b^{(e)} \) and intercept \( a^{(e)} \). Since these end points occur locally at \( x_1^{(e)} \) and \( x_2^{(e)} \), we have that

\[
\tilde{\varphi}_1^{(e)} = a^{(e)} + b^{(e)} x_1^{(e)}
\]

\[
\tilde{\varphi}_2^{(e)} = a^{(e)} + b^{(e)} x_2^{(e)}
\]

(15.86)

If we solve Eq. (15.86) for \( a^{(e)} \) and \( b^{(e)} \), substitute the results into Eq. (15.85), and group the terms appropriately, we discover that the approximation in Eq. (15.85) can be written—see Exercise 15.26—in the form

\[
\tilde{\varphi}^{(e)}(x) = \sum_{j=1}^{2} N_j^{(e)}(x) \tilde{\varphi}_j^{(e)}
\]

(15.87)

where

\[
N_1^{(e)}(x) = \frac{x_2^{(e)} - x}{l^{(e)}}
\]

(15.88)

\[
N_2^{(e)}(x) = \frac{x - x_1^{(e)}}{l^{(e)}}
\]

(15.89)

and \( l^{(e)} = x_2^{(e)} - x_1^{(e)} \) is the length of the element. Equations (15.88) and (15.89) define the shape functions for the \( e \)-th element. Graphs of these shape functions are shown in Fig. 15.9. Note that

\[
N_1^{(e)}(x_1^{(e)}) = 1 \quad N_1^{(e)}(x_2^{(e)}) = 0
\]

(15.90)

\[
N_2^{(e)}(x_1^{(e)}) = 0 \quad N_2^{(e)}(x_2^{(e)}) = 1
\]

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 between the right-hand side and the left-hand side when the approximate solution \( \tilde{\varphi} \) is substituted into the equation, i.e., by

\[
r(x) = - \frac{d}{dx} \left( \alpha \frac{d\tilde{\varphi}}{dx} \right) + \beta \tilde{\varphi} - f
\]

(15.91)

Were \( \tilde{\varphi}(x) \) an exact solution, the residual \( r(x) \) would be identically zero. Since \( \tilde{\varphi} \) is only an approximation to \( \varphi \), however, \( r(x) \) will be nonzero. We define the approximate solution by requiring

\(^{19}\)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.
that \( r(x) \) be as small as possible in some sense. We might, for example, choose to make \( r(x) = 0 \) at a discrete set of points. Still better, we can choose to make an appropriate number of weighted “averages” of \( r(x) \) equal to zero, i.e., we can choose \( \varphi_1^{(e)} \) and \( \varphi_2^{(e)} \) so that

\[
R_i^{(e)} = \int_{x_1}^{x_2} W_i^{(e)}(x) r(x) \, dx = 0 \quad ; \quad i = 1, 2
\]

(15.92)

where the \( R_i^{(e)} \) are the weighted residual integrals and the \( W_i^{(e)}(x) \) are as yet unspecified weighting functions for the \( e \)-th element. Note that we need as many weighting functions as we have nodes, since there are that many values of \( \tilde{\varphi}_i^{(e)} \) to determine. Different choices for these weighting functions yield different—though in the end essentially equivalent—methods of solution. In Galerkin’s method, the weighting functions are chosen as the shape functions \( N_i^{(e)}(x) \) used for the expansion of \( \tilde{\varphi} \) in Eq. (15.87).\(^2\) With that choice, the weighted residual integral for the \( e \)-th element is given by

\[
R_i^{(e)} = \int_{x_1}^{x_2} N_i^{(e)} \left[ -\frac{d}{dx} \left( \alpha(x) \frac{d\tilde{\varphi}_i^{(e)}}{dx} + \beta(x) \tilde{\varphi}_i^{(e)} \right) + \beta(x) N_i^{(e)} \tilde{\varphi}_i^{(e)} \right] dx - \int_{x_1}^{x_2} N_i^{(e)} f(x) \, dx \quad i = 1, 2
\]

(15.93)

where Eq. (15.91) was substituted into Eq. (15.92) for \( r(x) \). If the first term is integrated by parts, we find that\(^2\)

\[
R_i^{(e)} = \int_{x_1}^{x_2} N_i^{(e)} \left[ \alpha(x) \frac{dN_i^{(e)}}{dx} \frac{d\tilde{\varphi}_i^{(e)}}{dx} + \beta(x) N_i^{(e)} \tilde{\varphi}_i^{(e)} \right] dx
\]

\[
- \int_{x_1}^{x_2} N_i^{(e)} f(x) \, dx - \left( \alpha N_i^{(e)} \frac{d\tilde{\varphi}_i^{(e)}}{dx} \right) \bigg|_{x_1}^{x_2} \quad ; \quad i = 1, 2.
\]

(15.94)

Now if Eq. (15.87) is substituted into Eq. (15.94), we find that

\[
R_i^{(e)} = \sum_{j=1}^{2} \tilde{\varphi}_j^{(e)} \int_{x_1}^{x_2} \left( \alpha(x) \frac{dN_i^{(e)}}{dx} \frac{dN_j^{(e)}}{dx} + \beta(x) N_i^{(e)} N_j^{(e)} \right) dx
\]

\(2\) 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.

\(2\) Recall that \( \int u \, dv = uv - \int v \, du \) where for this case \( u = N_i^{(e)} \) and \( dv = (d/dx) (\alpha^{(e)} \tilde{\varphi}_i^{(e)}/dx) \) dx.
These equations for the weighted residual integrals can also be expressed in matrix form as

\[
\begin{bmatrix}
R_1^{(e)} \\
R_2^{(e)}
\end{bmatrix} = \begin{bmatrix}
K_{11}^{(e)} & K_{12}^{(e)} \\
K_{21}^{(e)} & K_{22}^{(e)}
\end{bmatrix} \begin{bmatrix}
\tilde{\phi}_1^{(e)} \\
\tilde{\phi}_2^{(e)}
\end{bmatrix} - \begin{bmatrix}
b_1^{(e)} \\
b_2^{(e)}
\end{bmatrix} - \begin{bmatrix}
g_1^{(e)} \\
g_2^{(e)}
\end{bmatrix}
\] (15.96)

or more compactly as

\[
\{R^{(e)}\} = [K^{(e)}]\{\tilde{\phi}^{(e)}\} - \{b^{(e)}\} - \{g^{(e)}\}
\] (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

\[
g_i^{(e)} = \left( \alpha N_i^{(e)} \frac{d\tilde{\phi}^{(e)}}{dx} \right) \bigg|_{x_2^{(e)}}^{x_1^{(e)}} = \left( \alpha N_i^{(e)} \frac{d\tilde{\phi}^{(e)}}{dx} \right) \bigg|_{x=x_2^{(e)}} - \left( \alpha N_i^{(e)} \frac{d\tilde{\phi}^{(e)}}{dx} \right) \bigg|_{x=x_1^{(e)}}
\] (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\(^{22}\)

\[
K_{11}^{(e)} = K_{22}^{(e)} = \frac{\alpha^{(e)} l^{(e)}}{l^{(e)}} + \frac{\beta^{(e)} l^{(e)}}{6}
\] (15.101)

\[
K_{12}^{(e)} = K_{21}^{(e)} = -\frac{\alpha^{(e)} l^{(e)}}{l^{(e)}} + \frac{\beta^{(e)} l^{(e)}}{6}
\] (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)

If, finally, we replace \(N_1^{(e)}\) and \(N_2^{(e)}\) in Eq. (15.100) with the shape functions in Eqs. (15.88) and (15.89), we have that

\[
g_1^{(e)} = -\alpha(x_1^{(e)}) \left. \frac{d\tilde{\phi}^{(e)}}{dx} \right|_{x_1^{(e)}}
\]

\[
g_2^{(e)} = \alpha(x_2^{(e)}) \left. \frac{d\tilde{\phi}^{(e)}}{dx} \right|_{x_2^{(e)}}
\] (15.104)

\(^{22}\)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.
With \([K], \{b\}, \text{and} \{g\}\) 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

\[
\begin{bmatrix}
K_{11}^{(e)} & K_{12}^{(e)} \\
K_{21}^{(e)} & K_{22}^{(e)}
\end{bmatrix}
\begin{bmatrix}
\varphi_1^{(e)} \\
\varphi_2^{(e)}
\end{bmatrix} =
\begin{bmatrix}
b_1^{(e)} \\
b_2^{(e)}
\end{bmatrix} +
\begin{bmatrix}
g_1^{(e)} \\
g_2^{(e)}
\end{bmatrix}
\]  \hspace{1cm} (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

\[
\begin{bmatrix}
K_{11}^{(1)} & K_{12}^{(1)} \\
K_{21}^{(1)} & K_{22}^{(1)}
\end{bmatrix}
\begin{bmatrix}
\varphi_1^{(1)} \\
\varphi_2^{(1)}
\end{bmatrix} =
\begin{bmatrix}
b_1^{(1)} \\
b_2^{(1)}
\end{bmatrix} +
\begin{bmatrix}
g_1^{(1)} \\
g_2^{(1)}
\end{bmatrix}
\]  \hspace{1cm} (15.106)

which is equivalent to the two equations

\[
K_{11}^{(1)} \varphi_1^{(1)} + K_{12}^{(1)} \varphi_2^{(1)} = b_1^{(1)} + g_1^{(1)}
\]  \hspace{1cm} (15.107)

\[
K_{21}^{(1)} \varphi_1^{(1)} + K_{22}^{(1)} \varphi_2^{(1)} = b_2^{(1)} + g_2^{(1)}
\]

Note that in Eqs (15.106) and (15.107) we have employed the local numbering system. For the other two elements—also with local node numbers—we have similarly that

\[
K_{11}^{(2)} \varphi_1^{(2)} + K_{12}^{(2)} \varphi_2^{(2)} = b_1^{(2)} + g_1^{(2)}
\]  \hspace{1cm} (15.108)

\[
K_{21}^{(2)} \varphi_1^{(2)} + K_{22}^{(2)} \varphi_2^{(2)} = b_2^{(2)} + g_2^{(2)}
\]

and

\[
K_{11}^{(3)} \varphi_1^{(3)} + K_{12}^{(3)} \varphi_2^{(3)} = b_1^{(3)} + g_1^{(3)}
\]  \hspace{1cm} (15.109)

\[
K_{21}^{(3)} \varphi_1^{(3)} + K_{22}^{(3)} \varphi_2^{(3)} = b_2^{(3)} + g_2^{(3)}
\]

From the relationship between the local and global numbering systems as defined in Eq. (15.84), however, it is clear that \(x_1^{(1)} = x_2^{(2)} = x_3^{(3)} = x_4\). Imposing the condition of continuity at nodes 2 and 3, we conclude then that \(\varphi_1^{(1)} = \varphi_2^{(2)} = \varphi_1^{(3)}\) and \(\varphi_2^{(2)} = \varphi_2^{(3)}\) and \(\varphi_4^{(3)}\) and \(\varphi_4^{(4)}\). Now if we once again invoke the relationship in Eq. (15.84) and recognize the correspondences

\[
\varphi_1^{(1)} \rightarrow \varphi_1, \hspace{0.5cm} \varphi_2^{(1)} \rightarrow \varphi_2, \hspace{0.5cm} \varphi_2^{(2)} = \varphi_2^{(3)} \rightarrow \varphi_3, \hspace{0.5cm} \text{and} \hspace{0.5cm} \varphi_2^{(3)} \rightarrow \varphi_4
\]

linking the (approximate) values at locally numbered nodes to the (approximate) values \(\varphi_1, \varphi_2, \varphi_3,\) and \(\varphi_4\) at the globally numbered nodes, Eqs. (15.107)–(15.109) become

\[
K_{11}^{(1)} \varphi_1 + K_{12}^{(1)} \varphi_2 = b_1^{(1)} + g_1^{(1)}
\]

\[
K_{12}^{(1)} \varphi_1 + K_{22}^{(1)} \varphi_2 = b_2^{(1)} + g_2^{(1)}
\]

\[
K_{11}^{(2)} \varphi_2 + K_{12}^{(2)} \varphi_3 = b_1^{(2)} + g_1^{(2)}
\]

\[
K_{21}^{(2)} \varphi_2 + K_{22}^{(2)} \varphi_3 = b_2^{(2)} + g_2^{(2)}
\]

\[
K_{11}^{(3)} \varphi_3 + K_{12}^{(3)} \varphi_4 = b_1^{(3)} + g_1^{(3)}
\]

\[
K_{21}^{(3)} \varphi_3 + K_{22}^{(3)} \varphi_4 = b_2^{(3)} + g_2^{(3)}
\]  \hspace{1cm} (15.110)
Thus, we have six equations but only four unknowns. Some of these equations must be redundant. If we choose carefully, we might ignore two of them. Alternatively, we can reduce the number of equations to four by replacing the second and third equations with their sum and, similarly, replacing the fourth and fifth equations with their sum to find that Eq. (15.110) becomes

$$
\begin{bmatrix}
K_{11}^{(1)} & K_{12}^{(1)} & 0 & 0 \\
K_{21}^{(1)} & K_{22}^{(1)} + K_{11}^{(2)} & K_{12}^{(2)} & 0 \\
0 & K_{21}^{(2)} & K_{22}^{(2)} + K_{11}^{(3)} & K_{12}^{(3)} \\
0 & 0 & K_{21}^{(3)} & K_{22}^{(3)}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_1 \\
\tilde{\varphi}_2 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= 
\begin{bmatrix}
b_1^{(1)} \\
b_2^{(1)} + b_1^{(2)} \\
b_2^{(2)} + b_1^{(3)} \\
b_2^{(3)} + b_1^{(4)}
\end{bmatrix} + 
\begin{bmatrix}
g_1^{(1)} \\
g_2^{(1)} + g_1^{(2)} \\
g_2^{(2)} + g_1^{(3)} \\
g_2^{(3)} + g_1^{(4)}
\end{bmatrix}
$$

or more compactly

$$
[K] \{ \tilde{\varphi} \} = \{ b \} + \{ g \}.
$$

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

$$
K_{11} = K_{11}^{(1)} = \frac{\alpha^{(1)} l^{(1)}}{l^{(1)}} + \beta^{(1)} \frac{l^{(1)}}{3}
$$

$$
K_{ii} = K_{22}^{(i-1)} + K_{11}^{(i)} = \frac{\alpha^{(i-1)} l^{(i-1)}}{l^{(i-1)}} + \frac{\beta^{(i-1)} l^{(i-1)}}{3} + \frac{\alpha^{(i)} l^{(i)}}{l^{(i)}} + \frac{\beta^{(i)} l^{(i)}}{3} ; \quad i = 2, 3, 4, \ldots, N - 1
$$

$$
K_{NN} = K_{22}^{M} = \frac{\alpha^{M} l^{M}}{l^{M}} + \frac{\beta^{M} l^{(M)}}{3}
$$

$$
K_{i+1,i} = K_{i,i+1} = K_{12}^{(i)} = \frac{\alpha^{(i)} l^{(i)}}{l^{(i)}} + \frac{\beta^{(i)} l^{(i)}}{6} ; \quad i = 1, 2, 3, \ldots, N - 1
$$

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

$$
b_1 = b_1^{(1)} = f^{(1)} \frac{l^{(1)}}{2}
$$

$$
b_i = b_2^{(i-1)} + b_1^{(i)} = f^{(i-1)} \frac{l^{(i-1)}}{2} + f^{(i)} \frac{l^{(i)}}{2} ; \quad i = 2, 3, 4, \ldots, N - 1
$$

$$
b_N = b_1^{(M)} = f^{(M)} \frac{l^{(M)}}{2}
$$

Finally, we need to evaluate the elements of $\{ g \}$. Note that all but the first and last entries in $\{ g \}$ can be written as

$$
g_i = g_2^{(i-1)} + g_1^{(i)}
$$

If Eq. (15.104) is substituted into Eq. (15.115), we have

$$
g_i = \alpha \left. \frac{d \tilde{\varphi}}{dx} \right|_{x = x_i^{2(i-1)}} - \alpha \left. \frac{d \tilde{\varphi}}{dx} \right|_{x = x_i^{(i)}}
$$
However, $x_2^{(i-1)} = x_1^{(i-1)} = x_i$ [see Eq. (15.84)] and $\alpha (d\tilde{\varphi}/dx)$ must be continuous at $x_i$. As a result, the right-hand side of Eq. (15.116) is zero for $i = 2, 3$ and we are left with the vector

$$
\{g\} = \left\{ -\alpha(x_1) \left. \frac{d\tilde{\varphi}}{dx} \right|_{x = x_1}, 0, 0, \left. \alpha(x_4) \frac{d\tilde{\varphi}}{dx} \right|_{x = x_4} \right\}
$$

(15.117)

In general for a problem with $M$ elements ($N$ nodes), we would find that

$$
g_1 = -\alpha(x_1) \left. \frac{d\tilde{\varphi}}{dx} \right|_{x = x_1 = 0}
g_i = 0 \quad i = 2, 3, 4, \ldots, N-1
$$

$$
g_N = \alpha(x_N) \left. \frac{d\tilde{\varphi}}{dx} \right|_{x = x_N = L}
$$

(15.118)

In consequence of the considerations to this point, our system of equations for $\{\tilde{\varphi}\}$—the (approximate) values at the nodes—now assumes the form

$$
\begin{bmatrix}
  K_{11} & K_{12} & K_{13} & K_{14} \\
  K_{21} & K_{22} & K_{23} & K_{24} \\
  K_{31} & K_{32} & K_{33} & K_{34} \\
  K_{41} & K_{42} & K_{43} & K_{44}
\end{bmatrix}
\begin{bmatrix}
  \tilde{\varphi}_1 \\
  \tilde{\varphi}_2 \\
  \tilde{\varphi}_3 \\
  \tilde{\varphi}_4
\end{bmatrix}
= \begin{bmatrix}
  b_1 \\
  b_2 \\
  b_3 \\
  b_4
\end{bmatrix}
+ \begin{bmatrix}
  g_1 \\
  0 \\
  0 \\
  g_4
\end{bmatrix}
$$

(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 $\{\tilde{\varphi}\}$, we need to incorporate the boundary conditions. First consider the boundary condition of the third kind given in Eq. (15.69). If we replace $\varphi$ by its approximation $\tilde{\varphi}$, solve for $\alpha (d\tilde{\varphi}/dx)$ at $x = L$, and substitute this into Eq. (15.118), we have that

$$
g_N = \alpha \left. \frac{d\tilde{\varphi}}{dx} \right|_{x = L} = q - \gamma \tilde{\varphi}_N
$$

(15.120)

or specifically for our case that

$$
g_4 = \alpha \left. \frac{d\tilde{\varphi}}{dx} \right|_{x = L} = q - \gamma \tilde{\varphi}_4
$$

(15.121)

Thus, the last equation in our system becomes

$$
K_{41} \tilde{\varphi}_1 + K_{42} \tilde{\varphi}_2 + K_{43} \tilde{\varphi}_3 + K_{44} \tilde{\varphi}_4 = b_4 + q - \gamma \tilde{\varphi}_4
$$

(15.122)

which we can recast as

$$
K_{41} \tilde{\varphi}_1 + K_{42} \tilde{\varphi}_2 + K_{43} \tilde{\varphi}_3 + (K_{44} + \gamma) \tilde{\varphi}_4 = b_4 + q
$$

(15.123)

---

23We elect to introduce a symbol for every element in $K$ even though many of those elements are known in advance to be zero.
With this rewriting of the fourth equation, our system of equations becomes
\[
\begin{bmatrix}
K_{11} & K_{12} & K_{13} & K_{14} \\
K_{21} & K_{22} & K_{23} & K_{24} \\
K_{31} & K_{32} & K_{33} & K_{34} \\
K_{41} & K_{42} & K_{43} & K_{44} + \gamma
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_1 \\
\tilde{\varphi}_2 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= \begin{bmatrix}
b_1 \\
b_2 \\
b_3 \\
b_4 + q
\end{bmatrix} + \begin{bmatrix}
g_1 \\
0 \\
0 \\
0
\end{bmatrix}
\tag{15.124}
\]

In general for a domain with \(N\) nodes, we have
\[
K_{NN} \to K_{NN} + \gamma = \frac{\alpha^{(M)}}{l^{(M)}} + \frac{\beta^{(M)}}{3} + \gamma
\tag{15.125}
\]
\[
b_N \to b_N + q = f^{(M)} \frac{l^{(M)}}{2} + q
\tag{15.126}
\]
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 \(\tilde{\varphi}_i (i = 1, 2, 3, \ldots, N)\) and \(M + 1\) equations. However, the Dirichlet boundary condition given in Eq. (15.68) specifies the value of one of these unknowns, specifically \(\tilde{\varphi}_1\). 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.\(^{24}\)

The new system of equations for our three-element example is therefore
\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
K_{21} & K_{22} & K_{23} & K_{24} \\
K_{31} & K_{32} & K_{33} & K_{34} \\
K_{41} & K_{42} & K_{43} & K_{44}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_1 \\
\tilde{\varphi}_2 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= \begin{bmatrix}
p \\
b_2 \\
b_3 \\
b_4
\end{bmatrix}
\tag{15.127}
\]
where \(K_{44}\) and \(b_4\) have been modified according to Eqs. (15.125) and (15.126).

The symmetry of the coefficient matrix is easily restored. Consider, for example, the second equation in Eq. (15.127), namely
\[
K_{21} \tilde{\varphi}_1 + K_{22} \tilde{\varphi}_2 + K_{23} \tilde{\varphi}_3 + K_{24} \tilde{\varphi}_4 = b_2
\tag{15.128}
\]
Since \(\tilde{\varphi}_1 = p\), however, this equation can be recast as
\[
K_{22} \tilde{\varphi}_2 + K_{23} \tilde{\varphi}_3 + K_{24} \tilde{\varphi}_4 = b_2 - K_{21} p
\tag{15.129}
\]
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 inhomogenieties are now known, these four equations uniquely determine the vector \(\{\tilde{\varphi}\}\) of unknowns.\(^{25}\)

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.

\(^{24}\)Actually the original first equation \(K_{11} \tilde{\varphi}_1 + K_{12} \tilde{\varphi}_2 + K_{13} \tilde{\varphi}_3 + K_{14} \tilde{\varphi}_4 = b_1 + g_1\) is merely reinterpreted as an equation determining \(g_1\) after the solution \(\{\tilde{\varphi}\}\) has been found.

\(^{25}\)We assume, of course, that the coefficient matrix is not singular.
## 15.10 Using IDL to Solve 1D PDEs via FEMs

### 15.10.1 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 $\alpha, \alpha', \beta$, 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.\(^{26}\) 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 (BLOCK 1) of the program we wish to write will, then, request input of all the constants needed in the remainder of the program. Appropriate IDL statements are\(^{27}\)

\begin{verbatim}
read, prompt='Enter number of segments (M): ', M & M = fix(M)
read, prompt='Enter alpha: ', alpha & alpha = float(alpha)
read, prompt='Enter beta: ', beta & beta = float(beta)
read, prompt='Enter f: ', f & f = float(f)
read, prompt='Enter l: ', l & l = float(l)
read, prompt='Enter p: ', p & p = float(p)
read, prompt='Enter gamma: ', gamma & gamma = float(gamma)
\end{verbatim}

The next step (BLOCK 2) 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\(^{28,29}\)

\begin{verbatim}
K = fltarr(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[0,0] = S ; Set diagonal elements of K
for i = 1, M-1 do K[i,i] = S2
K[M,M] = S
for i = 0, M-1 do K[i+1,i] = T ; Set elements above and below
for i = 0, M-1 do K[i,i+1] = T ; main diagonal of K
\end{verbatim}

To complete the assembly (BLOCK 3), we create the vector $\{b\}$ as given by Eq. (15.114) in a similar fashion, invoking the statements\(^{26}\)

\begin{verbatim}
K = fltarr(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[0,0] = S ; Set diagonal elements of K
for i = 1, M-1 do K[i,i] = S2
K[M,M] = S
for i = 0, M-1 do K[i+1,i] = T ; Set elements above and below
for i = 0, M-1 do K[i,i+1] = T ; main diagonal of K
\end{verbatim}

\(^{26}\)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.

\(^{27}\)The `fix` and `float` functions guarantee that entered quantities will be stored in the proper format (integer or real) for the subsequent calculations, regardless of how they are entered.

\(^{28}\)Even though the coefficient matrix in this case is tridiagonal, we elect here not to use IDL's routine `trisol` but instead to illustrate an approach that will work even if the coefficient matrix is not tridiagonal.

\(^{29}\)Remember that indices in IDL start at zero rather than one, so all IDL indices are one less than the corresponding subscript on $K$. More subtly, remember that $K_{ij}$ is the element in the $i$th row and the $j$th column, which is denoted $K[j-1,i-1]$ in IDL; i.e., the row and column indices are reversed in IDL's notation. We must here pay attention to that issue.
Finally (BLOCK 4), 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

\[ K[M,M] = K[M,M] + \gamma \]
\[ b[M] = b[M] + q \]

Then (BLOCK 5), we impose the Dirichlet condition as described in Section 15.9.5 with the statements

\[ K[0,0] = 1.0 \]
\[ b[0] = p \]
\[ \text{for } j = 1, M \text{ do } K[j,0] = 0.0 \]
\[ \text{for } i = 1, M \text{ do } b[i] = b[i] - K[0,i]*p \]
\[ \text{for } i = 1, M \text{ do } K[0,i] = 0.0 \]

As in Eq. (15.130), \( K_{11} \), or \( K[0,0] \) in IDL terminology, is set equal to one and \( b_1 \), or \( b[0] \), is set equal to \( p \). Then, we assign the value zero to \( K_{1j} \) for \( j = 2, 3, 4, \ldots, M \). Finally, the last two lines multiply all but the first entry in the first column of \( K \) by \( p \), subtract each of those values from the corresponding element of \( b \) and then—note the order of operations—set all but the first element in the first column of \( K \) to zero. As a result, we now have in \( K \) and \( b \) the coefficient matrix and vector of inhomogenieties as in Eq. (15.130), a symmetric system of equations ready to be solved.

To solve this system of equations (BLOCK 6), we use LU decomposition, invoking IDL’s routines \texttt{ludc} and \texttt{lusol} with the statements

\texttt{ludc, K, index ; Do LU decomposition of K}
\texttt{phi = lusol(K, index, b) ; Solve equation}

The first of these routines \texttt{ludc} replaces \( K \) with the LU decomposition of a row-wise permutation of itself.\(^{30}\) This routine also creates the vector \texttt{index}, which contains a record of the row permutations which occurred as a result of partial pivoting. Along with the vector of inhomogenieties \( b \), the revised matrix \( K \) and vector \texttt{index} are then passed as input to \texttt{lusol}, which uses back substitution to compute the solution. The solution itself is returned in the vector \texttt{phi}.

At this point, \texttt{phi} contains the (approximate) solution to the boundary value problem defined in Section 15.1.7. To plot this solution as part of post-processing, however, we must have the values of the nodes to which the elements in \texttt{phi} correspond. Since we have chosen equally spaced nodes, we can use IDL’s function \texttt{findgen} to generate the proper single-precision, floating-point array. Since each element in this array is set to the value of its one-dimensional subscript, we can produce a vector with \( M+1 \) values, each separated from its predecessor by an amount \( l \), with the statement

x = l*findgen(M+1)

This vector, of course, now contains the coordinates of the $M + 1$ nodes we have introduced.

A more fully commented command file containing these statements is named fem1d.pro, is listed in Appendix 15.C.1, and can be copied from the directory $\text{HEAD/idl}$.

15.10.2 An Example: Simple Harmonic Motion

We now illustrate the application of fem1d to a simplification of Eq. (15.67), specifically the equation

$$m \frac{d^2 \varphi}{dt^2} + k \varphi = 0 \quad ; \quad 0 \leq t \leq T$$  \hspace{1cm} (15.131)

for the simple harmonic motion of a mass $m$ attached to a spring of stiffness $k$. To match this situation to that discussed in Section 15.1.7, we must interpret $\alpha$ as a constant equal to $-m$, $\beta$ as a constant equal to $k$, $L$ as the time $T$ at the end of the interval of interest, and the independent variable $x$ as the independent variable $t$. Further, we must set $f = 0$. The dependent variable $\varphi$ gives the displacement of the oscillator from its equilibrium position. Basically, we are interested in the position over a range of times $0 \leq t \leq T$. In a consistent set of units, we will take

$$m = -\alpha = 4.0 \, \text{kg} \quad ; \quad k = \beta = 3.0 \, \text{N/m}$$

To complete the definition of the problem, we need to specify the boundary conditions. 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 $t = 0$ 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 $\varphi(0) = 0 \, \text{m}$ and the rest of which (with $\alpha = -4.0 \, \text{kg}$) reduce Eq. (15.69) to $d\varphi(t)/dt|_{t=T} = 1.0 \, \text{m/s}$. In this example, we will seek a solution over the interval $0 \leq t \leq 10$ 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 $1.0 \, \text{m/s}$.

With these choices, we are now ready to execute the command file and solve the problem. The statement

IDL> @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 $^{33}$

Enter number of segments (M): 20
Enter alpha: \hspace{1cm} -4.0
Enter beta: \hspace{1cm} 3.0
Enter f: \hspace{1cm} 0.0
Enter l: \hspace{1cm} 0.5
Enter p: \hspace{1cm} 0.0
Enter gamma: \hspace{1cm} 0.0
Enter q: \hspace{1cm} -4.0

$^{31}$At some sites, the file may also be located in the IDL directory structure such that IDL can find it when it is identified only by its name without a prepended path.

$^{32}$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.

$^{33}$Remember that we have written the command file fem1d.pro to expect $l = L/M$. 
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 no further interaction with the program and, presently, the IDL 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 IDL’s main prompt. In particular, we could plot the solution by invoking the single statement

\[
\text{IDL} > \text{plot}, x, \phi, \text{thick}=3, \text{title}='M=20', \text{ticklen}=1.0
\]

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

\[
\begin{align*}
\text{IDL} & > x020 = x \& \phi020 = \phi \\
\text{IDL} & > @\text{fem1d} \\
\text{Enter number of segments (M): 50} \\
& \quad \text{(rest same as for M=20, except } l = 10/50 = 0.2) \\
\text{IDL} & > x050 = x \& \phi050 = \phi \\
\text{IDL} & > @\text{fem1d} \\
\text{Enter number of segments (M): 100} \\
& \quad \text{(rest same as for M=20, except } l = 10/100 = 0.1) \\
\text{IDL} & > x100 = x \& \phi100 = \phi
\end{align*}
\]

Then, to plot all solutions in a single display, we execute the statements

\[
\begin{align*}
\text{IDL} & > !\text{p.multi} = [0, 2, 2] \\
\text{IDL} & > \text{plot}, x020, \phi020, \text{thick}=3, \text{title}='M=20', \text{ticklen}=1.0 \\
\text{IDL} & > \text{plot}, x050, \phi050, \text{thick}=3, \text{title}='M=50', \text{ticklen}=1.0 \\
\text{IDL} & > \text{plot}, x100, \phi100, \text{thick}=3, \text{title}='M=100', \text{ticklen}=1.0 \\
\text{IDL} & > \text{plot}, x020, \phi020, \text{thick}=3, \text{title}='M=20, 50, 100', \text{ticklen}=1.0 \\
\text{IDL} & > \text{oplot}, x050, \phi050, \text{thick}=3 \\
\text{IDL} & > \text{oplot}, x100, \phi100, \text{thick}=3
\end{align*}
\]

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

\[
\begin{align*}
\text{IDL} & > \text{diff} = \text{fltarr}(51) \\
\text{IDL} & > \text{for } i=0,50 \text{ do } \text{diff}[i] = \phi100[2*i] - \phi050[i] \\
\text{IDL} & > \text{print}, \text{max}(\text{diff}), \text{min}(\text{diff}) \\
& \quad 0.0190166 \quad -0.0160106 \\
\text{IDL} & > \text{print}, \text{max}(\phi100), \text{min}(\phi100) \\
& \quad 1.60363 \quad -1.60462
\end{align*}
\]

which suggests that the solution for \( M = 50 \) is accurate to an absolute error of about ±0.02 and that the solution falls in the range \(-1.6 \leq \varphi \leq 1.6\). If we generate a solution for \( M = 200 \), its difference from the solution for \( M = 100 \) ranges from \(-0.00449562 \) to \(0.00537550\), so the solution for \( M = 100 \) evidently has an absolute error of about ±0.005.
Figure 15.10: Harmonic motion via finite element analysis. This graph was produced with IDL.

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

```
IDL> print, (phi100[1]-phi100[0]) / $(x100[1]-x100[0])
-1.38757

IDL> print, (phi100[100]-phi100[99]) / $(x100[100]-x100[99])
0.957034
```

(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 $-1.38584$ m/s and $0.978902$ m/s, respectively; had we used $M = 500$, we would have found those velocities to be $-1.38651$ and $0.991608$ 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. **FINITE DIFFERENCE METHODS (FDMS) IN TWO DIMENSIONS**

### 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),

\[
\frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \tag{15.132}
\]

which is to be solved subject both to the initial conditions

\[
u(x, 0) = f(x) ; \quad \frac{\partial u}{\partial t}(x, 0) = g(x) \tag{15.133}
\]

that specify how the motion of the string is initiated and to appropriate boundary conditions that specify how the medium is “fastened” at its two ends, say at \( x = 0 \) and \( x = L \). To illustrate two distinct types of boundary conditions, we suppose here that the displacement of the medium must be zero for all time at \( x = 0 \) and the derivative of that displacement must be zero for all time at \( x = L \), i.e., that

\[
u(0, t) = 0 ; \quad \frac{\partial u}{\partial x}(L, t) = 0 \tag{15.134}
\]

For a string, for which \( u(x, t) \) represents a displacement, these conditions stipulate physically that the string is tied down at \( x = 0 \) and free at \( x = L \); for an air column, for which \( u(x, t) \) also represents a displacement, these conditions stipulate that the pipe containing the column is closed at \( x = 0 \) (where there will therefore be a displacement node and a pressure antinode) and open at \( x = L \) (where there will be a displacement antinode and a pressure node).

In one numerical approach to this problem via an FDM, we introduce a set of equally spaced nodes in the spatial dimension but leave the temporal dimension continuous. Suppose we divide the interval \( 0 \leq x \leq L \) into \( N \) elements by \( N + 1 \) nodes, each separated from its nearest neighbor(s) by the distance \( \Delta x = L/N \). We denote the coordinates of those nodes by \( x_i = i \Delta x \) with \( i = 0, 1, 2, \ldots, N \). In particular, \( x_0 = 0 \) and \( x_N = L \). Then, we introduce the set of functions \( u_i(t) = u(x_i, t) \), evaluate the PDE of Eq. (15.132) at \( x_i \), and express the spatial derivative at \( x_i \) in terms of finite differences as in Eq. (15.75) to find that

\[
d \frac{\partial^2 u_i}{\partial t^2} = c^2 \frac{\Delta x^2}{\Delta t^2} \left( u_{i+1} - 2u_i + u_{i-1} \right) \tag{15.135}
\]

The boundary conditions can now be invoked to resolve problems with this equation when \( i = 0 \) and \( i = N \), at which points the equation refers to \( u \) at points outside the domain of the problem, i.e., to \( u_{-1} \) or \( u_{N+1} \), respectively. At \( x = 0 \), the boundary condition implies that \( u_0 = 0 \), so we can replace Eq. (15.135) at \( i = 0 \) with the equation \( u_0 = 0 \), though—for the sake of a similar treatment of all nodes—it is more appropriate to regard \( u_0 = 0 \) as the solution to the initial value problem

\[
d \frac{\partial^2 u_0}{\partial t^2} = 0 ; \quad u_0(0) = 0 , \quad \frac{\partial u_0}{\partial t}(0) = 0 \tag{15.136}
\]

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 \( d^2 u_N / d t^2 \). 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
\[
\begin{align*}
\frac{d^2 u_0}{dt^2} &= 0 \quad (15.138) \\
\frac{d^2 u_i}{dt^2} &= \frac{c^2}{\Delta x^2} \left( u_{i+1} - 2 u_i + u_{i-1} \right), \quad i = 1, 2, 3, \ldots, N - 1 \quad (15.139) \\
\frac{d^2 u_N}{dt^2} &= \frac{c^2}{\Delta x^2} \left( -2 u_N + 2 u_{N-1} \right) \quad (15.140)
\end{align*}
\]
of ODEs subject to the initial conditions
\[
\begin{align*}
u_i(0) &= f(x_i) \quad ; \quad \frac{du_i}{dt}(0) = g(x_i) \quad (15.141)
\end{align*}
\]
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
\[
\frac{u_{i,j+1} - 2 u_{i,j} + u_{i,j-1}}{\Delta t^2} = \frac{c^2}{\Delta x^2} \left( u_{i+1,j} - 2 u_{i,j} + u_{i-1,j} \right) \quad (15.142)
\]
Solving for \( u_{i,j+1} \) and introducing \( \alpha = \frac{c^2 \Delta t^2}{\Delta x^2} \), we find the simpler form
\[
u_{i,j+1} = \alpha u_{i+1,j} + 2(1 - \alpha) u_{i,j} + \alpha u_{i-1,j} - u_{i,j-1} \quad (15.143)
\]
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
\[
u_{0,j} = 0 \quad \text{and} \quad \frac{\partial u}{\partial x}(L, t_j) \approx \frac{u_{N+1,j} - u_{N-1,j}}{2 \Delta x} = 0 \quad \Rightarrow \quad u_{N+1,j} = u_{N-1,j} \quad (15.144)
\]
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
\[
\begin{align*}
u_{0,j+1} &= 0 \quad (15.145) \\
u_{i,j+1} &= \alpha u_{i+1,j} + 2(1 - \alpha) u_{i,j} + \alpha u_{i-1,j} - u_{i,j-1}, \quad i = 1, 2, \ldots, N - 1 \quad (15.146) \\
u_{N,j+1} &= 2\alpha u_{N-1,j} + 2(1 - \alpha) u_{N,j} - u_{N,j-1} \quad (15.147)
\end{align*}
\]
In their turn, the initial conditions start the process by stipulating that
\[
u_{i,0} = f(x_i) \quad \text{and} \quad u_{i,1} = u_{i,0} + g(x_i) \Delta t \quad (15.148)
\]
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 \( \alpha \leq 1 \).
15.16. FINITE DIFFERENCE METHODS (FDMS) IN TWO DIMENSIONS

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.

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

\[
u(x,0) = f(x)
\]

that specifies the initial temperature distribution in the medium and to appropriate boundary conditions that specify the way the temperature is controlled at the two ends, say at \( x = 0 \) and \( x = L \). To illustrate both possibilities in a single example, we suppose that, at \( x = 0 \), the temperature is maintained at the value \( T_0 \) and that, at \( x = L \), the medium is insulated so that no heat flow takes place either into or out of the rod at that end, i.e., we suppose that

\[
u(0,t) = T_0 \quad \text{and} \quad \frac{\partial u}{\partial x}(L,t) = 0
\]

In one numerical approach to this problem via an FDM, we introduce a set of equally spaced nodes in the spatial dimension but leave the temporal dimension continuous. Suppose we divide the interval \( 0 \leq x \leq L \) into \( N \) elements by \( N+1 \) nodes, each separated from its nearest neighbor(s) by the
distance $\Delta x = L/N$. We denote the coordinates of those nodes by $x_i = i \Delta x$ with $i = 0, 1, 2, \ldots, N$. In particular, $x_0 = 0$ and $x_N = L$. Then, we introduce the set of functions $u_i(t) = u(x_i, t)$, evaluate the PDE of Eq. (15.149) at $x_i$, and express the spatial derivative at $x_i$ in terms of finite differences as in Eq. (15.75) to find that

$$\frac{du_i}{dt} = \frac{\alpha^2 c}{\Delta x^2} \left( u_{i+1} - 2u_i + u_{i-1} \right) \quad (15.152)$$

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

$$\frac{du_0}{dt} = 0 \quad ; \quad u_0(0) = T_0 \quad (15.153)$$

To address Eq. (15.152) at $i = N$, we look to a central difference approximation of the spatial derivative, finding that

$$\frac{\partial u}{\partial x}(L, t) \approx \frac{u_{N+1}(t) - u_{N-1}(t)}{2\Delta x} = 0 \quad \Rightarrow \quad u_{N+1} = u_{N-1} \quad (15.154)$$

Then, we write Eq. (15.152) for $i = N$, substitute $u_{N-1}$ for $u_{N+1}$, and solve the result for $du_N/dt$. In this approach, in which we have discretized only the spatial variable, the problem of solving a PDE in two variables becomes one of solving the coupled set

$$\frac{du_0}{dt} = 0 \quad (15.155)$$

$$\frac{du_i}{dt} = \frac{\alpha^2}{\Delta x^2} \left( u_{i+1} - 2u_i + u_{i-1} \right) , \quad i = 1, 2, 3, \ldots, N-1 \quad (15.156)$$

$$\frac{du_N}{dt} = \frac{2\alpha^2}{\Delta x^2} \left( u_{N-1} - u_N \right) \quad (15.157)$$

of ODEs subject to the initial condition

$$u_i(0) = f(x_i) \quad (15.158)$$

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

$$\frac{u_{i,j+1} - u_{i,j}}{\Delta t} = \frac{\alpha^2}{\Delta x^2} \left( u_{i+1,j} - 2u_{i,j} + u_{i-1,j} \right) \quad (15.159)$$

Solving for $u_{i,j+1}$ and introducing $\gamma = \alpha^2 \Delta t/\Delta x^2$, we find the simpler form

$$u_{i,j+1} = \gamma u_{i-1,j} + (1 - 2\gamma)u_{i,j} + \gamma u_{i+1,j} \quad (15.160)$$

Evidently, if we know values of $u_{i,j}$ for $j = 0$, we can determine values for $j = 1$, then for $j = 2$, $\ldots$. That is, if we know $u_i$ along one horizontal line in Fig. 15.12, we can step the solution forward
Figure 15.12: Geometry for full discretization of the diffusion equation. Equation 15.160 involves the points marked with solid figures and has been solved to express the solution at the (one) point marked with a solid square in terms of the solutions at the (three) points marked with solid circles.

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

$$ u_{0,j} = T_0 \quad \text{and} \quad \frac{\partial u}{\partial x}(L, t_j) \approx \frac{u_{N+1,j} - u_{N-1,j}}{2\Delta x} = 0 \implies u_{N+1,j} = u_{N-1,j} \quad (15.161) $$

Free of values outside the interval $0 \leq x \leq L$, the equations we use to step the solution to $t_{j+1}$ from knowledge of the solution at previous times then are

$$ u_{0,j+1} = T_0 \quad (15.162) $$

$$ u_{i,j+1} = \gamma u_{i-1,j} + (1 - 2\gamma)u_{i,j} + \gamma u_{i+1,j} \quad , \quad i = 1, 2, \ldots, N - 1 \quad (15.163) $$

$$ u_{N,j+1} = 2\gamma u_{N-1,j} + (1 - 2\gamma)u_{N,j} \quad (15.164) $$

In their turn, the initial conditions start the process by stipulating that

$$ u_{i,0} = f(x_i) \quad (15.165) $$

For consistency with the imposed boundary conditions, we must, of course, require that $f(0) = T_0$. We leave it to Exercise 15.19 to demonstrate that the method described in this paragraph will be unstable unless $\gamma \leq 1/2$.

The discussion at the end of Section 15.2 on discretization error and roundoff errors in finite difference methods applies as much in two dimensions as in one. Since solution of the diffusion equation involves a direct method, however, convergence error as discussed in that earlier section is not here an issue.

We can proceed no further in this example without resorting to a specific coding language, so we postpone further discussion to later sections.
15.16.3 The Laplace Equation

As laid out in Section 15.1.3, the Laplace equation for the steady state temperature distribution in a two-dimensional plate involves two independent spatial variables \(x\) and \(y\) which, together, locate a point in the two-dimensional medium. The temperature \(u(x, y)\) at the point \((x, y)\) satisfies Eq. (15.27),

\[
\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0
\]

which is to be solved subject to boundary conditions that specify the way the temperature is controlled at the boundaries of the region.\(^{34,35}\)

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 \((x, y) = (0, 0), (L, 0), (L, L), (0, L)\) in the \(xy\) plane. Then, so as to illustrate both possible types of boundary condition, we will suppose that

\[
\begin{align*}
    u(x, 0) &= 0 \quad ; \quad u(x, L) = 100 \quad ; \quad u(0, y) = 100 \frac{y}{L} \quad ; \quad \frac{\partial u}{\partial x}(L, y) = 0
\end{align*}
\]

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

\[
\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta x^2} = 0
\]

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:

\(^{34}\)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.

\(^{35}\)Note that this approach to the Laplace equation was also discussed in Section 9.3.1, though the example discussed there involved only Dirichlet boundary conditions.
15.16. **FINITE DIFFERENCE METHODS (FDMS) IN TWO DIMENSIONS**

Figure 15.13: Geometry for full discretization of the Laplace equation. Equation 15.169 involves the points marked with solid figures and has been solved to express the solution at the (one) point marked with a solid square in terms of the solutions at the (four) points marked with solid circles.

- **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

\[
\left. \frac{\partial u}{\partial x} \right|_{x=L} = 0 \quad \Rightarrow \quad \frac{u_{N+1,j} - u_{N-1,j}}{2 \Delta x} = 0 \quad \Rightarrow \quad u_{N+1,j} = u_{N-1,j} \quad (15.170)
\]

For these nodes, Eq. (15.169) becomes

\[
u_{N,j} = \frac{1}{4} \left( u_{N+1,j} + u_{N-1,j} + u_{N,j+1} + u_{N,j-1} \right)
= \frac{1}{4} \left( 2u_{N-1,j} + u_{N,j+1} + u_{N,j-1} \right) \quad (15.171)
\]

and, as long as \( j \) stays in the specified range, this equation no longer involves points outside the region defined by the boundaries.

- **the node at the lower left corner** for which \( i = 0 \) and \( j = 0 \). This point lies on two of the bounding edges. Fortunately, in the present case, the boundary conditions on those two edges are consistent, and we simply replace Eq. (15.169) with \( u_{0,0} = 0 \). In fact, however, this point will never enter into any equation we will need to consider, so the value we assign at this one
CHAPTER 15. SOLVING PARTIAL DIFFERENTIAL EQUATIONS

point is of no consequence. Indeed, in some problems, there may be a discontinuity in the
temperature at an isolated point such as this one. Conveniently, the method we have adopted
is not upset by such a discontinuity.

- the node at the upper left corner, for which \( i = 0 \) and \( j = N \). This point is similar to the
  point \((i,j) = (0,0)\), even to the consistency of the values on the two edges to which it belongs.
  We set \( u_{0,N} = 100 \).

- the node at the lower right corner, for which \( i = N \) and \( j = 0 \). Here, the boundary condition
  on the lower edge dictates that we should set \( u_{N,0} = 0 \). Applied to this node, however,
  Eq. (15.171) with \( j = 0 \) suggests that we should require that

\[
    u_{N,0} = \frac{1}{4} \left( 2u_{N-1,0} + u_{N,1} + u_{N,-1} \right) \tag{15.172}
\]

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

\[
    u_{N,-1} = -u_{N,1} \tag{15.173}
\]

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

\[
    \begin{align*}
    u_{i,j} &= \frac{1}{4} \left( u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} \right) ; \quad 1 \leq i, j \leq N - 1 \\
    u_{i,0} &= 0 \quad ; \quad 1 \leq i \leq N - 1 \\
    u_{i,N} &= 100 \quad ; \quad 1 \leq i \leq N - 1 \\
    u_{0,j} &= 100 \frac{y_j}{L} \quad ; \quad 1 \leq j \leq N - 1 \\
    u_{N,j} &= \frac{1}{4} \left( 2u_{N-1,j} + u_{N,j+1} + u_{N,j-1} \right) \quad ; \quad 1 \leq j \leq N - 1 \tag{15.174}
    \\
    u_{0,0} &= 0 \\
    u_{0,N} &= 100 \\
    u_{N,0} &= 0 \\
    u_{N,N} &= 100
    \end{align*}
\]

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.

15.17 Using IDL 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 IDL.
15.17.1 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 IDL, we would begin by requesting input of all necessary parameters and assuring that each is stored with the proper data type. The statements

\begin{verbatim}
read, prompt='Number of segments: '  , N  & N = fix(N)
read, prompt='Time step: '            , dt & dt = float(dt)
read, prompt='Number of time steps ' , T  & T = fix(T)
read, prompt='Speed of propagation: ' , c  & c = float(c)
read, prompt='Length of string: '    , L  & L = float(L)
\end{verbatim}

accomplish those objectives. Then, we would determine the length of each segment, establish values for the coordinates at which solutions will be generated, and evaluate and display the one parameter $\alpha$ with the statements

\begin{verbatim}
dx = L/N
x = dx*findgen(N+1)
alpha = c^2*dt^2/dx^2
print, 'alpha = ', alpha
\end{verbatim}

Prudence dictates the wisdom of adding the statement

\begin{verbatim}
if alpha GT 1.0 then begin $
    print, 'alpha > 1; execution halted' & stop & $
endif
\end{verbatim}

to halt execution if the parameter $\alpha$ has a value that guarantees an unstable—and hence inaccurate—solution.

As we saw in the general discussion in Section 15.16.1, generation of the solution at the next time instant requires knowledge of the solution at the current time instant and of the solution at the immediate past time instant. At each step of the way, we need preserve only the current and immediate past solutions, but we must keep those solutions until the solution at the next time instant has been generated. In essence, we need three vectors of dimension $N + 1$ for storing solutions. We declare, therefore, that at any step in the process, the vector $u_1$ will store the past solution, the vector $u_2$ will store the current solution, and the vector $u_3$ 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{verbatim}
u1 = fltarr( N+1 ) ; For past solution
u2 = fltarr( N+1 ) ; For current solution
u3 = fltarr( N+1 ) ; For next solution
\end{verbatim}

Next, before coding the loop that will generate the solution, step by step, we must initialize the values in $u_1$ to reflect the initial 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 undisplaced over the first three-eighths
and the last three eighths of its length. Thus, we suppose that

\[
f(x) = \begin{cases} 
0 & 0 \leq x \leq \frac{3}{8}L \\
1 + \cos \frac{8\pi}{L} \left( x - \frac{L}{2} \right) & \frac{3}{8}L \leq x \leq \frac{5}{8}L \\
0 & \frac{5}{8}L \leq x \leq L 
\end{cases}
\]  

(15.175)

which results in a smooth transition from zero displacement outside the center one quarter of the string and the sinusoidal displacement in that interval. Further, we suppose that the string is released from rest so that

\[
g(x) = 0 \tag{15.176}
\]

or, equivalently—see Eq. (15.148)—\(u_{i,0} = u_{i,1}\) or \(u_2 = u_1\). The coding that will impose these initial conditions, display the initial solution, and then display the solution after the first time step then is:

\[
b = 8.0*!\pi/L \\
\text{for } i=3*N/8+1,5*N/8 \text{ do } u1[i] = 1.0 + \cos(b*(x[i]-L/2.0))
\]

!y.range = [ -2.0,2.0 ]
plot, x, u1, thick=3.0
u2 = u1
plot, x, u2, thick=3.0

Here, we have recognized that the displacement over time should range from \(-2.0\) to \(+2.0\), and we have set the vertical scale to display that same range for each anticipated graph.

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

\[
\text{for } j=2,T \text{ do begin }$
\text{\quad } u3[0] = 0 \text{ \& }$
\text{\quad } \text{for } i=1,N-1 \text{ do } u3[i] = alpha*u2[i+1] \text{ + }$
\text{\quad } \quad 2.0*(1.0-alpha)*u2[i] + alpha*u2[i-1] - u1[i] \text{ \& }$
\text{\quad } u3[N] = 2*alpha*u2[N-1]+2*(1.0-alpha)*u2[N]- u1[N] \text{ \& }$
\text{\quad } \text{plot, x, u3, thick=3.0 \ &}$
\text{\quad } u1 = u2 & u2 = u3 \text{ \&}$
\text{\quad } \text{endfor}
\]

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 \textit{fdmwave1d.pro}, is listed in Appendix 15.D.1, and can be copied from the directory \$\text{HEAD/idl}$.\footnote{At some sites, the file may also be located in the IDL directory structure such that IDL can find it when it is identified only by its name without a prepended path.}

Two problems emerge when we run the coding developed to this point with a trial set of values, say \(N = 100, dt = 0.1, T = 20, c = 0.5\), and \(L = 10\) (for which \(\alpha = 0.25\)). First, the solution is

\footnote{The calculation of the range of \(i\) to be used is complicated. Because of integer arithmetic, the quantities \(3\times N/8\) and \(5\times N/8\) 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.}

```
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

\[
\text{wait, delay}
\]

after each plot statement and inserting the statement

\[
\text{read, prompt='Delay between plots (secs): ', delay & delay = float(delay)}
\]

at the beginning so the user can specify the desired delay at execution time.

We can, of course, fix the second problem simply by requesting a larger number of time steps or 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, however, 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

\[
\text{read, prompt='Plot frequency: ', f & f = fix(f)}
\]

to request the number of solutions whose display should be suppressed after a particular solution has been displayed and, in addition, we modify the plot statement in the final loops to make the display conditional upon \( j \)—the loop index—being a multiple of \( f \). In effect, we simply condition all plot statements except the first in a way that allows each step to be calculated but only every \( f \)-th step to be displayed. The second plot statement, including the associated wait statement will be executed only if \( f = 1 \), i.e., all steps are to be displayed. All of the remaining plot and wait statements will be recast in the form

\[
\text{if } f*(j/f) \text{ EQ } j \text{ then begin$
\text{plot, x, u3, thick=3.0 & wait, delay$}
\text{endif}
\]

(See the full listing in Appendix 15.D.1 for the details.) With this embellishment, we might run the program and—with a bit of trial and error—submit the controlling values

\[
\text{IDL> @fdmwave1d}
\text{Number of segments: 100}
\text{Time step: 0.1}
\text{Number of time steps 800}
\text{Speed of propagation: 0.5}
\text{Length of string: 10}
\text{Delay between plots (secs): 0.1}
\text{Plot frequency: 5}
\text{alpha = 0.250000}
\]

By examining the output, we conclude that these parameters advance the solution through about what we would 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

---

38Since both $f$ and $j$ are integers, the division $(j/f)$ will be an integer division. Thus, $f*(j/f)$ will be equal to $j$ only when $j$ is a multiple of $f$. 
exactly cancel each other, should superpose at \( t = 40 \) s to exactly invert the initial pulse, \( \ldots \). 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 \texttt{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

```idl
IDL> @fdmwave1d
Number of segments: 100
Time step: 0.1
Number of time steps 200
Speed of propagation: 0.5
Length of string: 10
Delay between plots (secs): 0.1
Plot frequency): 20
alpha = 0.250000
IDL> u2_100 = u2 & x_100 = x
IDL> print, max(u2_100), min(u2_100)
  0.123434  -0.123497
```

Here, at the next to the last IDL prompt, we save the final values for later plotting and, in the last IDL prompt, we display the range of values assumed by the solution at \( t = 20 \) s. Then, with the input,

```idl
IDL> @fdmwave1d
Number of segments: 200
Time step: 0.05
Number of time steps 400
Speed of propagation: 0.5
Length of string: 10
Delay between plots (secs): 0.1
Plot frequency): 40
alpha = 0.250000
IDL> u2_200 = u2 & x_200 = x
IDL> print, max(u2_200), min(u2_200)
  0.0437220  -0.0436869
```

and again with the input

```idl
IDL> @fdmwave1d
Number of segments: 400
Time step: 0.025
Number of time steps 800
Speed of propagation: 0.5
Length of string: 10
Delay between plots (secs): 0.1
Plot frequency): 80
alpha = 0.250000
IDL> u2_400 = u2 & x_400 = x
IDL> print, max(u2_400), min(u2_400)
  0.0190846  -0.0190193
```

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, producing the
Figure 15.15: Shape of the string at $t = 20$ s for different segment sizes. This graph was produced with IDL.

same number of 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

IDL> !p.multi=[0,1,3]
IDL> !p.charsize=2.0
IDL> plot, x_100, u2_100, thick = 5, title = "N=100", yrange=[-0.2,0.2]
IDL> plot, x_200, u2_200, thick = 5, title = "N=200", yrange=[-0.2,0.2]
IDL> plot, x_400, u2_400, thick = 5, title = "N=400", yrange=[-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 $N = 400$.

15.17.2 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 IDL, we would begin by requesting input of all necessary parameters and assuring that each is stored with the proper data type. The statements

```idl
read, prompt='Number of segments: ', N & N = fix(N)
read, prompt='Time step: ', dt & dt = float(dt)
read, prompt='Number of time steps ', T & T = fix(T)
```
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

$$dx = L/N$$
$$x = dx*\text{findgen}(N+1)$$
$$\gamma = \alpha^2*dt/dx^2$$
$$\text{print, 'gamma = ', gamma}$$

Prudence dictates the wisdom of adding the statement

$$\text{if gamma GT 0.5 then begin}$
$$\text{print, 'gamma > 0.5; execution halted' & stop &}$
$$\text{endif}$$

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

$$u_1 = \text{fltarr( N+1 )} \quad ; \text{For current solution}$$
$$u_2 = \text{fltarr( N+1 )} \quad ; \text{For next solution}$$

Next, before coding the loop that will generate the solution, step by step, we must initialize the values in $u_1$ to reflect the initial temperature distribution in the rod and then display that distribution. To be specific, let us suppose that, initially, the temperature varies like one hump of a sine wave, but only in the middle quarter of its length, and is zero over the first three-eighths and the last three eighths of its length. Thus, we suppose that

$$f(x) = \begin{cases} 
0 & 0 \leq x \leq \frac{3}{8}L \\
1 + \cos \frac{8\pi}{L} \left(x - \frac{L}{2}\right) & \frac{3}{8}L \leq x \leq \frac{5}{8}L \\
0 & \frac{5}{8}L \leq x \leq L 
\end{cases}$$

(15.177)

which results in a smooth transition from zero temperature outside the center one quarter of the string to the sinusoidal displacement in that interval. The coding that will impose these initial conditions and then display and label the initial solution is\footnote{The calculation of the range of $i$ to be used is complicated. Because of integer arithmetic, the quantities $3*N/8$ and $5*N/8$ 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.}
\[ b = 8.0 \pi / L \]

for \( i = 3 \times N/8 + 1, 5 \times N/8 \) do \( u1[i] = 1.0 + \cos(b \times (x[i]-L/2.0)) \)

\[ !y.range = [0.0, 2.0] \]
\[ plot, x, u1, thick=5.0, title='t = 0 s', chsize=1.5 \]
\[ wait, delay \]

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

\[ \text{for } j=1,T \text{ do begin } \]
\[ \quad u2[0] = 0.0 & \]
\[ \quad \text{for } i=1,N-1 \text{ do } u2[i] = gamma \times u1[i-1] \]
\[ \quad \quad + (1.0-2.0 \times gamma) \times u1[i] + gamma \times u1[i+1] \]
\[ \quad u2[N] = 2 \times gamma \times u1[N-1] + (1-2 \times gamma) \times u1[N] \]
\[ \quad \text{tmp = 't '='+string(j*dt,format='(I5)')+' s' } \]
\[ \quad \text{plot, x, u2, thick=5.0, title=tmp, chsize=1.5 } \]
\[ \quad u1 = u2 & \]
\[ \text{endfor} \]

Here, each pass through the outermost loop advances the solution by one time step. Within that loop, we (1) construct the solution at the next time instant by exploiting Eqs. (15.162)–(15.164), (2) plot the solution, and (3) shift the values to prepare for the next pass through the loop. A more fully commented command file containing these statements—and a few others to be discussed in a moment—is named \texttt{fdmdiffus1d.pro}, is listed in Appendix 15.D.2, and can be copied from the directory \$HEAD/idl.\textsuperscript{40}

Two problems emerge when we run the coding developed to this point with a trial set of values, say \( N = 100, dt = 0.1, T = 25, \alpha = 0.1, \) and \( L = 10 \) (for which \( \gamma = 0.1 \)). First, the solution is generated and displayed so quickly that we cannot follow its evolution. Second, even with 25 time steps, the solution is not advanced very far.

We fix the first of these problems by introducing a time delay after each plot is produced by inserting the statement

\[ \text{wait, delay} \]

after each \texttt{plot} statement and inserting the statement

\[ \text{read, prompt='Delay between plots (secs): ', delay \& delay = float(delay)} \]

at the beginning so the user can specify the desired delay at execution time.

We can, of course, fix the second problem simply by requesting a larger number of time steps or by enlarging the time step (or both—though we must be careful not to increase \( \gamma \) beyond 0.5). As the number of time steps requested increases, however, the program as we now have it will display the solution at every step of the way. With the more accurate solutions (numerous segments, short time interval between generated solutions, many time steps to be computed), we would be wise to introduce a mechanism for suppressing the display of many of the generated solutions. To do so, we add at the beginning the statement\textsuperscript{40}

\begin{footnotesize}
\begin{enumerate}
  \item At some sites, the file may also be located in the IDL directory structure such that IDL can find it when it is identified only by its name without a prepended path.
\end{enumerate}
\end{footnotesize}
15.17. USING IDL TO SOLVE 2D PDES VIA AN FDM

read, prompt='Plot frequency: ', f & f = 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 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 plot
statement everywhere it appears with the conditional statement

\[
\text{if } f*(j/f) \text{ EQ } j \text{ then begin }
\]
\[
\begin{align*}
\text{plot, } x, u2, \text{ thick=3.0 } & \text{ & wait, delay } \\
\text{endif}
\end{align*}
\]

(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

IDL> @fdmdiffus1d
Number of segments:  100
Time step:  0.4
Number of time steps  4400
Value of alpha:  0.1
Length of rod:  10
Delay between plots (secs):  0.1
Plot frequency:  50
gamma =  0.400000

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
\( u = 0 \, ^\circ\text{C} \) 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 \( t = 1200 \, \text{s} \). 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 \( t = 160 \, \text{s} \) for
\( N = 100, 200, \) and \( 400 \). To recreate the frame at \( t = 160 \, \text{s} \) in Fig. 15.16, we would execute the
statements and provide the input

IDL> @fdmdiffus1d
Number of segments:  100
Time step:  0.4
Number of time steps  400
Value of alpha:  0.1
Length of rod:  10
Delay between plots (secs):  0.1
Plot frequency:  25
gamma =  0.400000
IDL> u1_100=u1 & x_100 = x

Here, at the last IDL prompt, we save the final values for later plotting. Then, with the input,
Figure 15.16: Temperatures in a rod at various times as it approaches equilibrium. Note that the vertical scale after the graph for \( t = 0 \) s has been expanded the better to show the gradual evolution of the temperature distribution. This graph was produced with IDL.

\[
\begin{align*}
\text{IDL} & > \text{@fdmdiffusid} \\
\text{Number of segments:} & \quad 200 \\
\text{Time step:} & \quad 0.1 \\
\text{Number of time steps} & \quad 1600 \\
\text{Value of alpha:} & \quad 0.1 \\
\text{Length of rod:} & \quad 10 \\
\text{Delay between plots (secs):} & \quad 0.1 \\
\text{Plot frequency:} & \quad 100 \\
\gamma & = 0.400000 \\
\text{IDL} & > u1_200=u1 \ & \text{x}_200 = x
\end{align*}
\]

and again with the input

\[
\begin{align*}
\text{IDL} & > \text{@fdmdiffusid} \\
\text{Number of segments:} & \quad 400 \\
\text{Time step:} & \quad 0.025 \\
\text{Number of time steps} & \quad 6400 \\
\text{Value of alpha:} & \quad 0.1 \\
\text{Length of rod:} & \quad 10 \\
\text{Delay between plots (secs):} & \quad 0.1 \\
\text{Plot frequency:} & \quad 400 \\
\gamma & = 0.400000
\end{align*}
\]
15.17. USING IDL TO SOLVE 2D PDES VIA AN FDM

Figure 15.17: Temperature distribution at \( t = 160 \) s for different segment sizes. This graph was produced with IDL.

IDL> 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, producing the same number of intermediate graphs, preserving the parameter \( \gamma \) at the value 0.4.\(^ {41} \) For graphical comparison (and choosing a vertical scale to reveal the differences most clearly), we generate Fig. 15.17 with the statements

IDL> !p.multi=[0,1,3]
IDL> plot, x_100, u1_100, thick = 5, title = "N=100, t = 160 s", $ yrange=[0.0,0.6], charsize=1.5
IDL> plot, x_200, u1_200, thick = 5, title = "N=200, t = 160 s", $ yrange=[0.0,0.6], charsize=1.5
IDL> plot, x_400, u1_400, thick = 5, title = "N=400, t = 160 s", $ yrange=[0.0,0.6], charsize=1.5

The three frames in Fig. 15.17 appear to be identical. Indeed, were the solutions overplotted, 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 IDL 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\(^ {42} \)

41Note that, since \( \gamma = \alpha^2 dt/dx^2 \), doubling \( N \) which halves \( dx \) and reduces \( dx^2 \) to one quarter of its original value requires that \( dt \) be reduced by a factor of four as well if \( \gamma \) is to retain its original value.

42Remember that the solution for \( N = 200 \) has twice as many values at the solution for \( N = 100 \).
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 \(\pm 7 \times 10^{-5}\).

### 15.17.3 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 IDL, we would begin by requesting input of the two necessary parameters and assuring that each is stored with the proper data type. The statements

```
read, prompt='Number of segments: ', N & N = fix(N)
read, prompt='Length of side: ', L & L = float(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

```
dx = L/N
x = dx*findgen(N+1) & 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

```
u = fltarr(N+1, N+1)
u[*,0] = 100.0
u[0,*] = 100.0 - 100.0*x/L
print, u
```

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

```
100.000 100.000 100.000 100.000 100.000 100.000
80.0000 0.00000 0.00000 0.00000 0.00000 0.00000
60.0000 0.00000 0.00000 0.00000 0.00000 0.00000
40.0000 0.00000 0.00000 0.00000 0.00000 0.00000
20.0000 0.00000 0.00000 0.00000 0.00000 0.00000
0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
```

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 15.17.1 and 15.17.2. 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 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 15.17.1 and 15.17.2, 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.\(^{43}\) 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.\(^{44}\) For the moment, let us simply carry out a user-specified number of iterations. Thus, we would add to the read statements the coding

```idl
read, prompt='Maximum number of iterations: ', maxits & maxits=fix(maxits)
```

to obtain the desired number of iterations, we would change the print statement above to read

```idl
print, 'Iteration 0' & print, u & print
```

so as to label the output and space it away from what precedes and follows and, finally, we would add the multiple loop

```idl
for itcnt=1, maxits do begin $
    for i = 1, N-1 do $
        for j = 1, N-1 do $
            u[i,j] = 0.25*(u[i+1,j] + u[i-1,j] + u[i,j+1] + u[i,j-1]) & $
        for j = 1, N-1 do u[N,j] = 0.25*(2.0*u[N-1,j] + u[N,j+1] + u[N,j-1]) & $
        print, 'Iteration ', itcnt & print, u & print & $
    endfor
endfor
```

to effect the solution. Here, each pass through the outermost loop will effect one iteration in the algorithm described above. A total of maxits iterations will be executed before the loop terminates. Within that outermost loop, the double loop on $i$ and $j$ steps through the “interior” points in the grid, replacing the value at each with the average of the values at its four nearest neighbors. Then, the single loop on $j$ 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 print statement displays the solution after the current iteration is completed. A more fully commented command file containing these statements—and a few others to be discussed in a moment—is named fdmlap2d.pro, is listed in Appendix 15.D.3, and can be copied from the directory $\texttt{SHEAD/idl}$.

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

\(^{43}\)This iterative approach has already been described—though for slightly different boundary conditions—in Section 9.3.1 and implemented in IDL in Section 9.8.

\(^{44}\)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.
IDL> @fdmlap2d
Number of segments: 5
Length of side: 10
Maximum number of iterations: 5

obtaining the output

<table>
<thead>
<tr>
<th>Iteration</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100.000</td>
<td>100.000</td>
<td>100.000</td>
<td>100.000</td>
<td>100.000</td>
<td>100.000</td>
</tr>
<tr>
<td></td>
<td>80.000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td>60.000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td>40.000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td>20.000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100.000</td>
<td>100.000</td>
<td>100.000</td>
<td>100.000</td>
<td>100.000</td>
</tr>
<tr>
<td></td>
<td>80.000</td>
<td>45.0000</td>
<td>36.2500</td>
<td>34.0625</td>
<td>33.5156</td>
</tr>
<tr>
<td></td>
<td>60.000</td>
<td>26.2500</td>
<td>15.6250</td>
<td>12.4219</td>
<td>11.4844</td>
</tr>
<tr>
<td></td>
<td>40.000</td>
<td>16.5625</td>
<td>8.04688</td>
<td>5.11719</td>
<td>4.15039</td>
</tr>
<tr>
<td></td>
<td>20.000</td>
<td>9.14062</td>
<td>4.29688</td>
<td>2.35352</td>
<td>1.62598</td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100.000</td>
<td>100.000</td>
<td>100.000</td>
<td>100.000</td>
</tr>
<tr>
<td></td>
<td>80.000</td>
<td>60.6250</td>
<td>52.5781</td>
<td>49.6289</td>
</tr>
<tr>
<td></td>
<td>60.000</td>
<td>38.2031</td>
<td>27.8125</td>
<td>23.5107</td>
</tr>
<tr>
<td></td>
<td>40.000</td>
<td>23.8477</td>
<td>15.2686</td>
<td>11.3208</td>
</tr>
<tr>
<td></td>
<td>20.000</td>
<td>12.0361</td>
<td>7.41455</td>
<td>5.09033</td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100.000</td>
<td>100.000</td>
<td>100.000</td>
</tr>
<tr>
<td></td>
<td>80.000</td>
<td>67.6953</td>
<td>61.2842</td>
</tr>
<tr>
<td></td>
<td>60.000</td>
<td>44.8389</td>
<td>36.2256</td>
</tr>
<tr>
<td></td>
<td>40.000</td>
<td>28.0359</td>
<td>20.7492</td>
</tr>
<tr>
<td></td>
<td>20.000</td>
<td>13.8626</td>
<td>9.92554</td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100.000</td>
<td>100.000</td>
</tr>
<tr>
<td></td>
<td>80.000</td>
<td>71.5308</td>
</tr>
<tr>
<td></td>
<td>60.000</td>
<td>48.9481</td>
</tr>
<tr>
<td></td>
<td>40.000</td>
<td>30.8900</td>
</tr>
<tr>
<td></td>
<td>20.000</td>
<td>15.2039</td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100.000</td>
</tr>
<tr>
<td></td>
<td>80.000</td>
</tr>
<tr>
<td></td>
<td>60.000</td>
</tr>
<tr>
<td></td>
<td>40.000</td>
</tr>
<tr>
<td></td>
<td>20.000</td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
</tr>
</tbody>
</table>
The progression in the solution from iterate to iterate is quite clear, though it is also clear from the changes taking place from one iterate to the next that we haven't gone far enough to achieve convergence.

While the coding developed to this point certainly displays the essence of the algorithm for solving the Laplace equation, the above output also makes clear that we must carry the solution through a larger number of iterates to achieve convergence. That change, however, will result in much more output unless we suppress all but every \( f \)-th iteration by adding the coding

\[
\text{read, \quad prompt='Display frequency:', \quad f \quad \& \quad f=\text{fix}(f)}
\]

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 the \text{print} statement everywhere it appears with the conditional statement

\[
\text{if } f \ast (\text{itcnt}/f) \text{ EQ itcnt then begin} \$
\quad \text{print, 'Iteration ', itcnt \& print, u \& print \&}
\text{endif}
\]

(See the full listing in Appendix 15.D.3 for the details.) With this embellishment, we might submit the starting values

IDL> @fdmlap2d
Number of segments: \hspace{1cm} 5
Length of side: \hspace{1cm} 10
Maximum number of iterations: \hspace{1cm} 60
Display frequency: \hspace{1cm} 15

obtaining the output

<table>
<thead>
<tr>
<th>Iteration</th>
<th>0</th>
<th>100.000</th>
<th>100.000</th>
<th>100.000</th>
<th>100.000</th>
<th>100.000</th>
<th>100.000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80.0000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td>60.0000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td>40.0000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td>20.0000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration</th>
<th>15</th>
<th>100.000</th>
<th>100.000</th>
<th>100.000</th>
<th>100.000</th>
<th>100.000</th>
<th>100.000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80.0000</td>
<td>79.5749</td>
<td>79.2885</td>
<td>79.1384</td>
<td>79.1087</td>
<td>79.1753</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60.0000</td>
<td>59.3947</td>
<td>58.9870</td>
<td>58.7732</td>
<td>58.7309</td>
<td>58.8257</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40.0000</td>
<td>39.4674</td>
<td>39.1086</td>
<td>38.9204</td>
<td>38.8832</td>
<td>38.9666</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration</th>
<th>30</th>
<th>100.000</th>
<th>100.000</th>
<th>100.000</th>
<th>100.000</th>
<th>100.000</th>
<th>100.000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80.0000</td>
<td>79.9908</td>
<td>79.9846</td>
<td>79.9814</td>
<td>79.9807</td>
<td>79.9822</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60.0000</td>
<td>59.9869</td>
<td>59.9781</td>
<td>59.9735</td>
<td>59.9726</td>
<td>59.9746</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td></td>
</tr>
</tbody>
</table>
In this output, convergence to a steady solution is more evident. Indeed, the solution at iteration 60 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.\(^{45}\) We would prefer to use a more refined grid and to use graphical rather than tabular display of the solution. To achieve those objectives, we need simply specify a larger number of segments, a larger number of iterations, a larger gap between displayed iterates, and the plotting of the iterates with (say) `surface` as an alternative to the printing of a table with `print`. A more fully commented command file containing these modifications—and a few others to be discussed in a moment—is named `fdmlap2d_plot.pro`, is listed in Appendix 15.D.4, and can be copied from the directory `$HEAD/idl`.

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

```
wait, delay
```

after each `surface` statement and inserting at the beginning the statement

```
read, prompt='Delay between plots (secs): ', delay & delay = float(delay)
```

so the user has control over that delay at execution time.

In a first execution of this coding, however, the graphical displays are oriented so as to tempt the eye into viewing the surface from the wrong side. To repair that problem, we replace each occurrence of `u` in a `surface` statement with `reverse(u,2)` to reflect `u` in a horizontal axis (the second argument of the IDL function `reverse`) before creating the surface, thereby moving the edge maintained at 100 °C from the front of the display to the back. (See the listing in Appendix 15.D.4 for the details of this change.)

Once all of these changes have been implemented (see Appendix 15.D.4), we generate displays with the statements

\(^{45}\)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.
Figure 15.18: Approach of temperature distribution to equilibrium. The panels show the initial distribution (upper left), the distribution at iterate 25 (upper right), the distribution at iterate 50 (lower left), and the distribution at iterate 75 (lower right). This graph was produced with IDL.

IDL> @fdmlap2d_plot
Number of segments: 20
Length of side: 10
Maximum number of iterations: 70
Display frequency: 10
Delay between plots (secs): .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 left edge towards the insulated (right) 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.pro` to incorporate this embellishment is the subject of Exercise 15.23.

Convergence of the iterative process to a particular tolerance, however, only provides evidence that the discretized equations have been solved to a particular accuracy. That characteristic assures us only that convergence error is adequately small. (Probably roundoff error is also under control at that point.) By itself, convergence of the iterative process offers only marginal assurance that discretization error is adequately small. To assess discretization error, we would need to compare not successive iterates on a given grid but converged iterates on at least two grids, one of which is finer than the other. Multigrid algorithms to achieve that objective are beyond the scope of this chapter.

15.23 Finite Element Methods (FEMs) in Two Dimensions

While their implementation is more involved in two dimensions than in one dimension, finite element approaches to problems in two dimensions involve the same several steps as those identified at the 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

46 In this example, we elect to begin all numbering of elements and nodes—both global and local—with zero,
15.23. **FINITE ELEMENT METHODS (FEMS) IN TWO DIMENSIONS**

Figure 15.19: Discretization of a general region with triangular elements.

Figure 15.20: Subdivision of a rectangular domain into eight triangular elements.

Figure 15.21: Rectangular domain of Fig. 15.20 with global (left) and local (right) node numbers.
Table 15.2: The connectivity matrix for the example in Section 15.23.

<table>
<thead>
<tr>
<th>Element</th>
<th>Global number of local node 0</th>
<th>Global number of local node 1</th>
<th>Global number of local node 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>

A global label 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, from top to bottom and from left to right. The local node numbers are assigned counterclockwise starting at the lower left node of each element. For element 0, for example, the global number of local node 0 is 1, the global number of local node 1 is 3, and the global number of local node 2 is 0. Similarly, for element 1, we have global node numbers 1, 4, and 3. The rest of the entries are generated in the same way and are recorded in Table 15.2.

15.23.2 Selecting the Interpolation or Shape Functions

The simplest interpolation or shape functions are linear in both $x$ and $y$. For element $e$, we write the approximating equation in the form

$$\tilde{\varphi}^{(e)}(x, y) = a^{(e)} + b^{(e)} x + c^{(e)} y \quad (15.178)$$

The constants are determined so that this equation gives the correct values at all three nodes of any element, i.e., so that

$$\begin{align*}
\tilde{\varphi}_0^{(e)} &= a^{(e)} + b^{(e)} x_0^{(e)} + c^{(e)} y_0^{(e)} \\
\tilde{\varphi}_1^{(e)} &= a^{(e)} + b^{(e)} x_1^{(e)} + c^{(e)} y_1^{(e)} \\
\tilde{\varphi}_2^{(e)} &= a^{(e)} + b^{(e)} x_2^{(e)} + c^{(e)} y_2^{(e)}
\end{align*} \quad (15.179)$$

where subscripts refer to local node numbers. Solution of this system of equations for $a^{(e)}$, $b^{(e)}$, and $c^{(e)}$ and substitution of the solution into Eq. (15.178) gives the approximating function

$$\tilde{\varphi}^{(e)}(x, y) = \sum_{j=0}^{2} N_j^{(e)}(x, y) \tilde{\varphi}_j^{(e)} \quad (15.180)$$

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.

$^{47}$See Exercise 15.15.
where, with $\Delta^{(e)}$—the determinant of the matrix of coefficients in Eq. (15.179)—defined by

$$
\Delta^{(e)} = \begin{vmatrix}
1 & x_0^{(e)} & y_0^{(e)} \\
1 & x_1^{(e)} & y_1^{(e)} \\
1 & x_2^{(e)} & y_2^{(e)}
\end{vmatrix} \quad (15.181)
$$

the interpolation or shape functions $N_j^{(e)}(x,y)$ are given by

$$
N_0^{(e)}(x,y) = \frac{1}{\Delta^{(e)}} \begin{vmatrix}
1 & x & y \\
1 & x_0^{(e)} & y_0^{(e)} \\
1 & x_1^{(e)} & y_1^{(e)} \\
1 & x_2^{(e)} & y_2^{(e)}
\end{vmatrix} \quad (15.182)
$$

$$
N_1^{(e)}(x,y) = \frac{1}{\Delta^{(e)}} \begin{vmatrix}
1 & x & y \\
1 & x_0^{(e)} & y_0^{(e)} \\
1 & x_2^{(e)} & y_2^{(e)}
\end{vmatrix} \quad (15.183)
$$

$$
N_2^{(e)}(x,y) = \frac{1}{\Delta^{(e)}} \begin{vmatrix}
1 & x & y \\
1 & x_0^{(e)} & y_0^{(e)} \\
1 & x_1^{(e)} & y_1^{(e)}
\end{vmatrix} \quad (15.184)
$$

From these expressions and the properties of determinants, we can quickly recognize that each interpolation function has the value one at the node corresponding to its index and the value zero at the other two nodes. Thus, the sum on the right side of Eq. (15.180) reduces at each node to the value given on the left side. Equally true but less obvious, each interpolation function is in fact zero along the entire line joining the two nodes at which it has the value zero. This property will turn out to be useful when the boundary conditions are taken into account.

### 15.23.3 Formulating the Equations for a Single Element

From the general form of Eq. (15.70), the expression for the residual $r$ can be written easily as

$$
\begin{aligned}
r &= -\frac{\partial}{\partial x} \left( \alpha_x \frac{\partial \hat{\varphi}}{\partial x} \right) - \frac{\partial}{\partial y} \left( \alpha_y \frac{\partial \hat{\varphi}}{\partial y} \right) + \beta \hat{\varphi} - f \\
&= -\frac{\partial}{\partial x} \left( \alpha_x \frac{\partial \hat{\varphi}}{\partial x} \right) + \frac{\partial}{\partial y} \left( \alpha_y \frac{\partial \hat{\varphi}}{\partial y} \right) + \beta \hat{\varphi} - f
\end{aligned} \quad (15.185)
$$

Evaluated for element $e$, the weighted residuals then are given by the integral

$$
R_i^{(e)} = \int \int_{\Omega^{(e)}} N_i^{(e)} r \, dx \, dy \quad ; \quad i = 0, 1, 2 \quad (15.186)
$$

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

$$
\begin{aligned}
R_i^{(e)} &= \int \int_{\Omega^{(e)}} N_i^{(e)} \left[ -\frac{\partial}{\partial x} \left( \alpha_x \frac{\partial \hat{\varphi}}{\partial x} \right) - \frac{\partial}{\partial y} \left( \alpha_y \frac{\partial \hat{\varphi}}{\partial y} \right) + \beta \hat{\varphi} - f \right] \, dx \, dy \\
&= -\int \int_{\Omega^{(e)}} N_i^{(e)} \left[ \frac{\partial}{\partial x} \left( \alpha_x \frac{\partial \hat{\varphi}}{\partial x} \right) + \frac{\partial}{\partial y} \left( \alpha_y \frac{\partial \hat{\varphi}}{\partial y} \right) \right] \, dx \, dy \\
&\quad + \int \int_{\Omega^{(e)}} N_i^{(e)} \beta \hat{\varphi} \, dx \, dy - \int \int_{\Omega^{(e)}} N_i^{(e)} f \, dx \, dy
\end{aligned} \quad (15.187)
$$

---

48See Exercise 15.15.
Then, interpreting \( \psi, V_x, \) and \( V_y \) in Eq. (15.232), the identity developed in Appendix 15.E, as \( N_i^{(c)}, \alpha_x \partial \tilde{\phi} / \partial x, \) and \( \alpha_y \partial \tilde{\phi} / \partial y, \) respectively, we can rewrite the first integral in Eq. (15.188) so as to recast the entirety of Eq. (15.188) in the form

\[
R_i^{(c)} = \int \int_{\Omega^{(c)}} \left( \alpha_x \frac{\partial N_i^{(c)}}{\partial x} \frac{\partial \tilde{\phi}^{(c)}}{\partial x} + \alpha_y \frac{\partial N_i^{(c)}}{\partial y} \frac{\partial \tilde{\phi}^{(c)}}{\partial y} + \beta N_i^{(c)} \tilde{\phi}^{(c)} \right) dx dy \\
- \int \int_{\Omega^{(c)}} N_i^{(c)} f dx dy - \oint_{\Gamma^{(c)}} N_i^{(c)} \left( \alpha_x \frac{\partial \tilde{\phi}^{(c)}}{\partial x} \hat{i} + \alpha_y \frac{\partial \tilde{\phi}^{(c)}}{\partial y} \hat{j} \right) \cdot \hat{n}^{(c)} dl
\]

(15.189)

Here, \( \Gamma^{(c)} \) is the path bounding the region \( \Omega^{(c)}, \) and \( \hat{n}^{(c)} \) is a unit vector lying in the plane of element \( e, \) perpendicular at each point to \( \Gamma^{(c)} \), and directed outward from the perspective of a viewer in \( \Omega^{(c)}. \) Substitution of the approximating function of Eq. (15.180) yields for the \textit{weighted residuals} the expression

\[
R_i^{(c)} = \sum_{j=0}^{2} \int \int_{\Omega^{(c)}} \left( \alpha_x \frac{\partial N_i^{(c)}}{\partial x} \frac{\partial \tilde{\phi}^{(c)}}{\partial x} + \alpha_y \frac{\partial N_i^{(c)}}{\partial y} \frac{\partial \tilde{\phi}^{(c)}}{\partial y} + \beta N_i^{(c)} \tilde{\phi}^{(c)} \right) dx dy \\
- \int \int_{\Omega^{(c)}} N_i^{(c)} f dx dy - \oint_{\Gamma^{(c)}} N_i^{(c)} \left( \alpha_x \frac{\partial \tilde{\phi}^{(c)}}{\partial x} \hat{i} + \alpha_y \frac{\partial \tilde{\phi}^{(c)}}{\partial y} \hat{j} \right) \cdot \hat{n}^{(c)} dl
\]

(15.190)

or, in the matrix form,

\[
[K^{(c)}] \{ \tilde{\phi}^{(c)} \} = \{ b^{(c)} \} - \{ g^{(c)} \}
\]

(15.191)

where \([..] \) denotes a matrix and \( \{..\} \) denotes a vector. The matrix \([K^{(c)}]\) is a 3 \( \times \) 3 matrix and \( \{R^{(c)}\}, \{\tilde{\phi}^{(c)}\}, \{b^{(c)}\}, \) and \( \{g^{(c)}\} \) are three-element vectors. The matrix elements \( K_{ij}^{(c)} \) and the vector elements \( b_i^{(c)} \) and \( g_i^{(c)} \) can be obtained from Eq. (15.190), namely

\[
K_{ij}^{(c)} = \int \int_{\Omega^{(c)}} \left( \alpha_x \frac{\partial N_i^{(c)}}{\partial x} \frac{\partial N_j^{(c)}}{\partial x} + \alpha_y \frac{\partial N_i^{(c)}}{\partial y} \frac{\partial N_j^{(c)}}{\partial y} + \beta N_i^{(c)} N_j^{(c)} \right) dx dy
\]

(15.192)

\[
b_i^{(c)} = \int \int_{\Omega^{(c)}} N_i^{(c)} f dx dy
\]

(15.193)

\[
g_i^{(c)} = \oint_{\Gamma^{(c)}} N_i^{(c)} \left( \alpha_x \frac{\partial \tilde{\phi}^{(c)}}{\partial x} \hat{i} + \alpha_y \frac{\partial \tilde{\phi}^{(c)}}{\partial y} \hat{j} \right) \cdot \hat{n}^{(c)} dl
\]

(15.194)

where, in all cases, \( i \) and \( j \) assume the values 0, 1, and 2.

All elements of the matrix \([K^{(c)}]\) and of the vector \( \{b^{(c)}\} \) are known at the start of the problem. The elements of the vector \( \{g^{(c)}\} \) 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

\[
[K^{(c)}] \{ \tilde{\phi}^{(c)} \} = \{ b^{(c)} \} + \{ g^{(c)} \}
\]

(15.195)

Written out in terms of the specific (local) nodes, this equation becomes

\[
\begin{bmatrix}
R_{00}^{(c)} & R_{01}^{(c)} & R_{02}^{(c)} \\
R_{10}^{(c)} & R_{11}^{(c)} & R_{12}^{(c)} \\
R_{20}^{(c)} & R_{21}^{(c)} & R_{22}^{(c)}
\end{bmatrix}
\begin{bmatrix}
\tilde{\phi}_0^{(c)} \\
\tilde{\phi}_1^{(c)} \\
\tilde{\phi}_2^{(c)}
\end{bmatrix}
= 
\begin{bmatrix}
b_0^{(c)} \\
b_1^{(c)} \\
b_2^{(c)}
\end{bmatrix}
+ 
\begin{bmatrix}
g_0^{(c)} \\
g_1^{(c)} \\
g_2^{(c)}
\end{bmatrix}
\]

(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 \( \{ \tilde{\varphi} \} \), 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

\[
[K] \{ \tilde{\varphi} \} = \{ b \} + \{ g \}
\]

(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 global numbers of the three nodes of element 0 are (in order) 1, 3, and 0. Thus,

\[
\tilde{\varphi}_0^{(0)} = \tilde{\varphi}_1 ; \quad \tilde{\varphi}_1^{(0)} = \tilde{\varphi}_3 ; \quad \tilde{\varphi}_2^{(0)} = \tilde{\varphi}_0
\]

(15.198)

and, written with global identifications on the unknown \( \tilde{\varphi} \)'s, Eq. (15.196) for element 0 becomes

\[
\begin{bmatrix}
K_{00}^{(0)} & K_{01}^{(0)} & K_{02}^{(0)} \\
K_{10}^{(0)} & K_{11}^{(0)} & K_{12}^{(0)} \\
K_{20}^{(0)} & K_{21}^{(0)} & K_{22}^{(0)}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_1 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_0
\end{bmatrix}
= 
\begin{bmatrix}
b_0^{(0)} \\
b_1^{(0)} \\
b_2^{(0)}
\end{bmatrix}
+ 
\begin{bmatrix}
g_0^{(0)} \\
g_1^{(0)} \\
g_2^{(0)}
\end{bmatrix}
\]

(15.199)

Similarly, the three nodes of element 1 are (in order) 1, 4, and 3. Thus,

\[
\tilde{\varphi}_0^{(1)} = \tilde{\varphi}_1 ; \quad \tilde{\varphi}_1^{(1)} = \tilde{\varphi}_4 ; \quad \tilde{\varphi}_2^{(1)} = \tilde{\varphi}_3
\]

(15.200)

and Eq. (15.196) for element 1 becomes

\[
\begin{bmatrix}
K_{00}^{(1)} & K_{01}^{(1)} & K_{02}^{(1)} \\
K_{10}^{(1)} & K_{11}^{(1)} & K_{12}^{(1)} \\
K_{20}^{(1)} & K_{21}^{(1)} & K_{22}^{(1)}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_1 \\
\tilde{\varphi}_4 \\
\tilde{\varphi}_3
\end{bmatrix}
= 
\begin{bmatrix}
b_0^{(1)} \\
b_1^{(1)} \\
b_2^{(1)}
\end{bmatrix}
+ 
\begin{bmatrix}
g_0^{(1)} \\
g_1^{(1)} \\
g_2^{(1)}
\end{bmatrix}
\]

(15.201)

We have, of course, found six equations constraining only four unknowns, namely \( \tilde{\varphi}_0, \tilde{\varphi}_1, \tilde{\varphi}_3, \) and \( \tilde{\varphi}_4 \). Two of the six are redundant.\(^{49}\) 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

\[
\begin{bmatrix}
K_{22}^{(0)} & K_{20}^{(0)} & K_{21}^{(0)} & 0 \\
K_{02}^{(0)} & K_{00}^{(0)} & K_{01}^{(0)} & 0 \\
K_{12}^{(0)} & K_{10}^{(0)} & K_{11}^{(0)} & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_0 \\
\tilde{\varphi}_1 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= 
\begin{bmatrix}
b_0^{(0)} \\
b_1^{(0)} \\
b_3^{(0)} \\
0
\end{bmatrix}
+ 
\begin{bmatrix}
g_0^{(0)} \\
g_1^{(0)} \\
g_3^{(0)} \\
0
\end{bmatrix}
\]

(15.202)

\(^{49}\)We assume that the equations are not contradictory.
Similarly, we augment Eq. (15.201) to obtain the result:

\[
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & K^{(1)}_{00} & K^{(1)}_{02} & K^{(1)}_{01} \\
0 & K^{(1)}_{20} & K^{(1)}_{22} & K^{(1)}_{21} \\
0 & K^{(1)}_{10} & K^{(1)}_{12} & K^{(1)}_{11}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_0 \\
\tilde{\varphi}_1 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
b^{(1)}_0 \\
b^{(1)}_2 \\
b^{(1)}_4
\end{bmatrix}
+ 
\begin{bmatrix}
0 \\
g^{(1)}_0 \\
g^{(1)}_2 \\
g^{(1)}_4
\end{bmatrix}
\] (15.203)

Basically, we reduce the original set of six equations to the required four by adding these equations, finding as the result of assembling the equations for elements 0 and 1:

\[
\begin{bmatrix}
K^{(0)}_{22} & K^{(0)}_{20} & K^{(0)}_{21} & 0 \\
K^{(0)}_{02} & K^{(0)}_{00} + K^{(1)}_{00} & K^{(0)}_{01} + K^{(1)}_{02} & K^{(1)}_{01} \\
K^{(0)}_{12} & K^{(1)}_{10} + K^{(1)}_{20} & K^{(1)}_{11} + K^{(1)}_{22} & K^{(1)}_{21} \\
0 & K^{(1)}_{10} & K^{(1)}_{12} & K^{(1)}_{11}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_0 \\
\tilde{\varphi}_1 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
b^{(1)}_0 + b^{(1)}_0 \\
b^{(1)}_1 + b^{(1)}_2 \\
b^{(1)}_4
\end{bmatrix}
+ 
\begin{bmatrix}
g^{(0)}_0 + g^{(1)}_0 \\
g^{(0)}_1 + g^{(1)}_2 \\
g^{(0)}_4
\end{bmatrix}
\] (15.204)

Furthermore, we can expand these equations to the full 9 × 9 set for all of the nodes, finding that, at this stage in the full assembly,

\[
\begin{bmatrix}
K^{(0)}_{22} & K^{(0)}_{20} & 0 & K^{(0)}_{21} & 0 & 0 & 0 & 0 & 0 \\
K^{(0)}_{02} & K^{(0)}_{00} + K^{(1)}_{00} & 0 & K^{(0)}_{01} + K^{(1)}_{02} & K^{(1)}_{01} & 0 & 0 & 0 & 0 \\
K^{(0)}_{12} & K^{(1)}_{10} + K^{(1)}_{20} & 0 & K^{(1)}_{11} + K^{(1)}_{22} & K^{(1)}_{21} & 0 & 0 & 0 & 0 \\
0 & K^{(1)}_{10} & 0 & K^{(1)}_{12} & K^{(1)}_{11} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_0 \\
\tilde{\varphi}_1 \\
\tilde{\varphi}_3 \\
\tilde{\varphi}_4 \\
\tilde{\varphi}_5 \\
\tilde{\varphi}_6 \\
\tilde{\varphi}_7 \\
\tilde{\varphi}_8
\end{bmatrix}
= 
\begin{bmatrix}
b^{(0)}_2 \\
b^{(0)}_0 + b^{(1)}_0 \\
b^{(0)}_1 + b^{(1)}_2 \\
b^{(0)}_4
\end{bmatrix}
+ 
\begin{bmatrix}
g^{(0)}_2 \\
g^{(0)}_0 + g^{(1)}_0 \\
g^{(0)}_1 + g^{(1)}_2 \\
g^{(0)}_4
\end{bmatrix}
\] (15.205)

The pattern is now clear:

- The element \(K^{(e)}_{ij}\) of the 3 × 3 matrix \([K^{(e)}]\) contributes additively to the element \(K_{lm}\) of the 9 × 9 assembled matrix \([K]\), where \(l\) is the global number of the node with local number \(i\).
in element $e$ and $m$ is the *global* number of the node with *local* number $j$ in element $e$. For example, $K_{00}^{(0)}$ contributes to $K_{11}$ because local node 0 in element 0 coincides with global node 1; $K_{10}^{(1)}$ contributes to $K_{41}$ because local node 1 of element 1 coincides with global node 4 and local node 0 of element 1 coincides with global node 1.

- The element $b_i^{(e)}$ contributes additively to the element $b_l$, where $l$ is the *global* number of the node with *local* number $i$ in element $e$. For example, $b_1^{(0)}$ contributes to $b_3$ because local node 1 in element 0 coincides with global node 3.

- The element $g_i^{(e)}$ contributes additively to the element $g_l$, where $l$ is the *global* number of the node with *local* number $i$ in element $e$. For example, $g_2^{(1)}$ contributes to $g_3$ because local node 2 of element 1 coincides with global node 3.

From this point, continuing the assembly to include the contributions from the remaining elements is straightforward but tedious.

### 15.23.5 Incorporating the Boundary Conditions

Apart from the nodal values of $\tilde{\varphi}$, only the $g$'s in Eq. (15.205)—or better in what Eq. (15.205) becomes when all elements have been incorporated—remain unknown. To demonstrate how the boundary conditions either render *a priori* knowledge of these quantities unnecessary or specify these quantities explicitly, we begin by noting that the fully assembled vector $\{g\}$ is

$$
\{g\} = \begin{bmatrix}
g_2^{(0)} \\
g_0^{(0)} + g_0^{(1)} + g_2^{(2)} \\
g_0^{(2)} + g_0^{(3)} \\
g_1^{(0)} + g_2^{(1)} + g_2^{(4)} \\
g_1^{(1)} + g_1^{(2)} + g_2^{(3)} + g_0^{(4)} + g_0^{(5)} + g_2^{(6)} \\
g_1^{(3)} + g_0^{(6)} + g_0^{(7)} \\
g_1^{(4)} + g_2^{(5)} \\
g_1^{(5)} + g_1^{(6)} + g_2^{(7)} \\
g_1^{(7)}
\end{bmatrix} 
$$

(15.206)

Remember, now, that $g_i^{(e)}$ is given by the line integral

$$
g_i^{(e)} = \oint_{f_i^{(e)}} N_i^{(e)} \left( \alpha_x \frac{\partial \tilde{\varphi}^{(e)}}{\partial x} \hat{i} + \alpha_y \frac{\partial \tilde{\varphi}^{(e)}}{\partial y} \hat{j} \right) \cdot \hat{n}^{(e)} \, dl \tag{15.207}
$$

[see Eq. (15.194)] evaluated counterclockwise around the *entirety* of element $e$. Furthermore, notice that the various $g$'s that are added together in each row of Eq. (15.206) are the line integrals around each element that has a vertex at the *global* node with which the row in $\{g\}$ is associated.

Let us look at a few representative elements of $\{g\}$. Suppose we represent the line integrals by a notation like $(i \rightarrow j + j \rightarrow k + k \rightarrow i)$ 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 *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.\footnote{See Exercise 15.15.} Then, note that each of the other integrals appears twice, traversed once in each direction, and the two together cancel. This particular sum boils down to zero! The same behavior characterizes all nodes not on the boundary of $\Omega$. (There is only one such node—global node 4—in the present example.)

We can examine other elements in the vector $\{g\}$ using the same shorthand notation. For example, the element $g_2$ would be represented by

$$
(2 \rightarrow 4 + 4 \rightarrow 1 + 1 \rightarrow 2) + \\
(2 \rightarrow 5 + 5 \rightarrow 4 + 4 \rightarrow 2)
$$

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, 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

$$
\{g\} = \begin{cases} 
3 \rightarrow 0 + 0 \rightarrow 1 \\
0 \rightarrow 1 + 1 \rightarrow 2 \\
1 \rightarrow 2 + 2 \rightarrow 5 \\
6 \rightarrow 3 + 3 \rightarrow 0 \\
2 \rightarrow 5 + 5 \rightarrow 8 \\
7 \rightarrow 6 + 6 \rightarrow 3 \\
8 \rightarrow 7 + 7 \rightarrow 6 \\
5 \rightarrow 8 + 8 \rightarrow 7 
\end{cases}
$$

(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 $[K]\{\psi\} = \{b + g\}$ with the equation $\psi_m = p_m$, 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 $p_m$.\footnote{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.}

Then, to complete the incorporation of this boundary condition, we must reflect the influence of this known value $\psi_m$ 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:

$$
K_{n0} \psi_0 + K_{n1} \psi_1 + \cdots + K_{nm} \psi_m + \cdots + K_{n7} \psi_7 + K_{n8} \psi_8 = b_n + g_n
$$

(15.209)
Since $\tilde{\varphi}_m = p_m$, this expression can be recast as
\[ K_{n0} \tilde{\varphi}_0 + K_{n1} \tilde{\varphi}_1 + \cdots + 0 \tilde{\varphi}_m + \cdots + K_{n7} \tilde{\varphi}_7 + K_{n8} \tilde{\varphi}_8 = b_n + g_n - K_{nm} p_m \] (15.210)

Such recasting of all equations for $n \neq m$ is accomplished by

- Multiplying by $p_m$ all elements except the element associated with node $m$ in the column associated with node $m$ of the original matrix $[K]$,
- Subtracting each resulting product from the corresponding element in the original vector \{b + g\}, and then
- Substituting zero for all elements in the $m$-th column of $[K]$ except the element in the $m$-th row (which had already been set to 1).

So, for each node on which a Dirichlet condition is declared,

- The associated element in the vector \{b + g\} is replaced by the known value at the node corresponding to that element,
- The other elements in the vector \{b + g\} are adjusted as per Eq. (15.210),
- All elements in the associated row and the associated column of $[K]$ except the element in the intersection of that row and column are set to 0, and
- 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 $\mathbf{3}_i \rightarrow 0$, $0 \rightarrow 1$, etc. Thus, the $g$'s associated with portions of the boundary along which Neumann conditions are specified have values directly determinable from those boundary conditions. If Neumann conditions are specified on all boundaries, all parameters in the equation $[K]\{\tilde{\varphi}\} = \{b + g\}$ except the dependent variable $\tilde{\varphi}$ are known from the beginning, so again the problem is reduced to solving a fully determined set of simultaneous linear equations.

We can proceed no further in this example without resorting to a specific coding language, so we postpone further discussion to later sections.

### 15.24 Using IDL to Solve 2D PDEs via an FEM

#### 15.24.1 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

\[
\varphi(x, 0) = p_2 \quad ; \quad \varphi(x, L) = p_1 \tag{15.211}
\]

on the bottom \((y = 0)\) and top \((y = L)\) boundaries, the linear variation

\[
\varphi(0, y) = \frac{p_1 - p_2}{L} y + p_2 \tag{15.212}
\]

between the bottom and top values along the left edge \((x = 0)\), and the Neumann boundary condition

\[
\frac{\partial \varphi}{\partial x}(L, y) = q \tag{15.213}
\]

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 IDL statements are\(^{52}\)

\begin{verbatim}
read, prompt='Enter length of side (L): ', L & L = float(L)
read, prompt='Enter number of segments (d): ', d & d = fix(d)
read, prompt='Enter alpha_x: ', alpha_x & alpha_x = float(alpha_x)
read, prompt='Enter alpha_y: ', alpha_y & alpha_y = float(alpha_y)
read, prompt='Enter beta: ', beta & beta = float(beta)
read, prompt='Enter f: ', f & f = float(f)
read, prompt='Enter value for top edge: ', p1 & p1 = float(p1)
read, prompt='Enter value for bottom edge: ', p2 & p2 = float(p2)
read, prompt='Enter q: ', q & q = float(q)
\end{verbatim}

The number of elements, \(M\), is twice the square of the number of segments on each edge, the number of nodes, \(N\), is equal to the square of one plus the number of segments on each edge, i.e., to \((M+1)^2\), and the length of each segment of each edge is the length of the edge divided by the number of segments into which the edge is divided, all as reflected in the additional coding

\begin{verbatim}
M = 2*d^2 ; Calculate number of elements
N = (d+1)^2 ; Calculate number of nodes
dx = L/d ; Calculate segment size
\end{verbatim}

As illustrated in Fig. 15.22, the number of segments \(d\) also provides a general way to refer to element numbers and global node numbers for varying numbers of elements.

After calculating the number of elements and nodes involved, we find the coordinates of all of the nodes. We elect to store these coordinates in two vectors, \(x\) and \(y\), where each entry is the \(x\) or \(y\) coordinate of the node whose global node number corresponds to the index of the entry in the vector. These assignments can be made with two nested for loops, the inner loop incrementing row numbers down a column of nodes and the outer loop incrementing the column number. In this way the nodes are traversed in the sequence in which we numbered them. Further, we can keep track of the correct global node number with a simple counter. To accomplish the assignment of values to \(x\) and \(y\), we would use the coding\(^{53}\)

\(\text{\textsuperscript{52}}\)The \textit{fix} and \textit{float} functions guarantee that entered quantities will be stored in the proper format (integer or real) for the subsequent calculations, regardless of how they happen to be entered.

\(\text{\textsuperscript{53}}\)We assume that we are creating a \textit{procedure} pro-file, not a \textit{command} file. Thus we do not need ampersands and dollar signs to present loops as statements in a single line.
15.24. USING IDL TO SOLVE 2D PDES VIA AN FEM

Figure 15.22: General element and global node numbering based on the number of segments $d$ into which each edge is divided. Consistent with IDL conventions, we have here numbered elements and nodes starting at 0.

```
x = fltarr(N) ; Create array to store x values
y = fltarr(N) ; Create array to store y values
ct=0 ; Initialize a counter variable
for i = 0, d do begin ; Start row number loop
  for j = 0, d do begin ; Start column number loop
    x[ct] = i*dx ; Find x coordinate
    y[ct] = L - j*dx ; Find y coordinate
    ct=ct+1 ; Increment counter
  endfor ; End loops
endfor
```

We assign global and local node numbers following the pattern in Fig. 15.23. 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 15.2, 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 zeroth, first, and second local nodes of that element. Then, in developing the coding to create this matrix, we recognize that

- Integer division of an element number by $2\cdot d$ yields the index of the vertical line in Fig. 15.22 on which the zeroth node of the element lies. Here, the index of the vertical line is simply a count of the number of vertical lines to its left, so the first vertical line is at index 0, the
second at index 1, and the last at index $d$.

- On a given vertical line, the global node number of the zeroth 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 zeroth local node by
  
  - performing integer division of the element number by two,
  - adding one (since the sequence of zeroth local nodes begins with global node 1), 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.

Thus, the connectivity matrix $cm$ is created by the statements

```plaintext
cm = intarr(3,M) ; Create 3 by M null array
for e = 0, M-1 do begin ; Loop through all elements
  vl = e/(2*d) ; Find index of vertical line
  cm[0,e] = e/2 + 1 + vl ; Find global number of node 0
  if 2*(e/2) eq e then begin ; If e is even
    cm[1,e] = cm[0,e] + d ; find global number of node 1
    cm[2,e] = cm[0,e] - 1 ; find global number of node 2
  endif else begin ; If e is odd
    cm[1,e] = cm[0,e] + d + 1 ; find global number of node 1
    cm[2,e] = cm[0,e] + d ; find global number of node 2
  endelse
endfor
```

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 $\Delta x$, ...
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

\[
x_1^{(e)} = x_2^{(e)} = x_0^{(e)} + \Delta x \quad , \quad y_1^{(e)} = y_0^{(e)} \quad \text{and} \quad y_2^{(e)} = y_0^{(e)} + \Delta x \quad (e \text{ odd})
\]

In contrast, for even numbered elements, we set

\[
x_1^{(e)} = x_0^{(e)} + \Delta x \quad , \quad x_2^{(e)} = x_0^{(e)} \quad \text{and} \quad y_1^{(e)} = y_2^{(e)} = y_0^{(e)} + \Delta x \quad (e \text{ even})
\]

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 \( x_0^{(e)} \) to \( x_0^{(e)} + \Delta x \). 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 odd (see Fig. 15.22), the limits on \( y \) will run from \( y_0^{(e)} \) to \( y_0^{(e)} + x - x_0^{(e)} \), and we conclude that

\[
\int \int_{\Omega^{(e)}} \ldots \ dx \ dy = \int_{x_0^{(e)}}^{x_0^{(e)} + \Delta x} \left[ \int_{y_0^{(e)}}^{y_0^{(e)} + x - x_0^{(e)}} \ldots \ dy \right] \ dx \quad (e \text{ odd})
\]

When \( e \) is even, on the other hand, the limits on \( y \) will instead run from \( y = y_0^{(e)} + x - x_0^{(e)} \) to \( y_0^{(e)} + \Delta x \), and we conclude that

\[
\int \int_{\Omega^{(e)}} \ldots \ dx \ dy = \int_{x_0^{(e)}}^{x_0^{(e)} + \Delta x} \left[ \int_{y_0^{(e)} + x - x_0^{(e)}}^{y_0^{(e)} + \Delta x} \ldots \ dy \right] \ dx \quad (e \text{ even})
\]

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.\(^{54}\) 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

\[
K_{\text{odd}}^{(e)} = \begin{bmatrix}
\frac{\beta \Delta x^2 + 6\alpha_x}{24} & \frac{\beta \Delta x^2 - 12\alpha_x}{24} & \frac{\beta \Delta x^2}{24} \\
\frac{\beta \Delta x^2 - 12\alpha_x}{24} & \frac{\beta \Delta x^2 + 6\alpha_x + 6\alpha_y}{24} & \frac{\beta \Delta x^2 - 12\alpha_y}{24} \\
\frac{\beta \Delta x^2}{24} & \frac{\beta \Delta x^2 - 12\alpha_y}{24} & \frac{\beta \Delta x^2 + 6\alpha_y}{24}
\end{bmatrix}
\]

and

\[
K_{\text{even}}^{(e)} = \begin{bmatrix}
\frac{\beta \Delta x^2 + 6\alpha_y}{24} & \frac{\beta \Delta x^2}{24} & \frac{\beta \Delta x^2 - 12\alpha_y}{24} \\
\frac{\beta \Delta x^2}{24} & \frac{\beta \Delta x^2 + 6\alpha_x}{24} & \frac{\beta \Delta x^2 - 12\alpha_x}{24} \\
\frac{\beta \Delta x^2 - 12\alpha_y}{24} & \frac{\beta \Delta x^2 - 12\alpha_x}{24} & \frac{\beta \Delta x^2 + 6\alpha_x + 6\alpha_y}{24}
\end{bmatrix}
\]

The coding that constructs these two matrices in our program is

\(^{54}\)Details of this evaluation are described more fully in Appendix 15.F.
Kodd = fltarr(3,3) ; Create two 3 by 3 arrays
Keven = fltarr(3,3)

bx = beta*dx^2 ; Evaluate a common quantity

Kodd[0,0] = (bx + 6*alpha_x)/12 ; Assign the appropriate value to
each K[i,j]. Note that the array
Kodd[1,1] = (bx + 6*alpha_x + 6*alpha_y)/12 ; are symmetric, and that Keven
Kodd[2,2] = (bx + 6*alpha_y)/12 ; includes all of the same values as
Kodd[0,1] = Kodd[1,0] ; Kodd, but in different locations.
Kodd[1,2] = (bx - 12*alpha_y)/24
Kodd[2,1] = Kodd[1,2]
Kodd[0,2] = bx/24
Kodd[2,0] = bx/24

Keven[0,0] = Kodd[2,2]
Keven[1,1] = Kodd[0,0]
Keven[2,2] = Kodd[1,1]
Keven[0,1] = Kodd[0,2]
Keven[1,0] = Keven[0,1]
Keven[1,2] = Kodd[0,1]
Keven[2,1] = Keven[1,2]
Keven[0,2] = Kodd[1,2]
Keven[2,0] = Keven[0,2]

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 \(b^{(e)}\) can be found by using a symbol manipulating program to evaluate the integrals in Eq. (15.193). It turns out that all three elements \(b_0^{(e)}\), \(b_1^{(e)}\), and \(b_2^{(e)}\) 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\(^{56}\)

K = fltarr(N,N) ; Create arrays to store values of
b = fltarr(N) ; K[i,j] and b[i]
for e = 0,M-1 do begin ; Count through element numbers
    for i = 0,2 do begin ; 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 = 0,2 do $ ; correct locations in \([K]\) and \(\{b\}\)
            if 2*(e/2) ne e then $
                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]
            endfor
        endfor
    endfor

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.
\(^{56}\)Alternatively, we could make more liberal use of \texttt{begin-end} constructions to eliminate the need for the several dollar signs in the innermost loop.
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 $i$, on the top by $i(d+1)$, and on the bottom by $(i+1)(d+1) - 1$. Once we know these numbers, we can replace the appropriate equations by the given values of $\tilde{\varphi}$, and then reflect the influence of this relationship in the other equations as described in Section 15.23.5. The requisite coding is

for $i = 0, d$ do begin
  $u = i$ ; Nodes on the left boundary
  $s = i*(d+1)$ ; Nodes on the top boundary
  $t = (i+1)*(d+1)-1$ ; Nodes on the bottom boundary
  $p3 = (p1-p2)/L * y[u] + p2$ ; Find values of phi on left boundary
for $j = 0, N-1$ do begin
  $K[j,s] = 0$ ; 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 $j \neq s$ then $b[j] = b[j] - K[s,j]*p1$ ; Reflect influence
  if $j \neq t$ then $b[j] = b[j] - K[t,j]*p2$ ; phi in the other
  if $j \neq u$ then $b[j] = b[j] - K[u,j]*p3$
  $K(u,j) = 0$
endfor
  $K[s,s] = 1$ ; Set the appropriate entry to 1 in
  $K[t,t] = 1$ ; the rows where phi is known
  $K[u,u] = 1$
endfor

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 $g^{(e)}_i$. These two items and the fact that all elements with boundaries on the right edge have odd numbers reduce Eq. (15.194) to

$$g^{(e)}_i = \int_{y_0^{(e)}}^{y_0^{(e)}+\Delta x} N_i^{(e)} \alpha_x \frac{\partial \tilde{\varphi}}{\partial x} dy \quad (15.220)$$

Since we know that $\alpha_x(\partial \tilde{\varphi}/\partial x) = q$ and that $x = x_0^{(e)} + \Delta x$, we can easily evaluate this expression, provided we assume $\alpha_x$ can be treated as a constant over the area of each element. Again invoking a symbol manipulating program,\textsuperscript{57} we find that

$$g_0^{(e)} = 0 \quad g_1^{(e)} = \frac{q \Delta x}{2} \quad g_2^{(e)} = \frac{q \Delta x}{2} \quad (15.221)$$

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 $g_0^{(e)}$ since no node on the right edge is a zeroth local node. The necessary additional coding, including the consolidation of the inhomogenieties into a single vector, is

```idl
  g = fltarr(N) ; Create vector to store g values
for i = 1,d-1 do begin
```

\textsuperscript{57}See Appendix 15.F once more.
end = d*(d+1)+i ; Nodes on right boundary
g[end] = q * dx
dfor
b = b+g ; Store the vector {b+g} in {b}

At this point, K contains the coefficient matrix and b the inhomogeneities for the system of simultaneous linear equations we wish to solve. We need only perform LU decomposition and back substitution to solve the system and then write the results into an appropriate two-dimensional array. The remaining coding thus is

phi = fltarr(N) ; Create vector to store results
LUDC, K, INDEX ; Solve the system of equations and
phi = LUSOL(K, INDEX, b) ; write the results to phi

A = fltarr(d+1,d+1) ; Create a d+1 by d+1 array
cnt=0 ; Initialize a counter variable
for i = 0,d do for j = d,0,-1 do begin ; Use nested for loops to write all
A[i,j] = phi[cnt] ; entries in A
   cnt=cnt+1 ; Increment counter
endfor ; End loop

Since the nodes are, in this example, uniformly spaced over the region of the problem, we need not be concerned about knowing the nodal coordinates for purposes of plotting the solution. The coordinates of the node corresponding to each value in the two-dimensional array A are proportional to the indices of that value in the array.

All of the preceding IDL code has been incorporated into the procedure fem2d.pro, a listing of which can be found in Appendix 15.G.1. (The file itself can be copied from the directory $HEAD/idl.) As the listing indicates, expressing the code as a procedure pro-file rather than as a command file requires the first and last lines of the file to be

PRO fem2d, A

and

END

When run, the procedure simply asks for necessary input, calculates the solution, and returns that solution in the array A. Only the values stored in the array A will be accessible, once the procedure has returned control to IDL’s command level. Other variables used in the procedure are hidden after that return.

15.24.2 An Example: Isotropic Heat Flow

We will now demonstrate fem2d by applying it to an equation in isotropic heat flow. The steady-state temperature \( u(x, y) \) in a square plate of uniform composition is a solution to the two-dimensional Laplace equation

\[
\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0
\]

(15.222)

to which Eq. (15.70) reduces when we replace \( \tilde{\varphi} \) with \( u \) and set

\[
\alpha_x = 1 \quad ; \quad \alpha_y = 1 \quad ; \quad \beta = 0 \quad ; \quad f = 0
\]

(15.223)
We will take the plate to be ten units square \((L = 10)\), and position it in the first quadrant with a corner at the origin, and we will apply the boundary conditions

\[
\begin{align*}
    u(x, 0) &= 0 ; \\
    u(x, 10) &= 100 ; \\
    u(0, y) &= 10y ; \\
    \frac{\partial u}{\partial x}(10, y) &= 0
\end{align*}
\]

i.e., \(p_2 = 0\), \(p_1 = 100\), \(L = 10\), and \(q = 0\). With these choices, we run the program and solve the problem with the single statement

IDL> fem2d, T09

At the several prompts generated by this statement, we enter the values

Enter length of side (L): 10
Enter number of segments (d): 8
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

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 T09, which is returned to the main program and can easily be displayed by invoking the statements

IDL> x = 10.0*findgen(9)/8.0
IDL> surface, T09, x, x, charsize=2.0, thick=5.0, xtitle='x', ytitle='y'

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

IDL> for i=8,0,-1 do print, string( T09(*,i), format='(f6.2)' )
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

reveals even more clearly this anticipated linearity.

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
Figure 15.24: The steady-state temperature distribution of the isotropic plate. This graph was produced with IDL.

IDL> fem2d, T17
Enter length of side (L): 10
Enter number of segments (d): 16
Enter alpha_x: 1
Enter alpha_y: 1
Enter beta: 0
Enter f: 0
Enter value for top edge: 100
Enter value for bottom edge: 0
Enter q: 0

which will generate a solution on a $17 \times 17$ grid with twice as many divisions as the first solution. Then, we might extract from that more refined solution the solution on the $9 \times 9$ grid with the statements

IDL> T09_refined = fltarr(9,9)
IDL> for i=0,8 do for j=0,8 do T09_refined[i,j] = T17[2*i,2*j]

Then, we could display the refined solution on the original grid with the statement

IDL> for i=8,0,-1 do print, string( T09_refined(*,i), format='(f6.2)' )
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
We discover that, at least to two digits after the decimal place, the two solutions are identical,
and we conclude that discretization and roundoff error are under control to this level of accuracy.
(Because we have solved the algebraic equations by a direct method, we need not here be concerned
about convergence error.)

15.30 Exercises

Note: In these exercises, we refer to the programs developed in the text
without appending the file type, adopting this approach to leave to you the
choice of which language to use in addressing the exercise.

15.1. Recast fdm1d to solve the ODE of Eq. (pde:diffeq) when
\[ f(x) = kx, \]
the boundary conditions are
\[ \frac{\partial \phi}{\partial x}(0) = p ; \quad \phi(L) = q \]
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 \( \alpha = -4.0 ; \beta = 4.0 ; \ L = 10.0 ; \ p = q = 0 \)
Optional: (a) Find an analytic solution to the equation in this exercise and compare the exact
results with the approximate results generated by your modification of \( \text{fdm1d} \).
(b) Find the points
at which the solution has the value zero, both starting with the solution obtained in the main
exercise and working from the exact solution obtained in optional part (a).

15.2. Suppose the one-dimensional string of length \( l \) discussed in Section 15.1.1 hangs vertically
and is acted on by gravity. Suppose that \( u(x,t) \) and \( v(x,t) \) give the transverse (horizontal) and longitudinal
(vertical) displacements of the particle of the string nominally located at \( x \), which is measured
downward from the top of the string. (a) Examine the forces acting on this string, deduce the
general equations for both longitudinal and transverse motion of the string, and show ultimately
that, if the motion is entirely transverse and the amplitude of the motion is small, the equations
reduce to
\[ \rho(x) \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x} \left( \tau(x) \frac{\partial u}{\partial x} \right) ; \quad \tau(x) = -g \int_{x}^{L} \rho(x') dx' \]
where \( \rho(x) \) is the mass per unit length of the string. (b) Taking \( \rho \) to be constant,
show that the tension \( \tau(x) \), which is simply equal to the weight of the string below \( x \), is given by \( \tau(x) = \rho g (l - x) \).
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 \( \overline{x} = x/l \) and \( \overline{t} = t \sqrt{g/l} \).
(d) Suppose you seek a sinusoidal solution for which \( u(\overline{x}, \overline{t}) = f(\overline{x}) \cos \omega \overline{t} \). Find the ODE satisfied
by \( f(\overline{x}) \) and then introduce the new independent variable \( y \) defined by \( y^2 = 1 - \overline{x} \) to find that,
expressed as a function of \( y, f \) satisfies the Bessel equation
\[ y^2 \frac{d^2 f}{dy^2} + y \frac{df}{dy} - 4 \omega^2 y^2 f = 0 \]
Note: The variable transformation \( y^2 = 1 - x \) in effect recognizes that the vertical string is more
appropriately treated by locating points on the string relative to the bottom rather than the top of
the string!

15.3. Show that every second-order linear ODE can be cast in self-adjoint form, i.e., show that functions
\( \alpha(x), \gamma(x) \), and \( f(x) \) can be found such that the general second-order linear ODE
\[ a(x) \frac{d^2 \phi}{dx^2} + b(x) \frac{d\phi}{dx} + c(x) \phi = g(x) \]
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 $f_{dm1d}$ to solve the equation

$$\frac{d^2u}{dx^2} + kxu = 0$$

over the interval $0 \leq x \leq L$ subject to the boundary conditions $u(0) = 0.0$ and $du(L)/dx = 1.0$, arranging your program so the value of $k$ is entered at execution time. Then, explore the solution in some detail for various values of $k$, making sure to assess the accuracy of your solution.

15.5. In (two-dimensional) polar coordinates $(\nabla, \phi)$, the Laplace equation is

$$\nabla^2 u = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \phi^2} = 0 \quad \Rightarrow \quad \nabla^2 \frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial \theta^2} = 0$$

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 $\Delta \phi = 2\pi/m$ and then let $\nabla_i = a + i \Delta \nabla, i = 0, 1, 2, \ldots, n$ and $\phi_j = j \Delta \phi, j = 0, 1, 2, \ldots, m$. Finally let $u_{i,j} = u(\nabla_i, \phi_j)$. Discretize this equation. In particular use a central difference formula to approximate the derivative $\partial u/\partial \nabla$. 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

$$f(\phi) = 0 \quad ; \quad g(\phi) = \begin{cases} -100 & \quad -\pi < \phi < 0 \\ 100 & \quad 0 \leq \phi \leq \pi \end{cases}$$

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

$$\frac{d^2 u}{dx^2} + k^2 u = 0 \quad ; \quad u(0) = u(L) = 0$$

(a) Find the analytic solution to this problem and identify the special values of $k$ that permit non-trivial solutions. (b) Show that the finite difference approach to the problem leads to the matrix eigenvalue problem $Au = -k^2 L^2 u/N^2$ where $N$ is the number of equal-length segments into which
the length $L$ of the domain is divided and

$$
A = \begin{pmatrix}
-2 & 1 & 0 & \cdots & 0 & 0 & 0 \\
1 & -2 & 1 & \cdots & 0 & 0 & 0 \\
0 & 1 & -2 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -2 & 1 & 0 \\
0 & 0 & 0 & \cdots & 1 & -2 & 1 \\
0 & 0 & 0 & \cdots & 0 & 1 & -2 \\
\end{pmatrix}
; \quad u = \begin{pmatrix}
u_1 \\
u_2 \\
u_3 \\
\vdots \\
u_{N-3} \\
u_{N-2} \\
u_{N-1}
\end{pmatrix}
$$

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

$$\frac{d^2u}{dx^2} + k^2u = 0 \quad ; \quad u(0) = u(L) = 0$$

(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

$$
A = \begin{pmatrix}
-2 & 1 & 0 & \cdots & 0 & 0 & 0 \\
1 & -2 & 1 & \cdots & 0 & 0 & 0 \\
0 & 1 & -2 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -2 & 1 & 0 \\
0 & 0 & 0 & \cdots & 1 & -2 & 1 \\
0 & 0 & 0 & \cdots & 0 & 1 & -2 \\
\end{pmatrix}
; \quad B = \begin{pmatrix}
4 & 1 & 0 & \cdots & 0 & 0 & 0 \\
1 & 4 & 1 & \cdots & 0 & 0 & 0 \\
0 & 1 & 4 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 4 & 1 & 0 \\
0 & 0 & 0 & \cdots & 1 & 4 & 1 \\
0 & 0 & 0 & \cdots & 0 & 1 & 4 \\
\end{pmatrix}
$$

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

$$N_1^{(e)}(x) = \frac{1}{\Delta} \begin{vmatrix} 1 & x_1^{(e)} & x^2 \\ 1 & x_2^{(e)} & (x_2^{(e)})^2 \\ 1 & x_3^{(e)} & (x_3^{(e)})^2 \end{vmatrix}$$

$$N_2^{(e)}(x) = \frac{1}{\Delta} \begin{vmatrix} 1 & x_1^{(e)} & (x_1^{(e)})^2 \\ 1 & x & x^2 \\ 1 & x_3^{(e)} & (x_3^{(e)})^2 \end{vmatrix}$$

$$N_3^{(e)}(x) = \frac{1}{\Delta} \begin{vmatrix} 1 & x_1^{(e)} & (x_1^{(e)})^2 \\ 1 & x_2^{(e)} & (x_2^{(e)})^2 \\ 1 & x & x^2 \end{vmatrix}$$

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

$$N_1^{(e)}(x) = 2(\xi - 1)\left(\xi - \frac{1}{2}\right) ; \ N_2^{(e)}(x) = 4\xi(1 - \xi) ; \ N_3^{(e)}(x) = 2\xi\left(\xi - \frac{1}{2}\right)$$

when $x_2^{(e)}$ is midway between $x_1^{(e)}$ and $x_3^{(e)}$, i.e., when $x_2^{(e)} = x_1^{(e)} + \frac{1}{2}l^{(e)}$ and $x_3^{(e)} = x_1^{(e)} + l^{(e)}$, and (d) obtain graphs of these three shape functions over the interval $0 < \xi < 1$. Hint: You may find a symbolic manipulating program to be useful at many points in this problem.

15.11. Accepting the shape functions given in part (c) of Exercise 15.10, (a) find the $3 \times 3$ matrices

$$^1K_{ij}^{(e)} = \int_{x_1^{(e)}}^{x_3^{(e)}} \frac{dN_i}{dx} \frac{dN_j}{dx} \ dx \quad \text{and} \quad ^2K_{ij}^{(e)} = \int_{x_1^{(e)}}^{x_3^{(e)}} N_i N_j \ dx$$

and then construct the $3 \times 3$ matrix whose elements are

$$K_{ij}^{(e)} = \alpha^{(e)} ^1K_{ij}^{(e)} + \beta^{(e)} ^2K_{ij}^{(e)}$$

to produce the (elemental) stiffness matrix analogous to Eq. (15.98). (b) Find the three-element vector whose elements are

$$b_i^{(e)} = f^{(e)} \int_{x_1^{(e)}}^{x_3^{(e)}} N_i^{(e)} \ dx$$

analogous to Eq. (15.99) and the three element vector whose elements are

$$g_i^{(e)} = \alpha^{(e)} N_i^{(e)} \frac{d^2}{dx^2} \bigg|_{x=x_3^{(e)}} - \alpha^{(e)} N_i^{(e)} \frac{d^2}{dx^2} \bigg|_{x=x_1^{(e)}}.$$
analogous to Eq. (15.100). By the time you are done, you should discover that the elemental
equation applicable to this three-node element is
\[
\begin{bmatrix}
K_{11}^{(e)} & K_{12}^{(e)} & K_{13}^{(e)} \\
K_{21}^{(e)} & K_{22}^{(e)} & K_{23}^{(e)} \\
K_{31}^{(e)} & K_{32}^{(e)} & K_{33}^{(e)}
\end{bmatrix}
\begin{bmatrix}
\tilde{\varphi}_1^{(e)} \\
\tilde{\varphi}_2^{(e)} \\
\tilde{\varphi}_3^{(e)}
\end{bmatrix}
= \begin{bmatrix}
b_1^{(e)} \\
b_2^{(e)} \\
b_3^{(e)}
\end{bmatrix}
+ \begin{bmatrix}
g_1^{(e)} \\
g_2^{(e)} \\
g_3^{(e)}
\end{bmatrix}
\]
where
\[
g_1^{(e)} = -\alpha^{(e)} \frac{d\tilde{\varphi}_1^{(e)}}{dx} \bigg|_{x_1^{(e)}} ; \quad g_3^{(e)} = \alpha^{(e)} \frac{d\tilde{\varphi}_3^{(e)}}{dx} \bigg|_{x_3^{(e)}} .
\]
Except for factors of $\alpha^{(e)}$, $\beta^{(e)}$, $f^{(e)}$, and $l^{(e)}$, you should have explicit values for $K_{ij}^{(e)}$ and $b_i^{(e)}$.
Hints: (1) Begin by recasting the integrals as integrals on the variable $\xi$. (2) You may find a symbolic
manipulating program to be useful at many points in this problem. (3) Assume as in
Exercise 15.10 that the point $x = x_2$ is midway between $x_1$ and $x_3$.

15.12. Continuing with the circumstances of the previous two problems, suppose the region of interest is
divided into three three-noded elements, $e = 1, 2, 3$ with global nodes 1, 2, 3, 4, 5, 6, 7, nodes 1, 2,
and 3 in element 1, nodes 3, 4, and 5 in element 2, and nodes 5, 6, and 7 in element 3. Suppose
node 2 is midway between nodes 1 and 3, node 4 is midway between nodes 3 and 5, and node 6 is
midway between nodes 5 and 7, but do not suppose the lengths of the three elements are the same.
Following the pattern in Section 15.9.4, assemble the elemental equations for these three elements
into an equation for the whole system analogous to Eq. (15.130) if the solution is required to satisfy
the boundary conditions of Eqs. (15.68) and (15.69).

15.13. Recast fem1d so it will generate a solution to Eq. (15.67) when

- $\alpha$ is a (positive) constant and $\beta = 0$,
- $f(x)$ varies with position in accordance with $f(x) = Ae^{-\sigma(x-L/x)^2}$, with $A$ and $\sigma$ constants,
  and
- the solution is to have the fixed value $\varphi = 0$ at $x = 0$ and the fixed value $\varphi = 100$ at $x = L$,
  i.e., we have a source that is concentrated near the middle of the region of interest and we
  impose Dirichlet boundary conditions on both ends.

Then, explore the character of the solutions for various values of $\sigma$. Use two-noded elements of equal
size. Note: This solution to this problem corresponds physically to the steady state temperature in
a one-dimensional rod whose ends are maintained at the fixed temperatures 0 and 100, respectively,
and whose middle is heated with a source that provides a constant energy input.

15.14. 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.

15.15. (a) Solve Eq. (15.179) for the constants $a^{(e)}$, $b^{(e)}$, and $c^{(e)}$ and then verify the expressions given
in Eqs. (15.182)–(15.184) for the shape functions. Remember that symbol manipulating programs
are available. (b) For each node $i$, verify that the function $N_i^{(e)}(x, y)$ as given by Eqs. (15.182) and
(15.184) is zero not only at the two nodes not identified by its index but also along the entire line
joining those two nodes. Hint: Set the determinant in the numerator of the expression giving $N_i^{(e)}$
to zero, thereby obtaining the equation of a line in the plane. Verify that 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

\[ N_0(x, y) = \frac{1}{\Delta} \begin{vmatrix} 1 & x & y & xy \\ 1 & x_1 & y_1 & x_1y_1 \\ 1 & x_2 & y_2 & x_2y_2 \\ 1 & x_3 & y_3 & x_3y_3 \end{vmatrix} \quad \text{where} \quad \Delta = \begin{vmatrix} 1 & x_0 & y_0 & x_0y_0 \\ 1 & x_1 & y_1 & x_1y_1 \\ 1 & x_2 & y_2 & x_2y_2 \\ 1 & x_3 & y_3 & x_3y_3 \end{vmatrix} \]

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

\[ N_0(\xi, \eta) = (1 - \xi)(1 - \eta) \quad ; \quad N_1(\xi, \eta) = \xi(1 - \eta) \quad ; \quad N_2(\xi, \eta) = \xi \eta \quad ; \quad N_3(\xi, \eta) = \eta(1 - \xi) \]

(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. \textit{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 \( u(x, 0) = 0 \), \( \partial u(x, 0)/\partial x = 0 \). The solution should, of course, be zero at all subsequent times, since we have started the string in its equilibrium position with zero velocity. Now, suppose that a computer roundoff error occurs such that, instead of being zero at all nodes, \( u(x, \Delta t) = 0 \) everywhere except at one node near the middle of the string and, at that node \( u(x, \Delta t) \) mistakenly acquires the value 1 (one). Use your program to solve this problem for choices of the parameters that make \( \alpha = 0.5 \) and solve it again for other choices that make \( \alpha = 1.5 \). Look at the solution with print or plot frequency 1 and compare the way in which that one disruption of a value at the end of the first time step propagates forward in time for the two values of \( \alpha \).

15.19. Proof that the full discretization of the diffusion equation as described towards the end of Section 15.16.2 leads to an unstable method unless \( \gamma \leq 1/2 \) is extremely difficult. (A proof is worked out in pages 92–98 of 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
\[ u_{i,j} \approx \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - \Delta x^2 r_{i,j}}{4 - k^2 \Delta x^2} \]

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. 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 relaxing the initial guess to a final solution. This formula estimates the next iterate at a particular node as the average of that node’s four nearest neighbors. An alternative approach, known as over-relaxation, includes a contribution from the current value in the node itself, specifically
\[ u_{i,j} = (1 - \alpha) u_{i,j} + \alpha \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}}{4} \]
where \( \alpha = 1 \) reduces this equation to Eq. (15.168) and \( 1 \leq \alpha < 2 \). 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 \textit{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 \textit{fdmlap2d} so that iteration stops when the largest change occurring at any single node from one iterate to the next does not exceed an externally prescribed tolerance. To avoid all possibility of an infinite loop, you should halt iteration either when the tolerance has been reached (or exceeded) or when some prescribed number of iterations has taken place. Code so that your program displays the actual tolerance achieved. Further, if execution terminates because the prescribed tolerance is not achieved, your program should print a message that alerts you to the fact that the prescribed tolerance was not achieved. Test your program with the same example as was used in the text. Hints: Before starting an iteration, set a variable, say \texttt{maxch}, equal to zero. As you calculate a new value for each node in the iteration, store the result in a temporary variable so you can compare that value with the old value it will replace, updating \texttt{maxch} to the absolute value of the difference between the new and the old values \textit{if and only if} that difference exceeds the difference already stored in \texttt{maxch}. Then, substitute the new value for the old in the array containing the evolving solution and go on to the next node. Once the iteration is completed, \texttt{maxch} will contain the absolute value of the largest change at any node during that single iteration. If \texttt{maxch} is less than the prespecified convergence criterion, stop the iteration; otherwise conduct one more iteration.

15.24. Consider the problem defined by the two-dimensional Poisson equation
\[ \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = -(1 - x^2)(1 - y^2) \]
to be solved in the square region \( R \) defined by \(-1 \leq x, y \leq 1\) subject to the Dirichlet boundary conditions requiring \( U = 0 \) on the entire boundary of \( R \). Assume all variables are dimensionless. (a) Obtain a surface plot of this inhomogeneity over the \( x, y \) plane. (b) Recasting \textit{fem2d} to incorporate the given boundary conditions and to address the non-zero inhomogeneity find, explore, and (using surface and contour plots as you deem appropriate) display the solution \( U(x, y) \) in the square region \( R \). Compare the resulting equation with Eq. (15.69) as a guide to interpreting the present problem as a modification of the problem addressed by \textit{fem2d}. (c) Find the two-dimensional vector field \( \mathbf{V} = -\text{grad} U = -\nabla U \) in \( R \) and display that field graphically in whatever ways seem appropriate.

\[^{58}\alpha \geq 2 \text{ generates an unstable algorithm.}\]
15.25. (a) Starting with the Taylor series

\[ \varphi(x + \Delta x) = \varphi(x) + \Delta x \varphi'(x) + O(\Delta x^2) \]

evaluate

\[ \varphi(x + \Delta x) - \varphi(x) \]

to show that

\[ \varphi'(x) = \frac{\varphi(x + \Delta x) - \varphi(x)}{\Delta x} + O(\Delta x) \]

That is, derive the forward difference formula in Eq. (15.74). (b) Deduce the more accurate forward difference formula

\[ \varphi'(x) = -\frac{\varphi(x + 2\Delta x) - 4\varphi(x + \Delta x) + 3\varphi(x)}{2\Delta x} + O(\Delta x^2) \]

*Hint:* Start by using the Taylor expansion

\[ f(x + h) = f(x) + hf'(x) + \frac{1}{2}h^2 f''(x) + O(h^3) \]

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

\[ u(0, t) = 0 \quad ; \quad u(L, t) = 0 \quad ; \quad u(x, 0) = f(x) \quad ; \quad \frac{\partial u(x, t)}{\partial t} \bigg|_{t=0} = 0 \]

These equations describe a string of length \( L \) that is fixed at both ends and set into motion by drawing it aside to the initial shape given by \( f(x) \) and released from rest. Of course, for consistency, \( f(0) \) and \( f(L) \) must both be zero, but \( f(x) \) is otherwise unconstrained. Then, develop a program analogous to `fdmwave1d` to solve this problem, testing your program with

\[ f(x) = A \sin \frac{n \pi x}{L} \]

for \( n = 1, 2, \) and 3. You should recognize that this exercise asks you to find the lowest three normal modes in the vibrations of this string. You should, of course, expect the results of your program to reveal what you know to be the periodic motion in these three modes of oscillation of the string. Your program should also allow you to confirm that the frequencies of the three modes \( f_1, f_2, \) and \( f_3 \) are related by \( f_2 = 2f_1 \) and \( f_3 = 3f_1 \). Agreement with these expectations provides some level of confidence in the adequacy of your choice of parameters.
15.A  Program(s) for FDM Approach to 1D Problem

15.A.1  Listing of fdm1d.pro (IDL)

; ***** Command file fdm1d.pro *****

; ***** 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 IDL’s main
; command level.

; ***** Request input of necessary parameters and assure that each
; is stored with the proper data type.

read, prompt='Enter number of segments (N): ', N & N = fix(N)
read, prompt='Enter alpha: ', alpha & alpha = float(alpha)
read, prompt='Enter beta: ', beta & beta = float(beta)
read, prompt='Enter f: ', f & f = float(f)
read, prompt='Enter L: ', L & L = float(L)
read, prompt='Enter p: ', p & p = float(p)
read, prompt='Enter gamma: ', gamma & gamma = float(gamma)
read, prompt='Enter q: ', q & q = float(q)

; ***** Calculate segment size, square of segment size, and
; coordinates of nodes.

dx = L/N & dx2 = dx^2
x = dx*findgen(N+1)

; ***** Determine vector of inhomogenieties and vectors containing
; diagonal, superdiagonal, and subdiagonal elements of
; coefficient matrix.

inhomo = fltarr(N+1) + 1.0
inhomo = f*dx2*inhomo
inhomo[0] = p & inhomo[N] = inhomo[N] + 2.0*q*dx

diag = fltarr(N+1) + 1.0
diag = (2.0*alpha + beta*dx2)*diag & diag[0]=1
diag[N] = diag[N] + 2.0*gamma*dx

supdiag = fltarr(N+1) + 1.0
supdiag = -alpha*supdiag & supdiag[0]=0

subdiag = fltarr(N+1) + 1.0
subdiag = -alpha*subdiag & subdiag[N] = -2.0*alpha

; ***** Solve system using trisol.

phi = trisol( subdiag, diag, supdiag, inhomo )
15.B Evaluating Integrals for 1D FEM Problem

15.B.3 ... using Mathematica

An appropriate batch file to use Mathematica to evaluate the important integrals in Section 15.1.7 is named FEM1DCalcs.mtm, can be copied from the directory $\$HEAD/mathematica, and invoked within Mathematica with the statement Get[ "FEM1DCalcs.mtm"] issued at Mathematica’s prompt for commands. The batch file is listed (without output) in the remainder of this subsection.

(* FEM1DCalcs.mtm *)

(* This batch file evaluates the several integrals that appear in *)
(* setting up coding to solve the 1D FEM problem in CPSUP. *)

(* Create interpolation functions and their derivatives *)

n[1] = (x2-x)/l; n[2] = (x-x1)/l;
dnx[1] = D[n[1], x]; dnx[2] = D[n[2], x];

(* Create integrand for K, then integrate and display *)

K1 = Table[dnx[i]*dnx[j], {i, 2}, {j, 2}];
K2 = Table[n[i]*n[j], {i, 2}, {j, 2}];
K = a*K1 + b*K2;
K = Integrate[K, {x, x1, x2}];
K = Simplify[K /. x2 -> x1 + l];
Print["K =", K // MatrixForm]

(* Evaluate and display b *)

b = Integrate[f*n[1]*n[2], {x, x1, x2}];
b = Simplify[b /. x2 -> x1 + l];
Print["b = ", b]

15.C Program(s) for FEM Approach to 1D Problem

15.C.1 Listing of fem1d.pro (IDL)

; ****** Command file fem1d.pro ******

; ****** 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 IDL’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.

read, prompt='Enter number of segments (M): ', M & M = fix(M)
read, prompt='Enter alpha: ', alpha & alpha = float(alpha)
read, prompt='Enter beta: ', beta & beta = float(beta)
read, prompt='Enter f: ', f & f = float(f)
read, prompt='Enter l: ', l & l = float(L)
read, prompt='Enter p: ', p & p = float(p)
read, prompt='Enter gamma: ', gamma & gamma = float(gamma)
read, prompt='Enter q: ', q & q = float(q)
length = l*M
x = l*findgen(M+1)

; BLOCK 2 Determine coefficient (stiffness) matrix before incorporation
; of boundary conditions.
K = fltarr(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[0,0] = S ; Set diagonal elements of K
for i = 1, M-1 do K[i,i] = S2
K[M,M] = S

for i = 0, M-1 do K[i+1,i] = T ; Set elements above and below
for i = 0, M-1 do K[i,i+1] = T ; main diagonal of K

; BLOCK 3 Create vector of inhomogeneties.
B = fltarr(M+1) ; Create M+1 element vector of zeros
U = f*l/2.0 ; Evaluate common quantities
U2 = 2.0*U
B[0] = U ; Set elements of B
for i = 1, M-1 do B[i] = U2
B[M] = U

; BLOCK 4 Incorporate mixed boundary conditions.

; BLOCK 5 Incorporate Dirichlet boundary conditions.
K[0,0] = 1.0
B[0] = p
for j = 1, M do K[j,0] = 0.0
for i = 1, M do B[i] = B[i] - K[0,i]*p
for i = 1, M do K[0,i] = 0.0

; BLOCK 6 Solve system using LU decomposition.
ludc, K, index ; Do LU decomposition of K
phi = lusol(K, index, B) ; Solve equation
15.D  IDL Programs for FDM Approach to 2D Problems

15.D.1  Listing of fdmwave1d.pro

; ***** Command file fdmwave1d.pro

; ***** Request input of necessary parameters and assure that
; each is stored with the proper data type.
read, prompt='Number of segments: ', N & N = fix(N)
read, prompt='Time step: ', dt & dt = float(dt)
read, prompt='Number of time steps ', T & T = fix(T)
read, prompt='Speed of propagation: ', c & c = float(c)
read, prompt='Length of string: ', L & L = float(L)
read, prompt='Delay between plots (secs): ', delay & delay = float(delay)
read, prompt='Plot frequency: ', f & f = fix(f)

; ***** Determine segment length, set independent variables,
; and evaluate and display parameter; test for stability.

dx = L/N
x = dx*findgen(N+1)
alpha = c^2*dt^2/dx^2

print, 'alpha = ', alpha
if alpha GT 1.0 then begin
   print, 'alpha > 1; execution halted' & stop &
endif

; ***** Define vectors for past, current, and new solutions

u1 = fltarr( N+1 ) ; For past solution
u2 = fltarr( N+1 ) ; For current solution
u3 = fltarr( N+1 ) ; For next solution

; ***** Set and display initial displacement

b = 8.0*!pi/L
for i=3*N/8+1,5*N/8 do u1[i] = 1.0 + cos(b*(x[i]-L/2.0))

!y.range = [ -2.0,2.0 ]
plot, x, u1, thick=3.0 & wait, delay

; ***** Set solution at time dt with initial velocity
; equal to zero; display if requested

u2 = u1
if f EQ 1 then begin plot, x, u2, thick=3.0 & wait, delay & endif

; ***** Calculate solution repeatedly, displaying every
; f-th result

for j=2,T do begin
   u3[0] = 0 &
end
for i=1,N-1 do u3[i] = alpha*u2[i+1] + 2.0*(1.0-alpha)*u2[i] + alpha*u2[i-1] - u1[i] &
  u3[N] = 2*alpha*u2[N-1]+2*(1.0-alpha)*u2[N]- u1[N] &
if f*(j/f) EQ j then begin $
   plot, x, u3, thick=3.0 & $
   wait, delay & $
endif & $
   u1 = u2 & u2 = u3 &$
endfor

15.D.2 Listing of fdmdiffus1d.pro

; ***** Command file fdmdiffus1d.pro

; ***** Request input of necessary parameters and assure that each
; is stored with the proper data type.
read, prompt='Number of segments: ', N & N = fix(N)
read, prompt='Time step: ', dt & dt = float(dt)
read, prompt='Number of time steps ', T & T = fix(T)
read, prompt='Value of alpha: ', alpha & alpha = float(alpha)
read, prompt='Length of rod: ', L & L = float(L)
read, prompt='Delay between plots (secs): ', delay & delay = float(delay)
read, prompt='Plot frequency): ', f & f = fix(f)

; ***** Determine segment length, set independent variables,
; and evaluate and display parameter; test for stability.
dx = L/N
x = dx*findgen(N+1)
gamma = alpha^2*dt/dx^2
print, 'gamma = ', gamma
if gamma GT 0.5 then begin $
   print, 'gamma > 0.5; execution halted' & stop & $
endif

; ***** Define vectors for current and new solutions
u1 = fltarr( N+1 ) ; For current solution
u2 = fltarr( N+1 ) ; For next solution

; ***** Set and display initial temperature distribution
b = 8.0*pi/L
for i=3*N/8,5*N/8 do u1[i] = 1.0 + cos(b*(x[i]-L/2.0))
!y.range = [ 0.0,2.0 ]
plot, x, u1, thick=5.0, title='t = 0 s', charsize=1.5
wait, delay

; ***** Calculate solution repeatedly, displaying every
for j=1,T do begin $
    u2[0] = 0.0 &$
    for i=1,N-1 do $u2[i] = gamma*u1[i-1]$
        + (1.0-2.0*gamma)*u1[i] + gamma*u1[i+1] &$
    u2[N] = 2*gamma*u1[N-1] + (1-2*gamma)*u1[n] &$
    if f*(j/f) EQ j then begin $
        tmp = 't ='+string(j*dt,format='(I5)')+' s' &$
        plot, x, u2, thick=5.0, title=tmp, charsize=1.5 &$
        wait, delay & endif &$
    u1 = u2 &$
endfor

15.D.3 Listing of fdmlap2d.pro

; ***** Command file fdmlap2d.pro *****

; ***** Acquire controlling parameters *****
read, prompt='Number of segments: ', N & N = fix(N)
read, prompt='Length of side: ', L & L = float(L)
read, prompt='Maximum number of iterations: ', maxits & maxits=fix(maxits)
read, prompt='Display frequency: ', f & f=fix(f)

; ***** Calculate grid spacing, values of x and y at grid points *****

dx = L/N
x = dx*findgen(N+1) & y = x

; ***** Create and initialize array for solution *****

u = fltarr(N+1, N+1)
u[*,0] = 100.0
u[0,*] = 100.0 - 100.0*x/L

; ***** Display solution on the screen *****

print, 'Iteration ', 0 & print, u & print

; ***** Iterate the specified number of times, displaying
; every f-th iterate on the screen *****

for itcnt=1, maxits do begin $
    for i = 1, N-1 do $
        for j = 1, N-1 do $
            u[i,j] = 0.25*(u[i+1,j] + u[i-1,j] + u[i,j+1] + u[i,j-1]) &$
    for j = 1, N-1 do $u[N,j] = 0.25*(2.0*u[N-1,j] + u[N,j+1] + u[N,j-1]) &$
    if f*(itcnt/f) EQ itcnt then begin $
        print, 'Iteration ', itcnt & print, u & print &$
    endif &$
endfor
15.4  Listing of fdmlap2d_plot.pro

```plaintext
read, prompt='Number of segments: ', N & N = fix(N)
read, prompt='Length of side: ', L & L = float(L)
read, prompt='Maximum number of iterations: ', maxits & maxits=fix(maxits)
read, prompt='Display frequency: ', f & f=fix(f)
read, prompt='Delay between plots (secs): ', delay & delay = float(delay)

dx = L/N
x = dx*findgen(N+1) & y = x

u = fltarr(N+1, N+1)
u[*,0] = 100.0
u[0,*] = 100.0 - 100.0*x/L

surface, reverse(u,2), x, y, title='Iterate 0', charsize=2.5, thick=3, xtitle='x', ytitle='y', ztitle='u'
wait, delay

for itcnt=1, maxits do begin
  for i = 1, N-1 do $
    for j = 1, N-1 do $
      u[i,j] = 0.25*(u[i+1,j] + u[i-1,j] + u[i,j+1] + u[i,j-1]) & $
  for j = 1, N-1 do u[N,j] = 0.25*(2.0*u[N-1,j] + u[N,j+1] + u[N,j-1]) & $
  if f*(itcnt/f) EQ itcnt then begin
    tmp = 'Iterate '+string(itcnt, format='(I4)') & $
    surface, reverse(u,2), x, y, title= tmp, charsize=2.5, xtitle='x', ytitle='y', ztitle='u' & $
    wait, delay & $
  endif & $
endfor

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

\[ I = \int \int_{\Omega} \psi \left( \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} \right) \, dx \, dy \]  \hspace{1cm} (15.225)

where \( \psi, V_x, \) and \( V_y \) are functions of \( x \) and \( y \), and \( \Omega \) is the region in the \( xy \) plane over which the 2D integral is to be evaluated. Recast the integral in the form

\[ I = \int \int_{\Omega} \left( \frac{\partial(\psi V_x)}{\partial x} + \frac{\partial(\psi V_y)}{\partial y} \right) \, dx \, dy - \int \int_{\Omega} \left( V_x \frac{\partial \psi}{\partial x} + V_y \frac{\partial \psi}{\partial y} \right) \, dx \, dy \]  \hspace{1cm} (15.226)

Now, introduce the vector \( \mathbf{Q} \) whose \( x \) and \( y \) components are \(-\psi V_y\) and \(\psi V_x\), respectively. Since

\[ \frac{\partial(\psi V_x)}{\partial x} + \frac{\partial(\psi V_y)}{\partial y} = \frac{\partial Q_y}{\partial x} - \frac{\partial Q_x}{\partial y} = (\nabla \times \mathbf{Q})_z = (\nabla \times \mathbf{Q}) \cdot \mathbf{k} \]  \hspace{1cm} (15.227)

the integral of concern then can be written in the form

\[ I = \int \int_{\Omega} (\nabla \times \mathbf{Q}) \cdot \mathbf{k} \, dx \, dy - \int \int_{\Omega} \left( V_x \frac{\partial \psi}{\partial x} + V_y \frac{\partial \psi}{\partial y} \right) \, dx \, dy \]  \hspace{1cm} (15.228)
```
which, upon invoking Stokes’ theorem, we can rewrite further in the form

\[ I = \oint_{\Gamma} Q \cdot \hat{t} \, dl - \int_{\Omega} \left( V_x \frac{\partial \psi}{\partial x} + V_y \frac{\partial \psi}{\partial y} \right) \, dx \, dy \quad (15.229) \]

where \( \Gamma \) is the path in the \( xy \) plane bounding the region \( \Omega \) and \( \hat{t} \) is a unit vector tangent to that path and pointing in the direction of the thumb of the right hand when the fingers grasp the path while piercing the region \( \Omega \) in the \( z \) direction. The one-dimensional integral in this result is the analog of the unintegrated term in the more familiar one-dimensional formula for integration by parts.

That one dimensional integral involves the component of the vector \( Q \) 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

\[ \hat{t} = \hat{k} \times \hat{n} = \hat{k} \times (n_x \hat{i} + n_y \hat{j}) = -n_y \hat{i} + n_x \hat{j} \quad (15.230) \]

Thus,

\[ Q \cdot \hat{t} = n_x Q_y - n_y Q_x = n_x \psi V_x + n_y \psi V_y = (\psi V_x \hat{i} + \psi V_y \hat{j}) \cdot \hat{n} = \psi (V_x \hat{i} + V_y \hat{j}) \cdot \hat{n} \quad (15.231) \]

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

\[ I = -\int_{\Omega} \left( V_x \frac{\partial \psi}{\partial x} + V_y \frac{\partial \psi}{\partial y} \right) \, dx \, dy + \oint_{\Gamma} (\psi V_x \hat{i} + \psi V_y \hat{j}) \cdot \hat{n} \, dl \quad (15.232) \]

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.3 ... using Mathematica

An appropriate batch file to use Mathematica to evaluate the necessary integrals when array indices start at 0 is named FEM2DCalcs0.mtm, can be copied from the directory $HEAD/mathematica, and invoked within Mathematica with the statement

\[ \text{Get} [ \ "FEM2DCalcs0.mtm" ] \]

issued at Mathematica’s prompt for commands. The batch file includes the statements

\[
(* \text{FEM2DCalcs0.mtm} *)
\]

\[
(* \text{This batch file evaluates the several integrals that appear in} *)
\]

\[
(* \text{setting up coding to solve the 2D FEM problem in CPSUP when} *)
\]

\[
(* \text{array indices start at} 0. *)
\]

\[
(* \text{Create interpolation functions and their derivatives} *)
\]

\[
r = \{ 1, x, y \}
\]

\[
r0 = \{ 1, x0, y0 \}
\]

\[
r1 = \{ 1, x1, y1 \}
\]

\[
r2 = \{ 1, x2, y2 \}
\]

\[
\text{CapDelta} = \text{Det}[ \{r0, r1, r2\} ]
\]
\[ \begin{align*}
\text{n0} &= \text{Det} \left[ \{ r, r1, r2 \} \right] / \text{CapDelta} \quad \text{(Evaluate functions)} \\
\text{n1} &= \text{Det} \left[ \{ r0, r, r2 \} \right] / \text{CapDelta} \\
\text{n2} &= \text{Det} \left[ \{ r0, r1, r \} \right] / \text{CapDelta} \\
\text{n} &= \{ \text{n0}, \text{n1}, \text{n2} \} \\
\end{align*} \]

\[ \text{(* Simplify functions and evaluate derivatives *)} \]
\[ \text{(* for odd-numbered elements *)} \]

\[ \text{sb} = \{ x1 \rightarrow x0 + dx, x2 \rightarrow x0 + dx, y1 \rightarrow y0, y2 \rightarrow y0 + dx \} \]
\[ \text{nodd} = \text{Simplify} \left[ \text{n} / \text{sb} \right] \]
\[ \text{dnoddx} = \text{D} \left[ \text{nodd}, x \right] \]
\[ \text{dnoddy} = \text{D} \left[ \text{nodd}, y \right] \]

\[ \text{(* Simplify functions and evaluate derivatives *)} \]
\[ \text{(* for even-numbered elements *)} \]

\[ \text{sb} = \{ x1 \rightarrow x0 + dx, x2 \rightarrow x0, y1 \rightarrow y0+dx, y2 \rightarrow y0 + dx \} \]
\[ \text{neven} = \text{Simplify} \left[ \text{n} / \text{sb} \right] \]
\[ \text{dnevenx} = \text{D} \left[ \text{neven}, x \right] \]
\[ \text{dneveny} = \text{D} \left[ \text{neven}, y \right] \]

\[ \text{(* Create integrand for Kodd, then integrate *)} \]
\[ \text{Kodd1} = \text{Table} \left[ \text{dnoddx}[[i]]*\text{dnoddx}[[j]], \{ i, 3 \}, \{ j, 3 \} \right] \]
\[ \text{Kodd2} = \text{Table} \left[ \text{dnoddy}[[i]]*\text{dnoddy}[[j]], \{ i, 3 \}, \{ j, 3 \} \right] \]
\[ \text{Kodd3} = \text{Table} \left[ \text{nodd}[[i]]*\text{nodd}[[j]], \{ i, 3 \}, \{ j, 3 \} \right] \]
\[ \text{Kodd} = \text{ax}*\text{Kodd1} + \text{ay}*\text{Kodd2} + \text{b}*\text{Kodd3} \]
\[ \text{Kodd} = \text{Integrate} \left[ \text{Kodd}, \{ y, y0, y0+x-x0 \} \right] \]
\[ \text{Kodd} = \text{Simplify} \left[ \text{Kodd} \right] \]
\[ \text{Print} [ "Kodd =", \text{Kodd} // \text{MatrixForm} ] \]

\[ \text{(* Create integrand for Kodd, then integrate *)} \]
\[ \text{Keven1} = \text{Table} \left[ \text{dnevenx}[[i]]*\text{dnevenx}[[j]], \{ i, 3 \}, \{ j, 3 \} \right] \]
\[ \text{Keven2} = \text{Table} \left[ \text{dneveny}[[i]]*\text{dneveny}[[j]], \{ i, 3 \}, \{ j, 3 \} \right] \]
\[ \text{Keven3} = \text{Table} \left[ \text{neven}[[i]]*\text{neven}[[j]], \{ i, 3 \}, \{ j, 3 \} \right] \]
\[ \text{Keven} = \text{ax}*\text{Keven1} + \text{ay}*\text{Keven2} + \text{b}*\text{Keven3} \]
\[ \text{Keven} = \text{Integrate} \left[ \text{Keven}, \{ y, y0+x-x0, y0+dx \} \right] \]
\[ \text{Keven} = \text{Simplify} \left[ \text{Keven} \right] \]
\[ \text{Print} [ "Keven =", \text{Keven} // \text{MatrixForm} ] \]

\[ \text{(* Evaluate bodd and beven *)} \]
\[ \text{bodd} = \text{Integrate} \left[ \text{f*nodd}, \{ y, y0, y0+x-x0 \} \right] \]
\[ \text{bodd} = \text{Integrate} \left[ \text{bodd}, \{ x, x0, x0+dx \} \right] \]
\[ \text{bodd} = \text{Simplify} \left[ \text{bodd} \right] \]
\[ \text{Print} [ "bodd =", \text{bodd} ] \]

\[ \text{beven} = \text{Integrate} \left[ \text{f*neven}, \{ y, y0+x-x0, y0+dx \} \right] \]
beven = Integrate[ beven, {x, x0, x0+dx} ]
beven = Simplify[ beven ]
Print[ "beven = ", beven ]

(* Evaluate needed g's *)

g = nodd /. x -> x0+dx

g = Integrate[ q*g, {y, y0, y0+dx} ]
Print[ "g = ", g ]

An appropriate batch file to use Mathematica to evaluate the necessary integrals when array indices start at 1 is named FEM2DCalcs1.mpl, can be copied from the directory $HEAD/mathematica, and invoked within Mathematica with the statement batch( "FEM2DCalcs1.mtm" ) issued at Mathematica's prompt for commands. This batch file includes the statements
15.G Program(s) for FEM Approach to 2D Problems

15.G.1 Listing of fem2d.pro (IDL)

PRO fem2d, A

; ***** Enter parameters *****
read, prompt='Enter length of side (L): ', L & L = float(L)
read, prompt='Enter number of segments (d): ', d & d = fix(d)
read, prompt='Enter alpha_x: ', alpha_x & alpha_x = float(alpha_x)
read, prompt='Enter alpha_y: ', alpha_y & alpha_y = float(alpha_y)
read, prompt='Enter beta: ', beta & beta = float(beta)
read, prompt='Enter f: ', f & f = float(f)
read, prompt='Enter value for top edge: ', p1 & p1 = float(p1)
read, prompt='Enter value for bottom edge: ', p2 & p2 = float(p2)
read, prompt='Enter q: ', q & q = float(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 = fltarr(N) ; Create array to store x values
y = fltarr(N) ; Create array to store y values
ct=0 ; Initialize a counter variable
for i = 0, d do begin ; Start row number loop
  for j = 0, d do begin ; Start column number loop
    x[ct] = i*dx
    y[ct] = L - j*dx
    ct=ct+1
  endfor
endfor ; End loops

; ***** Create connectivity matrix *****

M = intarr(3,M) ; Create 3 by M null array
for e = 0, M-1 do begin ; Loop through all elements
  vl = e/(2*d)
  cm[0,e] = e/2 + 1 + vl
  if 2*(e/2) eq e then begin
    cm[1,e] = cm[0,e] + d
    cm[2,e] = cm[0,e] - 1
  endif else begin
    cm[1,e] = cm[0,e] + d + 1
    cm[2,e] = cm[0,e] + d
  endelse
endfor

; ***** Find [K] for odd and even numbered elements *****
Kodd = fltarr(3,3) ; Create two 3 by 3 arrays
Keven = fltarr(3,3)

bx = beta*dx^2 ; Evaluate a common quantity
Kodd[0,0] = (bx + 6*alpha_x)/12 ; Assign the appropriate value to each K[i,j]. Note that the array
Kodd[2,2] = (bx + 6*alpha_y)/12 ; are symmetric, and that Keven
Kodd[0,1] = (bx - 12*alpha_x)/24 ; includes all of the same values as
Kodd[1,2] = (bx - 12*alpha_y)/24 ; Kodd, but in different locations.
Kodd[0,2] = bx/24
Kodd[2,0] = bx/24

Keven[0,0] = Kodd[2,2]
Keven[1,1] = Kodd[0,0]
Keven[2,2] = Kodd[1,1]
Keven[0,1] = Kodd[0,2]
Keven[1,0] = Keven[0,1]
Keven[1,2] = Kodd[0,1]
Keven[2,1] = Keven[1,2]
Keven[0,2] = Kodd[1,2]
Keven[2,0] = Keven[0,2]

; ***** Create stiffness matrix and vector \{b\} *****

K = fltarr(N,N) ; Create arrays to store values of K[i,j] and b[i]
b = fltarr(N)

for e = 0,M-1 do begin ; Count through element numbers
  for i = 0,2 do begin ; 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 = 0,2 do $ ; correct locations in \[K\] and \{b\}
      if 2*(e/2) ne e then $
        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]
      $endfor
  endfor

; ***** Incorporate Dirichlet boundary conditions *****

for i = 0, d do begin
  u = i ; Nodes on the left boundary
  s = i*(d+1) ; Nodes on the top boundary
  t = (i+1)*(d+1)-1 ; Nodes on the bottom boundary
  p3 = (p1-p2)/L * y[u] + p2 ; Find values of phi on left boundary
  for j = 0, N-1 do begin
    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
15.G. PROGRAM(S) FOR FEM APPROACH TO 2D PROBLEMS

\[ \begin{align*}
  b[t] &= p2 \\
  b[u] &= p3 \\
  \text{if } j \neq s \text{ then } b[j] &= b[j] - K[s,j]p1 \quad ; \text{Reflect influence} \\
  K[s,j] &= 0 \quad ; \text{of known values of} \\
  \text{if } j \neq t \text{ then } b[j] &= b[j] - K[t,j]p2 \quad ; \text{phi in the other} \\
  K[t,j] &= 0 \quad ; \text{equations} \\
  \text{if } j \neq u \text{ then } b[j] &= b[j] - K[u,j]p3 \\
  K[u,j] &= 0
\end{align*} \]

}whilefor

\[ \begin{align*}
  K[s,s] &= 1 \quad ; \text{Set the appropriate entry to 1 in} \\
  K[t,t] &= 1 \quad ; \text{the rows where phi is known} \\
  K[u,u] &= 1
\end{align*} \]

}whilefor

; ***** Incorporate Neumann boundary condition *****

\[ \begin{align*}
  g &= \text{fltarr}(N) \quad ; \text{Create vector to store g values} \\
  \text{for } i = 1,d-1 \text{ do begin} \\
  \quad \text{nd} &= d*(d+1)+i \quad ; \text{Nodes on right boundary} \\
  \quad g[nd] &= q \times dx \\
  \text{endfor} \\
  b &= b + g \quad ; \text{Store the vector \{b+g\} in \{b\}}
\end{align*} \]

; ***** Solve the system of equations *****

\[ \begin{align*}
  \phi &= \text{fltarr}(N) \quad ; \text{Create vector to store results} \\
  \text{LUDC, K, INDEX} \quad ; \text{Solve the system of equations and} \\
  \phi &= \text{LUSOL}(K, INDEX, b) \quad ; \text{write the results to phi}
\end{align*} \]

; ***** Store solution and node locations in an array *****

\[ \begin{align*}
  A &= \text{fltarr}(d+1,d+1) \quad ; \text{Create a d+1 by d+1 array} \\
  \text{cnt}=0 \quad ; \text{Initialize a counter variable} \\
  \text{for } i = 0,d \text{ do for } j = d,0,-1 \text{ do begin} \\
  \quad A(i,j) &= \phi(cnt) \quad ; \text{Use nested for loops to write all} \\
  \quad \text{cnt} &= \text{cnt}+1 \quad ; \text{entries in A} \\
  \text{endfor} \\
  \text{End loop}
\end{align*} \]

END
Appendix A

Introduction to \LaTeX

Note: All \LaTeX\ source files (\texttt{.tex}, \texttt{.template}), all Windows batch files (\texttt{.bat}) program (\texttt{.pro}) files, and all UNIX command files (no file type) referred to in this chapter are available in the directory \texttt{$\texttt{SHEAD/tex}$}, where (as defined in the \textit{Local Guide}) \texttt{SHEAD} must be replaced by the appropriate path for your site.

During the 1970s, when Donald Knuth (a Stanford University computer scientist) was creating his monumental, several-volume series of books on all aspects of computer science, he recognized the need for a computer-based type-setting/word-processing system tailored to the needs of authors of technical manuscripts. Interrupting his main project, he developed \TeX\ as the essential engine to support computerized technical type setting. The first version of \TeX\ was made available to the using public in 1978 and a revised (and final) version was published in 1984. Written by Leslie Lamport, \LaTeX\ is a versatile and extensive collection of macros that sit on top of \TeX\ and facilitate exploitation of \TeX\’s capabilities. Broadly, \LaTeX\ is a document preparation system that formats equations, tables, and illustrations as easily as plain text. That the whole complex (\TeX, \LaTeX, and many other components) has been deliberately placed in the public domain,\footnote{The primary site for information (history, current plans, downloads, ...) for \TeX, \LaTeX, and numerous other publicly available components of \TeX and its derivatives is the web site of the \TeX Users’ Group (TUG), \url{www.tug.org}. This organization maintains CTAN (the Comprehensive \TeX Archive Network), which has a handful of backbone machines around the world and a number of mirror sites, from any of which an enormous number of files associated with the \TeX/\LaTeX system can be downloaded.} 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 \textit{wysiwyg},\footnote{As per Knuth’s instruction (see the first two paragraphs of Chapter 1 of \textit{The \TeX\book} as identified in Section A.18), \TeX is pronounced ‘tech’ as in ‘technology’.} 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 \textit{\LaTeX\ Companion}, and/or to any of the several other \TeX and \LaTeX books.\footnote{You may also find the web pages of TUG at the URL \url{http://www.tug.org} to be valuable. One link on that page points to an engaging description of the history of the development of \TeX.} 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.\footnote{What you see is what you get.}

\footnote{Here, individuals disagree on the pronunciation of the ‘la’, and Leslie Lamport offers no guidance. Some say \textit{lah\TeX} while others say \textit{lay\TeX}, with the emphasis on the first syllable in both cases. We simply must become accustomed to each other’s preferences.}

\footnote{You may also find the web pages of TUG at the URL \url{http://www.tug.org} to be valuable. One link on that page points to an engaging description of the history of the development of \TeX.}
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. Finally, the PostScript or PDF file will be viewed on the screen or sent to a printer. In this section, we describe how to structure the code for a very simple document and then how to carry out the remaining steps to convert that code into a displayed or printed document. The remainder of this Appendix will explain numerous embellishments that might be invoked in the code. Processing that code to produce the final document is (more or less) independent of the complexity of the code.

To keep the discussion comparatively simple, we assume initially that your document incorporates no figures, does not have internal cross references, and includes neither a table of contents or an index. Those embellishments will be described in later sections.

A.1.1 Structuring a \LaTeX{} Source File

As input, \LaTeX{} requires an ASCII file, which can be created with any text editor (\texttt{xemacs}, \texttt{notepad}, \texttt{wordpad}, \texttt{winedt}, \texttt{gedit}, ...) and which contains both the text of the desired document and embedded formatting commands. All of \LaTeX{}’s commands begin with a backslash (\texttt{\}). \LaTeX{} distinguishes upper and lower case letters; while most standard commands and parameters use exclusively lower case letters, we must nonetheless pay attention to case.

Three commands are sufficient to create the simplest document containing straight text formatted with \LaTeX{}’s defaults. The very first required command in the code for \textit{any} \LaTeX{} document specifies the document \textit{class} and has the form\footnote{Be aware that some of the features described in this Appendix are specific to \LaTeX{}2ε and will not work with the previous version (\LaTeX{} 2.09). \LaTeX{}2ε, however, will automatically enter 2.09 emulation mode if the first line in the code is the now obsolete command \texttt{\documentstyle} instead of the new command \texttt{\documentclass}.}

\begin{verbatim}
\documentclass{Class} or \documentclass[options]{Class}
\end{verbatim}

where

- \textit{Class} is any one of the options shown in Table A.1 and
- \textit{options} may be omitted altogether if the defaults are acceptable. The most commonly used options, one or more of which may be separated by commas in a string containing no spaces, are listed in Table A.2.

The document itself must be enclosed between the second and third of the essential commands, the second marking the beginning of the document proper and having the form

\begin{verbatim}
\begin{document}
\end{verbatim}

and the third marking the end of the document and having the form

\begin{verbatim}
\end{document}
\end{verbatim}

\footnote{Other formats (HTML, ePub, ...) may also be produced, though this Appendix will not address those alternatives.}
Table A.1: Document classes valid in standard \texttt{\LaTeX}. Other available classes may be described in the \textit{Local Guide}.

\begin{verbatim}
\textbf{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 \textit{The \LaTeX Manual}.

\textbf{report} This class is generally used for longer documents. See Sections 2.2.2 and C.5.1 in \textit{The \LaTeX Manual}.

\textbf{book} This class is meant for actual books. See Sections 5.1 and C.5.1 in \textit{The \LaTeX Manual}.

\textbf{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 \textit{The \LaTeX Manual}.

\textbf{letter} This class facilitates creating letters. See Sections 5.3 and C.5.1 in \textit{The \LaTeX Manual}.
\end{verbatim}

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 \textit{mandatory} in all documents. Supplemented by the additional feature that a blank line in the code triggers a new paragraph, they are also \textit{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.

\documentclass{article}
\begin{document}

This sample illustrates the simplest code containing the mandatory commands and a brief text. Since no optional formatting commands have been included, the text will be formatted using all of the built-in defaults (margins, paragraph indent, type size and font, etc.).

The blank line preceding this line in the code will trigger a new paragraph. Each paragraph is indented, but note that there is no extra space between paragraphs. Note also that the lines in the code can be quite ragged; they will be filled and justified in the processing that produces the final copy.

\end{document}

A.1.2 Creating a PostScript File

Converting the \LaTeX source file into a PostScript file describing the desired document involves two steps:

1. First, we must “compile” the code by processing the file with \LaTeX, a task accomplished either (1) by typing a command like

latex filename

( Default extension for the input file is .tex.)

where filename specifies the input file to be processed, or (2) by selecting an item from a menu in a graphical user interface (GUI).\footnote{See the Local Guide for the precise command at your site.} However \LaTeX is invoked, it will create several output files, all of which will have the same name as the input file except for the extension: (1) a binary output file, with extension .dvi, which contains the translation of the original file into \TeX’s \texttt{d}e\texttt{v}i\texttt{e}ndependent 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.\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.} The resulting .dvi file can be directly displayed on the screen (see Section A.1.4).\footnote{Remember that we have limited the present discussion to documents containing no figures. See Section A.7 for the adjustments to be made if PostScript or PDF figures are included.}

2. Second, we must translate the .dvi file into a file that can be understood by the printer on which the document is to be printed. Most often, the .dvi file will be translated into
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 \texttt{dvips}. The necessary translation will be achieved either (1) by typing a command like

\begin{verbatim}
dvips -o filename.ps -t letter filename
\end{verbatim}

which translates the entire file into PostScript, stores it in a file, and (with the \texttt{-o} option here used) gives the stored file the same name as the input file but with extension \texttt{.ps} and with the \texttt{-t} option forces use of letter size paper\footnote{If the explicit specification of the output file is omitted, \texttt{dvips} will attempt to send the file directly to the printer—and the whole operation will probably fail. Further, the PostScript file will then not be stored for subsequent use.} or (2) by selecting an item from a menu.\footnote{This option is often the default. The option \texttt{-t landscape} will override the default portrait orientation.}

To be more explicit, the file \texttt{texsample1.tex}, contains the \LaTeX\ code in Table A.3. Once this file has been copied to the default directory, we would produce the \texttt{.dvi} and \texttt{.ps} files with statements like

\begin{verbatim}
latex texsample1
dvips -o texsample1.ps -t letter texsample1
\end{verbatim}

or equivalent selections from a menu.

\subsection*{A.1.3 Creating a PDF File}

PDF files can be created from the \LaTeX\ code in at least two ways:

1. Starting with \LaTeX\ source file created in Section A.1.1, one submits to the operating system the single statement\footnote{Additional information about \texttt{dvips} can be found by typing the command \texttt{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 \texttt{dvips}, accessed in many systems by typing the statement \texttt{texdoc dvips} at a \textit{Shell} prompt or, in the UNIX and LINUX operating systems, the statement \texttt{man dvips} at a \textit{Shell} prompt.}

\begin{verbatim}
pdflatex filename
\end{verbatim}

or, for the sample above,

\begin{verbatim}
pdflatex texsample1
\end{verbatim}

\footnote{Check the \textit{Local Guide} for more details. In particular, some sites may have implemented the shorthand command \texttt{dvip filename} for the command \texttt{dvips -o filename.ps filename}.}

The program \texttt{pdflatex} is normally installed automatically whenever \LaTeX\ is installed.
which will produce directly a PDF file named \texttt{filename.pdf} or \texttt{texsample1.pdf} as well as two additional files, specifically the \texttt{.aux} (which we can ignore for now) and the \texttt{.log} file containing an expanded log of the messages that appear on the screen as \texttt{latex} works its magic.

2. Starting with the \texttt{.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 \texttt{filename.pdf} or \texttt{texsample1.pdf} from the \texttt{.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 \texttt{.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 \TeX on the screen exist. Among the more common are the following:

- (for displaying a \texttt{.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.¹⁷

### A.1.5 Printing the Document

Programs (\texttt{xdvi}, \texttt{yap}, \texttt{ghostview}, \texttt{acroread}) for displaying a file on the screen often have an item in the \textit{File} menu to request printing of the file on an available printer, and many times the print utility thereby invoked offers options for double-sided printing, scaling of the output, etc. Often, a simple command like

\begin{align*}
\text{UNIX/LINUX/MAC} & \quad \text{Windows} \\
\texttt{lp filename.ps} & \quad \texttt{print filename.ps}
\end{align*}

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.¹⁸

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.¹⁹ In any case, the resulting output for our sample file is shown in Table A.4.

---

¹⁷See the \textit{Local Guide} for specifics at your site.

¹⁸Printers without this capability are rare.

¹⁹Again, check the \textit{Local Guide} for details on how to print a file.
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{LengthParameter}{Value}

to set the length parameter identified by \textit{LengthParameter} to the value specified by \textit{Value}. A length can be a positive or negative value given in points\textsuperscript{20} (abbreviated pt), centimeters (abbreviated cm), or inches (abbreviated in or, when we wish to make the dimension immune to global changes of scale, \texttt{truein}). Further, to give \LaTeX{} some flexibility in placing text on the page, rubber values can be specified. For example, a value of \texttt{9pt plus 1pt minus 2pt} gives \LaTeX{} authority to “cheat” on the value over the range from 7 points to 10 points. The most useful length parameters are identified in Table A.5. Thus, the commands

\begin{verbatim}
\setlength{\textheight}{9truein}
\setlength{\textwidth}{6truein}
\setlength{\oddsidemargin}{0.25truein}
\setlength{\topmargin}{-0.5truein}
\setlength{\parindent}{20pt}
\setlength{\parskip}{6pt plus 2pt minus 1pt}
\end{verbatim}

placed in the preamble will change the defaults to specify a 6” \times 9” area of text centered on an 8.5” \times 11” page with a 20 point paragraph indent and an extra 5–8 points between paragraphs. Indeed, these are (almost\textsuperscript{21}) the settings used for this book.

A few issues of global style are specified by an optional argument in the command \texttt{\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 command \texttt{\documentclass} to\textsuperscript{22}

\begin{verbatim}
\documentclass[TypeSize]{Class}
\end{verbatim}

where, in the \texttt{article}, \texttt{book}, and \texttt{report} classes, \textit{TypeSize} can be one of \texttt{10pt} (the default), \texttt{11pt}, or \texttt{12pt}. The \texttt{slides} class \textit{admits} these specifications of size, but \textit{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 \texttt{\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 \textit{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 \textit{templates} are enumerated in Table A.6. Consult the

\begin{footnotesize}
\begin{itemize}
\item\textsuperscript{20}The \textit{point}, which is a standard printer’s measure of length, is $1/72\"$ (72 points per inch).
\item\textsuperscript{21}The difference lies in the specification of the left margins. For this book (which uses double-sided printing), \LaTeX{}’s ability to specify one left margin (with \texttt{oddsidemargin}) for odd-numbered pages and a different left margin (with \texttt{evensidemargin}) for even-numbered pages has been exploited to keep the text more fully out of the binding than would otherwise be the case.
\item\textsuperscript{22}By convention in \LaTeX{}, \textit{optional} arguments to a command are enclosed in \textit{square brackets} and \textit{mandatory} arguments are enclosed in \textit{curly braces}.
\end{itemize}
\end{footnotesize}
Table A.5: Selected parameters that affect the size and positioning of the text area on a page. The
given default values apply to the \texttt{article} style. Further information can be found in Section 6.4.1,
at the very end of Section C.5.3, and in Fig. C.3 in \textit{The \LaTeX\ Manual}.

\begin{itemize}
\item \texttt{textheight} (default 7.375") Specifies the vertical dimension of the region of the page occupied by text, excluding the head and the foot of the page. Its value is commonly 9" for an 8.5"\times11" page.
\item \texttt{textwidth} (default 4.75") Specifies the horizontal dimension of the region of the page occupied by text. Its value is commonly 6" for an 8.5"\times11" page.
\item \texttt{oddsidemargin} (default 0.75") For all pages (single-sided output) or odd-numbered pages (double-sided output), specifies the left margin—distance from the left edge of the page to the left edge of the region occupied by text—to be 1" plus the specified value. The value 0.25" will center 6" wide text on an 8.5"\times11" page.
\item \texttt{evensidemargin} (default 0.75") Only for even-numbered pages with double-sided output, specifies the left margin—distance from the left edge of the page to the left edge of the region occupied by text—to be 1" plus the specified value. The value 0.25" will center 6" wide text on an 8.5"\times11" page.
\item \texttt{topmargin} (default 0.25") Specifies the top margin, i.e., the distance from the top edge of the page to the top edge of the header above the main region occupied by text, to be 1.0" plus the specified value. By default, the height of the header (specified by \texttt{headheight}) and the separation of the header from the main text (specified by \texttt{headsep}) together add to 0.5", so—if the defaults for \texttt{headheight} and \texttt{headsep} are accepted—we can pretend that \texttt{topmargin} specifies the distance from the top edge of the page to the top edge of "real" text to be 1.5" plus the specified value. Thus, for example, the value $-0.5$" will center 9" high text on an 8.5"\times11" page and place the header line—if any—0.5" above the first line of text.
\item \texttt{parindent} (default 15.0 points) Specifies the paragraph indent.
\item \texttt{parskip} (default 0.0 points) Specifies the \emph{extra} space between paragraphs.
\end{itemize}

Table A.6: Templates in the directory \$\texttt{HEAD/tex}$.

<table>
<thead>
<tr>
<th>Template</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{lu_article.template}</td>
<td>Specifies a general article style, including a 6&quot;\times9&quot; text area centered on an 8.5&quot;\times11&quot; page.</td>
</tr>
<tr>
<td>\texttt{lu_cpl.template}</td>
<td>Specifies the style of CPL publications.</td>
</tr>
<tr>
<td>\texttt{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>\texttt{lu_letter.template}</td>
<td>Supplies a starting format for business letters.</td>
</tr>
<tr>
<td>\texttt{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 (\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>{\rm text}</td>
<td>Sets text in medium, Roman, upright type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\TeX</td>
<td>{\sl text}</td>
<td>Sets text in medium, Roman, slanted type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\TeX</td>
<td>{\it text}</td>
<td>Sets text in medium, Roman, italic type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\TeX</td>
<td>{\tt text}</td>
<td>Sets text in medium, typewriter, upright type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\TeX</td>
<td>{\bf text}</td>
<td>Sets text in bold, Roman, upright type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\TeX</td>
<td>{\sc text}</td>
<td>Sets text in medium, Roman, small caps type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\LATEX</td>
<td>{\em text}</td>
<td>Sets text in emphasized type.</td>
<td>(See text.)</td>
</tr>
<tr>
<td>\LATEX</td>
<td>{\bf text}</td>
<td>Sets text in emphasized 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 cap) of the desired style. Each characteristic is specified independently of the other two. Repeated selection of all three characteristics, however, is cumbersome, so—as enumerated in Table A.7—\LaTeX provides simpler ways to select the more common combinations.

Some of these mechanisms make use of *declarations*, which change the type style until it is explicitly changed again. If, for example, we included the declaration \bf at some point in our code, everything in our document from that point until we change the style with another declaration would be set in bold, Roman, upright type. To limit the scope of this declaration to some portion of the text, the text to be “emboldened”, *including the declaration*, would be enclosed in curly braces (as shown in the table). Once \LaTeX’s processing has passed beyond the point of the closing curly brace, the declaration is no longer in effect and the type style reverts to what it was before the opening curly brace.

Emphasis of a word or a phrase now and then would, in normal use, be achieved with *italic* type. In \LaTeX, the declaration \it might be used. The declaration \em, however, is preferred because it is sensitive to the current type style. When the current type style is Roman, \em will shift to italic type; when the current type style is italic, \em will shift to Roman type. Thus, \em can be used inside a phrase that is already being emphasized. Emphasized text inside of emphasized text that is itself embedded in Roman text will be set in Roman text to contrast with the italic style of the text that immediately surrounds it.

---

23The default is medium series, Roman, upright shape—the style type 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 declaration \em but also with the command \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 \TeX 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, \textbackslash and\textbackslash. (The italic correction is not necessary in all situations, e.g., if the following character is a period or a comma, which explains why it cannot be automatically and always inserted.)

In using one or another of these simple specifications, we are accepting particular combinations of series, family, and shape. \TeX provides two ways to take advantage of the full flexibility with separate specification of these three characteristics. For example, we might use declarations to specify bold, sans serif, upright type with the construction

\begin{verbatim}
{\bfseries\sffamily\upshape text}
\end{verbatim}

Alternatively, we might use the commands

\begin{verbatim}
\textbf{ \textsf{ \textup{ text } } }
\end{verbatim}

to achieve the same end with a shorter phrase. The full set of available declarations and commands includes

\begin{verbatim}
\mdseries, \bfseries, \textmd{...}, \textbf{...}
\end{verbatim}

to specify medium and bold series, respectively,

\begin{verbatim}
\rmfamily, \sffamily, \ttfamily, \textrm{...}, \textsf{...}, \texttt{...}
\end{verbatim}

to specify Roman, sans serif, and typewriter families, respectively, and

\begin{verbatim}
\upshape, \itshape, \slshape, \scshape ,
\textup{...}, \textit{...}, \textsl{...}, \textsc{...}
\end{verbatim}

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.
A.3. IN-TEXT SPECIFICATION OF LOCAL STYLE

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>\text{\tiny\text}</td>
<td>Sets text in tiny type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\text{\scriptsize\text}</td>
<td>Sets text in scriptsize type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\text{\footnotesize\text}</td>
<td>Sets text in footnotesize type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\text{\small\text}</td>
<td>Sets text in small type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\text{\normalsize\text}</td>
<td>Sets text in normalsize type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\text{\large\text}</td>
<td>Sets text in Large type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\text{\Large\text}</td>
<td>Sets text in LARGE type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\text{\huge\text}</td>
<td>Sets text in huge type.</td>
<td>Sample</td>
</tr>
<tr>
<td>\text{\Huge\text}</td>
<td>Sets text in Huge type.</td>
<td>Sample</td>
</tr>
</tbody>
</table>

A.3.3 White Space

Some aspects of the way the components (text, tables, graphs, ...) of a document are placed in the available space are controlled by the length parameters set in the preamble and discussed in Section A.2. Additional features of \LaTeX\ that give control over this feature of a document include:

- The command \text\{noindent\} to suppress paragraph indentation for a single paragraph.
- The commands \text\{\newpage\} and \text\{\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 \text\{\baselineskip\} to specify the linespacing. This parameter is set with the command \text\{\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 \text\{\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 \text\{\quad\} and \text\{\qquad\}, which insert an en space and an em space, respectively. (An en space is about the width of the letter x, and an em space is about the width of the letter M, both in the current font.)
- The command \text\{\vspace\{Value\}\} to generate vertical space. Here, Value can be any defined length parameter, e.g., \text\{\vspace\{\parskip\}\}, or a specific length, e.g., \text\{\vspace\{2.0truein\}\}. If the space happens to occur at the top of a new page, it will be omitted, but the command \text\{\vspace*\{Value\}\} will force the space to be included even if it is at the top of a page.
- The command \text\{\hspace\{Value\}\} to generate horizontal space. Here, Value can be any defined length parameter, e.g., \text\{\hspace\{\parindent\}\}, or a specific length, e.g., \text\{\hspace\{36.0pt\}\}. If the space happens to occur at the end of a line, it will be omitted, but the command \text\{\hspace*\{Value\}\} will force the space to be included even if it is at the end of a line.
- The commands \text\{\settowidth\}, \text\{\settoheight\}, and \text\{\settodepth\}, which return lengths for the corresponding dimensions of the box that will enclose whatever is the argument of the
command. No printed output is produced by these commands. We might, for example, define a new length parameter \texttt{dblast} as the width of the box accommodating two asterisks with the two commands

\newlength{dblast}
\settowidth{dblast}{**}

and then insert the space for two asterisks (without actually displaying them) with the command

\hspace{dblast}

These commands are described in detail at the end of Sections 6.4.1 and C.13.1 in The \textit{L}a\textsc{tex} \textit{Manual}.

\section*{A.3.4 Environments}

Sometimes, the formatting to be accomplished requires a more sophisticated definition of a change than is possible within the framework of a simple declaration or command with a few arguments. \LaTeX{}'s \textit{environments} are more versatile and flexible than declarations and commands. For example, to present text in a narrower paragraph, as is conventionally done with extended quotations, we might invoke one of the constructions

\begin{quote}
Text of quotation.
\end{quote}
or
\begin{quotation}
Text of quotation.
\end{quotation}

Normally, the command \texttt{\begin{...}} will be placed on a line by itself, the text may spread over several lines, and the command \texttt{\end{...}} will be placed on a line by itself. Left and right margins are indented equally. Unless the text includes its own formatting specifications, however, the first construction will indent each new paragraph and will place 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{center}
First line \ \\
Second line \ \\
Third line \ \\
\ldots
\end{center}

Normally, the command \texttt{\begin{...}} will be placed on a line by itself, the text may spread over several lines, each line in the final output will in the code be separated from its predecessor with the command \texttt{\\}, and the command \texttt{\end{...}} will be placed on a line by itself.

These two examples introduce the construction called an \textit{environment}, which in general has the format

\begin{EnvironmentName} \ldots \end{EnvironmentName}

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 \textit{The}
\texttt{\LaTeX} \textit{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!}

\subsection{\texttt{\LaTeX} Packages}

As a publishing system, \texttt{\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{graphicx}, \texttt{graphpap}, \texttt{ifthen}, \texttt{latexsym}, \texttt{imakeidx}, \texttt{verbatim}, \texttt{pict2e}, and \texttt{showidx}. A brief description of many of these packages is included in Section C.5.2 of \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 \texttt{\LaTeX} source files. Usually, those programs make use of their own special \texttt{\LaTeX} commands and supply program-specific \texttt{\LaTeX} packages which must be accessible, either at the proper point in the \texttt{\LaTeX} directory structure or in the directory from which you run \texttt{\LaTeX} when processing a \texttt{\LaTeX} file produced by the programs. Details on those additional packages will be found in the manuals provided by the vendors of the programs.

\subsection{Miscellaneous Other Capabilities}

Several additional capabilities merit particular comment:

\begin{itemize}
  \item To force \texttt{\LaTeX} to keep words together on a single line and/or to prevent \texttt{\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 not end a sentence. For example, we should type \texttt{Mr.~Cook} rather than \texttt{Mr.\ 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 \texttt{\LaTeX} automatically inserts extra space after each period (which is assumed to mark the end of a sentence). \texttt{\LaTeX} may add even more space in the second case as it justifies the line between the margins.

  \item To generate a dash, use \texttt{--}, \texttt{---}, or \texttt{\ldots}, depending on the length of the required dash. One hyphen produces a hyphen, as in two-column; two hyphens in a row produce a slightly longer
\end{itemize}

\footnote{The symbol \texttt{\ldots} 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 \LaTeX{} is permitted to insert a hyphen as it fills and justifies lines. \LaTeX{} has its own rules for hyphenation, but they are not infallible. Sometimes \LaTeX{} needs help. Note that words containing even one indication of an optional hyphen will not be hyphenated at any other point(s) in the word.

- To suppress hyphenation of a single word altogether, place the word as the argument of an \mbox command, e.g., \mbox{customize}.

### A.4 Including Equations

Most scientific documents will have equations, all of which must be specified in \LaTeX{}’s math mode. Shifting from the default text mode to math mode can be accomplished in several ways. For short, unnumbered equations or mathematical symbols that appear in a line of text, the equation or symbol must be enclosed in one of the two constructions \(...\) 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 \texttt{displaymath} environment centers a single equation on a line by itself, does not number the equation, and can be specified by the shortened form \(...\). The \texttt{equation} environment centers a single equation on a line by itself and numbers the equation automatically. The \texttt{eqnarray} environment—see The \texttt{L\LaTeX} Manual—is used for a string of equations and for long equations that will not fit on one line; its syntax allows vertical alignment of two or more displayed and numbered equations. To facilitate references to numbered equations produced in the \texttt{equation} and \texttt{eqnarray} environments, \LaTeX{} provides the command \texttt{\label{ReferenceName}} for defining a symbolic label \texttt{within} the environment and the command \texttt{\ref{ReferenceName}} to permit referencing the equation by its number in the text. Further, the command \texttt{\pageref{ReferenceName}} permits referencing the page on which the equation occurs in the text.\footnote{When these references are used, the information written into the .aux file becomes important and the code must be processed by \texttt{latex} or \texttt{pdflatex} twice, once to write the correct information into that file and a second time to read and make use of that information. Other environments, such as \texttt{table} and \texttt{figure} also admit a \texttt{\label} command and can make use of this capability for symbolic internal references to those components. The label command can also be used inside the argument of \texttt{part}, \texttt{chapter}, \texttt{section}, \texttt{subsection}, \texttt{subsubsection}, and \texttt{\footnote} and in the text following an \texttt{item} command in an \texttt{enumerate} environment to facilitate reference to these components of the document.} (Warning: The \texttt{displaymath}, \texttt{equation}, and \texttt{eqnarray} environments will generate error messages if they contain blank lines.)
\LaTeX{} also has many symbols and structures, common in mathematical formulas, that can be used only in math mode. Subscripts and superscripts, roots, Greek letters, integral signs, summation signs, the `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 \LaTeX{} is in math mode. So, to insert the symbol $\Theta$ in the final output, we place the command $\text{	extbackslashTheta}$ in the code.\footnote{In math mode, Greek letters can be produced by commands that simply name the letter, e.g., $\text{	extbackslashTheta}$ for $\Theta$ and $\text{	extbackslashomega}$ 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., $\text{	extbackslashkappa}$, do not exist.} (Here, the first dollar sign puts \LaTeX{} in undisplayed math mode and the second returns \LaTeX{} to text mode.) Since \LaTeX{} follows its own rules about spacing when in math mode, spaces in math mode are usually ignored. (A space, however, is necessary in such contexts as $\text{	extbackslashomega} \text{t}$ to separate the command $\text{	extbackslashomega}$ from the character $t$.) The commands $\text{	extbackslash;}, \text{	extbackslash!}, \text{	extbackslash,},$ and $\text{	extbackslash;};$ 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., $(x + y) \text{ } dx \text{ } dy$ rather than $(x + y) dx \text{ } dy$.\footnote{Technically and officially, the differential element $dx$ should be written with a \textit{Roman} d and an \textit{italic} x, i.e., $dx$, but this convention is rarely followed.}

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 $\text{	extbackslashcos}$ and $\text{	extbackslashsin}$ should be used for these functions.\footnote{See Table 3.9 in \textit{The \LaTeX{} Manual} for a complete listing.}

Except for $\text{	extbackslashsc}$, 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 $\text{	extbackslashmathrm}, \text{	extbackslashmathit}, \text{	extbackslashmathbf}, \text{	extbackslashmathtt}$, 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 $\text{	extbackslashmathrm}$, for example, to include a \textit{two}-word phrase, will result in the words running together unless the space command $\text{	extbackslashAlt}$ 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 $\text{	extbackslashmbox}$ to escape temporarily to text mode.

Including bold characters in \textit{equations} is especially tricky. The command $\text{	extbackslashmathbf}$ “emboldens” only \textit{some} characters in the processed formula and, in particular, leaves lower case Greek symbols in light face. The $\text{	extbackslashboldmath}$ 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

$\text{	extbackslashmbox}{ \text{	extbackslashboldmath $\theta$} } \ldots \text{	extbackslashmbox}{ $\ldots$ }$

where the command $\text{	extbackslashmbox}$, which \textit{is} valid in math mode, temporarily shifts to text mode so the $\text{	extbackslashboldmath}$ declaration can be invoked and the math mode expression in the argument of the command $\text{	extbackslashmbox}$ will then be made bold; the bold presentation is confined by the command $\text{	extbackslashmbox}$ to the desired part of the total expression.

\midsection{Including Lists}

Among the most commonly used environments are those that facilitate creation of lists, including the $\text{	extbackslashitemize}$ environment for bulleted lists (such as the one in Section A.3.6) and the $\text{	extbackslashenumerate}$ 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 $\text{	extbackslashbegin} \ldots \text{	extbackslashend}$. Within the environment, each new item is introduced with the command $\text{	extbackslashitem}$. 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{enumerate}
\item Text of first item.
\item Text of second item.
\item Text of third item.
\end{enumerate}

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
Table A.9: Format for inserting a table.

\begin{table}
\caption{...}
\label{...}
\begin{center}
\begin{tabular}{clr}
Row 1, Col 1 & Row 1, Col 2 & Row 1, Col 3 \\
Row 2, Col 1 & Row 2, Col 2 & Row 2, Col 3 \\
Row 3, Col 1 & Row 3, Col 2 & Row 3, Col 3
\end{tabular}
\end{center}
\end{table}

be invoked anywhere in the code and the resulting table will be placed at the point at which the environment appears, perhaps even in the middle of a line of text. Conventionally, however, tables are placed at the top or bottom of the page,\textsuperscript{31} and \LaTeX{} provides the \texttt{table} environment to facilitate proper placement of a table,\textsuperscript{32} though the bare \texttt{table} environment contains nothing other than its name that suggests its use for tables. Any text at all can be incorporated in the \texttt{table} environment and will be placed on the page as would a table. Most commonly, however, the \texttt{table} environment will embrace a \texttt{tabular} environment defining the table itself and will use the command \texttt{\caption} to specify the caption of the table and the command \texttt{\label} to specify a symbolic label for use in referring to the table within the document.\textsuperscript{33} Frequently, the \texttt{tabular} environment will itself be embraced in a \texttt{center} environment to control the horizontal placement of the table on the page. Thus, the common structure for the code defining a table would be of the form shown in Table A.9.\textsuperscript{34,35}

Here, individual elements in a row are separated from one another by ampersands, and the end of each row—except the last—is marked with the command \texttt{\}. In this example, the caption, including an automatically generated phrase and number of the form ‘Table 3:’, will appear above the table. Positioning the commands \texttt{\caption} and \texttt{\label} after the \texttt{center} environment would place the caption below the table. Whichever position is adopted, the command \texttt{\label} must appear after the command \texttt{\caption}. Wherever the command \texttt{\label} is positioned, it will place an entry in the .aux file so that the commands \texttt{\ref} and \texttt{\pageref} will function here as described for equations at the beginning of Section A.4.

The illustrative argument \texttt{clr} of the \texttt{tabular} environment requires a bit more explanation.

\textsuperscript{31}Actually, the command \texttt{\begin{table}} has an optional argument (\texttt{\begin{table}[OptArg]}), which can assume any of the values h, b, or t for placement of the table at the place where the \texttt{table} environment appears (h, for here) or at the bottom (b) or top (t) of the page. In the absence of an explicit specification, \LaTeX{} makes its own decision about placement.

\textsuperscript{32}If there is space on the current page when the \texttt{table} environment is encountered, the table will be placed on that page. Otherwise, the table will be placed on a subsequent page.

\textsuperscript{33}The \texttt{\caption} and \texttt{\label} commands can be placed anywhere within the \texttt{table} environment. Captions will commonly be placed either above or below the captioned component. In some contexts (see the warning at the end of Section A.10), placing the caption above the item is preferable.

\textsuperscript{34}\LaTeX{} automatically uses the specified caption as an entry for a list of tables. There is, however, a limit to the length of such entries. Especially long captions may generate error messages. To avoid this problem, we can—and sometimes must—exploit an optional argument to the command \texttt{\caption\{\ldots\}} 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\{}\{\ldots\}}\texttt{\}, though we shall here not follow that recommendation. See The \LaTeX{} Manual for details.

\textsuperscript{35}See Sections 3.6 and C.10 in 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{??..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 \texttt{\label} must appear \textit{after} the command \texttt{\caption}. For this method, you need only have printed copies of your figures.\textsuperscript{38} 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.

\subsection{Using the graphicx Package}

Incorporation of graphic images in a \LaTeX{} document is viewed by \LaTeX{} as the responsibility of the device driver. To enable that process, \LaTeX{} allows the passing of commands to the device driver. Both the format of such commands and the variety of capabilities thus available depends entirely on the device driver. In essence, this process uses the \TeX{} (not \LaTeX{}) command \texttt{\special}, but the user is unaware of that underlying command because it is invoked behind the scenes.

The simplest means for importing graphical displays into a \LaTeX{} document—i.e., of commanding \LaTeX{} to pass the proper information on to the device driver—exploits the \texttt{graphicx} package,\textsuperscript{39} whose features are made available by including that package with the statement

\begin{verbatim}
\usepackage{graphicx}
\end{verbatim}

in the preamble of our \LaTeX{} code.\textsuperscript{40}

After placing the command \texttt{\usepackage{graphicx}} in our preamble, we incorporate the graphical display itself by placing a statement like

\begin{verbatim}
\includegraphics[height=\texttt{dimension}]{filename}
\end{verbatim}

at the point where the PostScript or PDF figure defined in by the file \texttt{filename} is to be inserted.\textsuperscript{41} In general, the image would be centered horizontally on the page by inserting this command in a \texttt{center} environment, e.g.,

\begin{verbatim}
\begin{center}
\includegraphics[height=\texttt{dimension}]{filename}
\end{center}
\end{verbatim}

For example, the code\textsuperscript{42}

\footnote{Appropriately translated, the substance of footnote 34 applies here to the entry placed automatically in the list of figures.}

\footnote{The \texttt{graphics} package might also be used, but the syntax of the commands it defines is marginally less convenient than the syntax of the parallel commands in the \texttt{graphicx} package, \textit{i.e.} the \texttt{extended graphics} package.}

\footnote{More generally, this statement admits an optional argument to the desired graphics driver explicitly. Frequently, that driver will be \texttt{dvips} and the statement would be \texttt{\usepackage[dvips]{graphicx}}, but numerous other drivers exist, and many may be installed at your site. See the \textit{Local Guide} for more information. For ultimate flexibility in the formatting of graphics files, we elect to leave that choice to the default.}

\footnote{If you use \texttt{latex}, \texttt{dvips}, and \texttt{ps2pdf} as described in Section A.1.2 and item 2 in Section A.1.3 to produce PostScript and PDF documents, your figures must be available as \texttt{.ps} or \texttt{.eps} files, and—even if the file type is omitted in the \texttt{\includegraphics} command—the processing will correctly identify and incorporate the figures. If you seek to produce a PDF document directly with \texttt{pdflatex} as described in item 1 in Section A.1.3, then—even if the file type \texttt{pdf} is omitted in the\texttt{\includegraphics} command—the processing will correctly identify and incorporate the figures. If your figures are PostScript, you must use the first route to the final document; if your figures are PDF, you must use the second route. Software to convert PostScript figures to PDF figures and \textit{vice versa} is described in Section A.11. If both PostScript and PDF files exist for all figures \textit{and} you omit the file type \textit{in} the \texttt{\includegraphics} \texttt{command}, then you can use either \texttt{latex} or \texttt{pdflatex} for the processing.}

\footnote{Appropriately translated, the substance of footnote 34 applies here to the entry in the list of figures.}
Figure A.1: Mesh surface representation of the irradiance produced by a square aperture.

\begin{figure}
\caption{Mesh surface representation of the irradiance produced by a square aperture.}
\label{LATEX:irrad}
\begin{center}
\includegraphics[height=3.0truein]{diffract}
\end{center}
\end{figure}

will produce Fig. A.1 if the file diffract is stored in the current default directory. Here, the specification of the parameter \texttt{height} in the optional argument of the command \texttt{\includegraphics} causes the figure to be scaled to the specified dimension vertically and scaled by the same fraction horizontally to preserve its aspect ratio. As in the example in Section A.7.1, the figure here has been inserted before the command \texttt{\caption}, so that the caption will appear above the figure. Further, the command \texttt{\label}, which must follow the command \texttt{\caption}, defines a symbolic label that can be used within the document to refer to the figure.

\textit{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 \texttt{.dvi} file but those produced by the \texttt{qt} toolkit \textit{may} not then translate with \texttt{dvips} to the subsequent PostScript file. When the latter is the objective, use \texttt{gnuplot} (or maybe \texttt{fltk}). A bit of testing may be necessary.

The command \texttt{\includegraphics} admits several optional arguments. The argument \texttt{width} can also be included. If \texttt{width} is specified \textit{instead of} \texttt{height}, the height will be scaled to preserve...
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 \LaTeX\ Graphics Companion.

A.7.3 Using the \texttt{tikz} Package

The \texttt{tikz} package, which is routinely included in present-day \LaTeX\ distributions, provides an assortment of \LaTeX\ macros to facilitate the drawing of simple—and even complicated—figures. Basically, the statement \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{Note that, when displayed on the screen with \texttt{xdvi} or \texttt{yap}—see Section A.1.4—the intermediate .dvi file produced in Section A.1.2 will not contain the correct figures.} Be aware that the code inserted in the \texttt{tikzpicture} environment describes the picture in the same way that \LaTeX\ provides the text and its formatting. Just as one imagines the final formatted document as the \LaTeX\ coding is constructed, so one has to imagine the resulting picture as one constructs the descriptive code to create it in the final document. \texttt{Tikz} is not a WYSIWIG drawing program like \texttt{PAINT} or \texttt{TGIF}. One of its advantages over the route described in Section A.7.2, which imports figures produced in other programs, is that the route of this section guarantees that the fonts used in the figures will be identical to those used in the rest of the document.

The graphics system provided by \texttt{tikz} is impressively elaborate and versatile. As such, learning its capabilities, especially its more sophisticated capabilities, will require substantial effort. Typing the command\footnote{The designation \texttt{pgf}—Portable Graphics Format—reflects the name of the engine underlying the entire system.} \texttt{texdoc pgf}

at a \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 \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.
• 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.\footnote{Note that the default unit—here inches—for radius can be overridden with an explicit stipulation, e.g., [radius=0.1cm].} Available colors include red, green, blue, cyan, magenta, yellow, and many others. You can also define your own color using a command like\footnote{The command \definecolor here used is defined within the \tikzpicture environment and does not require use of the \LaTeX package \texttt{color}.}:

\begin{verbatim}
definemcolor{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.\footnote{See the manuals for the details of how to specify the dimensions of these shapes.}

• The statement \draw [\texttt{<--}, 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 \texttt{<--} places an arrow at both ends of the line,\footnote{The separate arguments \texttt{->} and \texttt{<-} will place an arrow at the end or the beginning of the line, respectively.} and the argument \texttt{line width} specifies the weight of the line, by default, in points.\footnote{The argument \texttt{line width=2} could be replaced, among others, by \texttt{ultra thin}, \texttt{very thin}, \texttt{thin}, \texttt{thick}, \texttt{very thick}, and \texttt{ultra thick}, and the default unit could be overridden with an argument like \texttt{line width=0.1cm}.}

• Text is placed where specified with the statements \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 $\pi$ with the simple coding $\texttt{pi}$.\footnote{Another way to draw smooth curves is to use a calculational aid in another program to generate a probably long list of coordinates for many points on the curve and then to insert that list as the argument of a \texttt{draw} command.}

• 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 \{\ldots\} 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 \texttt{draw} is a versatile command that admits numerous embellishments through the use of keywords and/or optional arguments. Conveniently, \texttt{tikz} determines the size of the space to be used in the document to reflect the size of the picture as constructed with the measures provided in the \texttt{tikz} coding.

When features of the package \texttt{tikz} are invoked in the \LaTeX code, producing the formatted document must be done carefully. If there are no complications (external figures, table of contents, internal references) to be incorporated, the simple statement

\begin{verbatim}
\foreach \x in {0.0,1.0,2.0} {
  \draw (\x,0.1)--(\x,-0.1);
}
\end{verbatim}

\end{verbatim}
Figure A.2: An illustrative \texttt{tikz} figure. Note that specified colors will be converted to a gray scale unless the display device—screen or printer—can display colors.

![Tikz figure with x and y axes](image)

Table A.10: Coding to produce Fig. A.2.

\begin{verbatim}
\begin{figure}
\caption{An illustrative \texttt{tikz} figure.}
\label{LATEX:tigzfig}
\begin{center}
\begin{tikzpicture} \[x=1.0in,y=1.0in\]
\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}
\end{verbatim}

\texttt{pdflatex \textit{filename}} (Default file type \texttt{.tex})

will produce a PDF file and the statements

\texttt{latex \textit{filename}} (Default file type \texttt{.tex})
\texttt{dvips -o \textit{filename}.ps -t letter \textit{filename}} (Default file type \texttt{.dvi})
\texttt{ps2pdf \textit{filename}.ps}

will produce a PostScript and then a PDF file, though figures defined by \texttt{tikz} will not be properly
rendered in the intermediate .dvi file. If a table of contents or internal references or both are involved, two—and maybe three—passes through \texttt{latex} or \texttt{pdflatex} will be necessary. If there are PostScript figures and \textit{no} PDF figures to be incorporated, only \texttt{latex} will work; if there are PDF figures (and \textit{no} PostScript figures) or hyperlinks to be incorporated, only \texttt{pdflatex} will work. Finally, if there are both PostScript and PDF files defining figures, files of one type will have to be converted to the other type before either of these routes to a finished document will work.\footnote{See Section 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.

\section*{A.7.4 Using the \texttt{picture} Environment and \texttt{pict2e} Package}

Finally, we merely mention the \texttt{picture} environment, described in Sections 7.1 and C.14.1 of \textit{The \LaTeX\ Manual}, and the supplementary \texttt{pict2e} package, described in \textit{The \LaTeX\ Companion}. These components add several commands that facilitate the detailed construction of at least simple graphical displays.

\section*{A.8 Including a Table of Contents, a List of Figures, and a List of Tables}

A table of contents, a list of figures, and a list of tables can be included in the formatted document by the commands \texttt{\tableofcontents}, \texttt{\listoffigures}, and \texttt{\listoftables}. Each

- results in the output of an auxiliary ASCII file of information with file type \texttt{.toc}, \texttt{.lof}, and \texttt{.lot}, respectively,
- requires a second pass through \texttt{latex} or \texttt{pdflatex} to incorporate the list in the finished document, and
- places the list in the output at the point at which the command appears.

When these commands are invoked, each of the components of the text defined by the commands \texttt{\part}, \texttt{\chapter}, \texttt{\section}, \texttt{\subsection}, and \texttt{\subsubsection} will automatically generate a line in the \texttt{.toc} file.\footnote{Note that not all of these options are available in all document classes. For example, \texttt{\part} and \texttt{\chapter} are available only in the \texttt{book} class. Note also the counters \texttt{secnumdepth} and \texttt{tocdepth} that control the depth to which sections are numbered and the depth to which sections are catalogued in the table of contents. The default values of these counters can be overridden using \texttt{\setcounter} in the preamble.} each use of the \texttt{\caption} command in a \texttt{figure} or \texttt{table} environment will automatically generate a line in the \texttt{.lof} or \texttt{.lot} file. Then, in the second pass—which will always be necessary—through \texttt{latex} or \texttt{pdflatex}, these files are read and each line results in an entry in the corresponding list.

Two commands allow explicit entry of information into one or another of these files at the point where the command is inserted in the \LaTeX\ source file, specifically

- the command \texttt{\addcontentsline{file}{unit}{entry}}
- the command \texttt{\addtocontents{file}{text}}
A.9. INCLUDING AN INDEX

Here, *file* is one of *toc*, *lof*, and *lot*, *unit* is one of *part*, *chapter*, *section*, *subsection*, or *subsubsection* if *file* is *toc*, *figure* if *file* is *lof*, and *table* if *file* is *lot*.

These two commands have quite different effects. The first adds a *bona fide* entry to the corresponding file. For example, the line

\addcontentsline{toc}{subsection}{Specially marked point}

will add to the table of contents an entry that is left-justified at the subsection level, contains the text “Specially marked point”, and includes a row of dots and the appropriate page number.\(^{56}\) The second merely adds text—no dots or page number—to the corresponding file. For example, the line

\addtocontents{toc}{\vspace{24pt}This is a note.\vspace{24pt}}

inserts a note preceded and followed by a bit of extra vertical space. Any textual note is left-justified as dictated by the section level at which it appears in the document.

Additional details about the issues discussed in this section are laid out in Sections 4.1 and C.4 of *The \LaTeX Manual*. Note, in particular, the optional argument in the command \caption and the several sectioning commands (\chapter, \section, ...), which provide control over the entry each command places in the table of contents, list of figures, and list of tables.

## A.9 Including an Index

Among many capabilities, \LaTeX is able to generate an index, though not automatically. In essence, we must place the commands\(^ {57}\)

\usepackage{imakeidx}
\makeindex
\printindex

in the preamble and the command

\index{Item to be indexed}

at appropriate points throughout the document. (The detailed structure of the argument to the command \index is described in Section 4.5 and Appendix A of *The \LaTeX Manual*.) When the file *filename.tex* is processed by latex or pdflatex, the presence of these commands results in the addition of three auxiliary ASCII files named

- *filename.idx* containing one line for each index entry,
- *filename.ilg* containing a log of messages created when, behind the scenes, the idx file is automatically processed through makeindex, and

\(^{56}\)If you are creating a linked document, this entry in the table of contents will also appear in the navigation panel of that document.

\(^{57}\)Earlier versions of \LaTeX used the package makeidx, which is still available. We here recommend using imakeidx because it makes available a few more features than are available in the previous package and it eliminates the need to run an auxiliary program to format the index.
APPENDIX A. INTRODUCTION TO \LaTeX

- filename.ind produced by makeindex and containing the formatted index that is then read by \texttt{\textbackslash printindex} to create the actual index in the document.

With \texttt{imakeidx}, the index is automatically created and incorporated in the document because an execution of \texttt{makeindex} is inserted automatically at the appropriate point when \texttt{latex} or \texttt{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 too quickly to be comprehended).\footnote{With \texttt{makeidx}, a separate explicit processing of the .idx file with \texttt{makeindex} to produce the .ind file and a further pass through \texttt{latex} or \texttt{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 \texttt{\makeindex} command. For example, the command

\begin{verbatim}
\makeindex [options=-s ind_style, intoc]
\end{verbatim}

stipulates that the style defined by the file \texttt{ind_style.ist} should be invoked and the index should be provided—argument \texttt{intoc}—with an entry in the table of contents. The construction of the style file is fully described in Section 12.4 in \textit{The \LaTeX Companion} as identified in Section A.18.

To assist in constructing the index, an auxiliary package called showidx, which works both with \texttt{latex} and with \texttt{pdflatex}, can be invoked to print the index entries on each page as marginal notes. This package is invoked simply by including the command

\begin{verbatim}
\usepackage{showidx}
\end{verbatim}

in the preamble, though it appears as if this command must be inserted \textit{after} the \texttt{\makeindex} command described above. Unfortunately, when printing in a \textit{6”\times9”} area on \textit{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 \texttt{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 \texttt{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}.}
• (after first pass through \texttt{pdflatex}) \texttt{.aux}, \texttt{.idx}, \texttt{.ilg}, \texttt{.ind}, \texttt{.log}, \texttt{.out}, \texttt{.pdf}, and \texttt{.toc}. The \texttt{.idx}, \texttt{.ilg}, and \texttt{.ind} files will be present only if you are creating an index with \texttt{imakeidx}. The \texttt{.out} file contains information about hyperlinks; the rest are as described in item 1 of Section A.1.3, in Section A.8, and in Section A.9. At this point, the PDF file does not have the navigation panel or the table of contents but it does have the index. All of these files are ASCII text files, though the PDF file may contain some non-printing characters. As such they can all be displayed in an available text editor and understood.

• (after the second pass and, if necessary, the third pass through \texttt{pdflatex}) the same as in the previous bullet, though many of those files have been updated. At this point, the navigation panel, the internal hyperlinks, and the index have all been created.

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 \textit{and} the files in all \texttt{\includegraphics} commands are presented \textit{without} file type, then the file processed with \texttt{latex} will look for \texttt{.ps} and \texttt{.eps} graphics files and the file processed with \texttt{pdflatex} will look for \texttt{.pdf} files. All files to be sought must, of course, exist.\footnote{See Section A.11 for a means to convert \texttt{.ps} and \texttt{.eps} files to \texttt{.pdf} and vice versa.} This feature makes it easy to process the \textit{same} source file both with \texttt{latex} and with \texttt{pdflatex}.

\section*{A.11 Converting .eps and .ps Files to .pdf}

When files describing figures are created, it is often easier to output those files from the creating program as \texttt{.eps} or \texttt{.ps} files rather than as \texttt{.pdf} files. Unfortunately, if you desire to produce the final output with \texttt{pdflatex}, files describing pictures need to be \texttt{.pdf}. PostScript files must therefore be converted into PDF before running \texttt{pdflatex}. Short of asking someone who already knows how to achieve that transfer, rummaging on the web for guidance may be a frustrating experience. Once you find the right tools, however, the process is quite simple. Basically, to convert the files \texttt{figure.eps} and \texttt{figure.ps} to a useful PDF file involves the two steps

\begin{enumerate}
  \item \texttt{ps2pdf figure.eps} \quad (which will yield \texttt{figure.eps.pdf})
  \texttt{ps2pdf figure.ps} \quad (which will yield \texttt{figure.pdf})
  \item \texttt{pdfcrop figure.eps.pdf} \quad (which will yield \texttt{figure.eps-crop.pdf})
  \texttt{pdfcrop figure.pdf} \quad (which will yield \texttt{figure-crop.pdf})
\end{enumerate}

Here, step 1 creates the \texttt{.pdf} file and step 2 removes extraneous white space around the perimeter of the first-created \texttt{.pdf} file. Second arguments to both \texttt{ps2pdf} and \texttt{pdfcrop}, as in

\texttt{ps2pdf figure.eps figure.pdf}
\texttt{pdfcrop figure.pdf figurecrop.pdf}

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

\texttt{pdfcrop --margins 10 figure.pdf figurecrop.pdf}
will provide a 10-pixel margin of white space around the image. See the help message for further
details.

To simplify the process, the MS DOS batch files listed in Section A.A.1.1, respectively, and
the UNIX Shell scripts listed in Section A.A.2.1, respectively, exploit the second argument for both
\texttt{ps2pdf} and \texttt{pdfcrop} to control the file names and, when executed with statements like

\begin{verbatim}
MS DOS               UNIX
ceps2pdf figure     ./ceps2pdf figure (for .eps file)
cps2pdf figure      ./cps2pdf figure (for .ps file)
\end{verbatim}

will (1) leave in the involved directory a cropped PDF file whose name \texttt{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 \textit{Command} window (Windows) or a \textit{Shell} window (UNIX), the steps

\begin{itemize}
\item 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 \textit{and} UNIX) into that directory.
These files can be created by direct typing or they can be copied from the directory \texttt{$\$HEAD/tex$}.
\item Copy all .ps or all .eps files to be converted into that temporary directory. \textit{This step guards
against disaster should the process to come be run in the initial directory and fail.}
\item Create a file containing a list of the names of the .eps or .ps files in the directory. The
statements
\begin{verbatim}
MS DOS               UNIX
dir /b *.eps > dir.txt ls -1 *.eps > dir.txt
dir /b *.ps > dir.txt  ls -1 *.ps > dir.txt
\end{verbatim}
utilizing the option /b in Windows and the option -1 (one, not el) in UNIX produce the desired
file (filename, including file type).
\item Strip the file type by running the python program \texttt{ExtractFileName.py} (Section A.A.3.1). Here the same code works in both Windows and UNIX. This action will create the file \texttt{nameonly.txt} containing only the names of the files to be converted.
\item Effect the conversions by running the script \texttt{rdfile} with statements like
\begin{verbatim}
MS DOS               UNIX
rdfileeps (for eps files)          ./rdfileeps (for eps files)
rdfileps (for ps files)            ./rdfileps (for ps files)
\end{verbatim}
This action will convert each file in turn from .eps or .ps to .pdf and leave no intermediate
files in the directory.
\item Delete the *.eps or *.ps files in the temporary directory.
\item Move the *.pdf files to the directory from which the .eps and .ps files were moved temporarily,
leaving both the original .ps and .eps files intact but adding the .pdf files, so either \texttt{latex}
or \texttt{pdflatex} can then be used to create the final document.
\end{itemize}

will accomplish that conversion

The reverse process of converting a .pdf file to .ps or .eps format is easier. Specifically, the
statements
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 \texttt{\textbackslash and, \textbackslash or, and \textbackslash not.}
For example, if the Boolean flags NUMREC and FORTRAN are defined and set, then the composite
statement
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:

\begin{verbatim}
\usepackage{ifthen}
\newboolean{PRINT}
\typeout{} \typein\[trueorfalse\]{true (for print), false (for linked):}
\ifthenelse{ \equal{\trueorfalse}{true} \or \equal{\trueorfalse}{false}}{ \setboolean{PRINT}{\trueorfalse} }{\typeout{} \typeout{Must be either "true" or "false". Try again.} \end{document}}
\end{verbatim}

which define the boolean variable *PRINT*, ask for entry of either *true* or *false* at execution time, and terminate execution if an invalid value is entered. Then, the lines

\begin{verbatim}
\usepackage{color}
\ifthenelse{\boolean{PRINT}}{}{\definecolor{MyPurple}{rgb}{0.577,0.000,1.000} \usepackage[colorlinks=true, linkcolor={MyPurple}]{hyperref}}
\end{verbatim}

also in the preamble, will include the package *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

\begin{verbatim}
\IfFileExists{test.tex}{\input{test.tex}{}}
\end{verbatim}

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.,

\begin{verbatim}
\IfFileExists{test.tex}{\input{test.tex}{\typeout{File not found} \end{document}}
\end{verbatim}

\section*{A.13 Error Messages Generated by \LaTeX \textipa{X}}

However carefully the code is created, sooner or later \LaTeX \textipa{X} will generate an error message, which will have the general form

\begin{verbatim}
! error identification
1. number text read
\textbf{text not read}
\end{verbatim}

\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 \LaTeX to continue to read the file unless a fatal error has occurred (though \LaTeX 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 \LaTeX 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 \LaTeX Manual.

A.14 The Page Previewer for .dvi Files

Since \LaTeX 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 \LaTeX package. To use this previewer, we first create the .dvi file by running our code through \LaTeX as described in Section A.1. Then we enter a command like

\begin{verbatim}
xdvi filename or yap filename
\end{verbatim}

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 \LaTeX code, process it repeatedly with \LaTeX, 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, \LaTeX, screen previewer, printer, ...), one convenient strategy for using a text editor and a page previewer simultaneously involves invoking the previewer with a command like

\begin{verbatim}
xdvi filename
\end{verbatim}

The point at which \LaTeX encounters difficulty is conveyed by the downward displacement of the text after that point, though that point is not always where the offense actually occurs.

Yet Another Previewer

Additional information about xdvi on UNIX systems can be found in the on-line help, accessed in many systems by typing the command man xdvi.

If explicit specification of the input file is omitted, some implementations of xdvi will bring up a browser in which the desired .dvi file can be selected. Other versions may open the most recently displayed file, in which case the browser can be invoked by selecting ‘Open’ from the File menu.
Table A.11: Some of the features available through selection from menus (first column) in some versions of `xdvi`, through clicking ML on buttons in other versions of `xdvi`. The features actually available by these means may vary from version to version of `xdvi`.

<table>
<thead>
<tr>
<th>Menu</th>
<th>Button</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>File→Reload</td>
<td>Reread</td>
<td>Rereads current <code>.dvi</code> 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 <code>.dvi</code> file.</td>
</tr>
</tbody>
</table>

Table A.12: Keystrokes for controlling `xdvi`.

- `s` Size text to fit window.
- `ls` (one-s) Largest text.
- `ns` Apply shrink factor `n`.
- `g` Move to last page.
- `1g` (one-g) Move to first page.
- `ng` Move to (absolute) page `n`.
- `n` Move to next page.
- `p` Move to previous page.
- `nn` Advance `n` pages.
- `np` Back up `n` pages.
- `q` Exit program.

`xdvi filename &` (UNIX) or `yap filename &` (Windows)

which will start `xdvi` or `yap` but detach it from the launching Shell window, leaving the Shell window free for other uses. Then, with the `.tex` file in a text editor detached from the launching Shell window, we can edit the text, save the edited version, use the Shell window to invoke `\LaTeX` on the edited file, and then simply click ML at an appropriate point in the displayed text to instruct `xdvi` or `yap` to reread the `.dvi` file.\(^69\) Thus, the effect of each new set of edits can be examined quickly without repeatedly starting and exiting from the previewer.

**A.15 The Spell Checker in UNIX\(^70\)**

The programs `ispell` and `aspell`, either or both of which may be part of your `\LaTeX` distribution, are common spell checkers. If either is installed at your site (see the Local Guide), it can be invoked

---

\(^69\)In some environments, rereading an updated file will happen automatically without an explicit user request.

\(^70\)To the author’s knowledge, no comparable stand-alone spell checker exists for Windows.
with a command like\textsuperscript{71}

\texttt{ispell filename} \quad \text{or} \quad \texttt{aspell check filename}

or perhaps with a mouse click in a GUI. For files with extension .tex, \texttt{ispell} enters its \TeX\ mode and will not identify every \TeX\ or \LaTeX\ command as a misspelled word. Note, however, that these commands are simply ignored by \texttt{ispell} and \texttt{aspell}; their correctness or legitimacy as commands to \LaTeX is not assessed.

### A.16 A Sample Document

The following sample document demonstrates some of the commands explained above and also introduces some useful commands that have not yet been mentioned. (The explanatory comments following the \% sign can be included in the code but have no effect on the output produced.)

\begin{verbatim}
\documentclass{article} \% Mandatory command; select \% default 10 point type
\setlength{\textwidth}{6truein} \% Set width of text area
\setlength{\textheight}{9truein} \% Set height of text area
\setlength{\oddsidemargin}{0.25truein} \% Center text on 8.5'' width
\setlength{\topmargin}{-0.5truein} \% Center text on 11'' height
\setlength{\parskip}{6pt plus 1pt minus 1pt} \% Specify extra space
\setlength{\parindent}{40pt} \% Set paragraph indent
\begin{document} \% Mandatory command
\begin{center} \% Large, bold, centered title
\Large \bf A Sample Document for Your Perusal \end{center}
\begin{flushleft} \% Two lines; flush left
\em Author\/: J.~Q.~Student \% Note italic correction \% Tilde prevents extra space
\em Date\/: \today \% \\ forces a new line
\end{flushleft}
\end{document}
\end{verbatim}

In this document we have used the \texttt{\setlength} commands to modify \LaTeX’s default page setup to make fuller use of an 8.5''\times11'' page.\texttt{\footnote{Note the special command \textbackslash{}backslash\{tt LaTeX\}} \footnote{\texttt{\LaTeX} \{‘‘logo’’\}. The title could have been generated using the three commands \texttt{\title{\ldots\title{\ldots\title{\ldots}}}, \texttt{\author{\ldots\author{\ldots\author{\ldots}}}+, and \texttt{\date{\ldots\date{\ldots\date{\ldots}}}+ before the \texttt{\begin{document}+}}}}

\textsuperscript{71}Additional information about \texttt{ispell} or \texttt{aspell} can be found by typing the command \texttt{ispell} with no arguments or the command \texttt{aspell help}. Even more detailed information may be available in the on-line help, accessed in many systems by typing the command \texttt{man ispell} or the command \texttt{man aspell}. Googling \texttt{ispell} or \texttt{aspell} will surely also provide links to detailed descriptions.
command---i.e., in the preamble---followed by the command \verb*+\maketitle+ after the beginning of the document. The command \verb*+\verb*+ ... +? will cause \LaTeX{} to print whatever is between the plus signs (which could be virtually any specified character) exactly as it is. The asterisk is optional, and it makes \LaTeX{} highlight the spaces that occur inside the \verb+verbatim+ environment.

\noindent Notice that a blank line creates a new paragraph. The \verb+\noindent+ command \emph{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.~\eqref{LATEX:demo}. The equation
\begin{equation}
\frac{\partial e^{x_0 y^2}}{\partial x_0} = y^2 e^{x_0 y^2}
\end{equation}
demonstrates partial derivatives and fractions. Note the automatic generation and placement of equation numbers in the output. Also note that pages are automatically numbered, though the actual printing of page numbers can be suppressed with the commands \verb+\pagestyle+ and \verb+\thispagestyle+.

\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 \texttt{\setlength} commands to modify \LaTeX{}'s default page setup to make fuller use of an 8.5" \(\times\) 11" page.\(^a\) The title could have been generated using the three commands \texttt{\title{\ldots}}, \texttt{\author{\ldots}}, and \texttt{\date{\ldots}} before the \texttt{\begin{document}} \texttt{command---}i.e., in the preamble---followed by the command \texttt{\maketitle} after the beginning of the document. The command \texttt{\verb*+\ldots+} will cause \LaTeX{} to print whatever is between the plus signs (which could be virtually any specified character) exactly as it is. The asterisk is optional, and it makes \LaTeX{} highlight the spaces that occur inside the \texttt{verbatim} environment.

Notice that a blank line creates a new paragraph. The \texttt{\noindent} command suppresses the standard paragraph indentation. In this special space we will also demonstrate some math environment operations. Consider the equation

\[
\int_0^\infty e^{-x} \, dx = 1, \tag{A.1}
\]

where a medium sized space, specified by the command \texttt{\,}, is put between the integral sign \(\int_0^\infty\), and the integrand and a small space, specified by the command \texttt{\,}, is inserted between the integrand and the \(dx\) in Eq. (A.1). The equation

\[
\frac{\partial e^{x_0 y^2}}{\partial x_0} = y^2 e^{x_0 y^2} \tag{A.2}
\]

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}.

\(^a\)Note the special command \texttt{\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

\[
\texttt{\renewcommand{\today}{\number\day\space ifcase\month/or January\%}} \\
| or February| or March| or April| or May| or June| or July| or August\% \\
| or September| or October| or November| or December| fi \space\number\year}
\]

The percent signs at the end of the first two lines are \textit{not} superfluous; they guarantee that \TeX{} will see this command as a single line and hence that extraneous spaces will not appear in
A.17. MISCELLANEOUS OTHER FEATURES

the output for occasional dates. Note also that this modification invokes \TeX{}’s \texttt{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. 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.\footnote{Version 4.1, which was released in final form on 11 August 2010, is compatible with \LaTeX{}\,2ε. The APS website \url{publish.aps.org/revtex4} (no www) contains up-to-date information about Re\TeX{}4, links to an assortment of manuals (including one titled \textit{Rev\TeX{}4.1 Author’s Guide}, and a link from which that version can be downloaded. Re\TeX{} is automatically included in many standard \LaTeX{} distributions.}

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{\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{"{o}} will produce ö, the command \texttt{"{n}} will produce ü, and the command \texttt{\c{c}} will produce ˇc. 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} \input{FileName} \end{verbatim}

The problem comes because the backslash that introduces the command \texttt{\input} will, in the \texttt{verbatim} environment, be treated as an ordinary character and will not be recognized as introducing a command. If, instead, we invoke the alternative \texttt{alltt} environment with the \LaTeX{} commands

\begin{verbatim} \input{FileName} \end{verbatim}

the problem is solved, since \texttt{\{} \texttt{,} \texttt{\}}, 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

\texttt{\usepackage{alltt}}

placed in the preamble to the \LaTeX{} file.
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, \TeX embellished with alltt will complain that \n is an undefined command. For this case (and, of course, for the others as well), we need instead exploit the \verbatim package (not environment), which is made available by placing the command
\usepackage{verbatim}

in the preamble. In particular, this inclusion defines a command \verbatiminput, invoked with a statement like
\verbatiminput{FileName}

When the \verbatim package is invoked, we also have available a new environment—the \verbatim environment, which can be used to “bracket” extended text that one wants to exclude from processing by \TeX.73

13. The ten characters ($, &, {, }, %, #, ~, ^, \), which have special meanings to \TeX 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 \verbatim 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 \TeX) 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‘\, \char‘\^, and \tt \char‘\_ will produce characters ‘b’ , ‘a’ , ‘^’ , and ‘\’, respectively. Similarly, the command \tt \char‘> will produce ‘>’, which differs in sometimes desirable ways from the character ‘>’ produced by $\backslash$. 73

14. The commands that allow us to create boxes as described in Sections 6.4.3 and C.13.3 in \TeX 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 \TeX 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 \TeX, 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 \TeX to remember the value at which it ended and retrieve that

\footnote{The \verbatim package also redefines the \verbatim environment in ways, however, of little consequence unless the text in the environment is very extensive. If the \verbatim package is invoked, one particularly significant change in the \verbatim environment results in the complete ignoring of any characters following the statement \verbatim{verbatim} in the same line. It is best always to place the statement \verbatim{verbatim} on a line by itself.}
value when a new environment is entered. The remembered value must be stored in a \TeX\ variable called a \textit{counter}, which must be defined with the command
\newcounter{hold}

where \texttt{hold}, which cannot contain numeric characters, is the name chosen for the counter. Then, the structure
\begin{enumerate}
\item Text of first item.
\item Text of second item.
\setcounter{hold}{\value{enumi}}
\end{enumerate}

\begin{enumerate}
\setcounter{enumi}{\value{hold}}
\item Text of third item.
\item Text of fourth item.
\setcounter{hold}{\value{enumi}}
\end{enumerate}

will achieve the desired end. In the first \texttt{enumerate} environment, the counter \texttt{enumi}, which is used behind the scenes to keep track of the number of items, is initialized to zero. It is then incremented by 1 with each new item. At the end of the environment, \texttt{enumi} stores the number of the last item. The command \texttt{\setcounter} in the first environment saves the current value of \texttt{enumi} in the counter \texttt{hold}, which survives the exit from the environment. Thus, immediately on entry to the second environment, we can use a “reflection” of the first use of \texttt{\setcounter} to restore the value that \texttt{enumi} had reached at the end of the first environment.

16. Parameters influencing placement of “floats” (figures and tables). Sometimes \TeX\ seems to have a mind of its own with regard to placement of floating objects on the current or later pages. Many of these “problems” can be cleared up by tampering with the default values of the parameters that limit the number of floats that can be put at the top or bottom of the page or, even more, by changing the parameters that stipulate the maximum fraction of a page that can be devoted to floats or the minimum amount of text that must appear on a page. These parameters are enumerated at the end of Section C9.1 in \textit{The \TeX\ Manual}. For single column presentations, the most important of these parameters are \texttt{\topfraction}, \texttt{\bottomfraction}, \texttt{\textfraction}, and \texttt{\floatpagefraction}. The command \texttt{\renewcommand} must be used to change the values of these parameters.

17. The way to change section, page, figure, table, and footnote “numbering”. By default in the \texttt{article} class, sections, subsections, subsubsections, pages, figures, and footnotes are labeled with Arabic numbers, e.g., Section 3.1.2. The format of that label as determined from the underlying counters \texttt{section}, \texttt{subsection}, \texttt{subsubsection}, \texttt{page}, \texttt{figure}, \texttt{table}, and \texttt{footnote} can be changed. If, for example, we wished the sections to be labeled with upper-case letters, the subsections with lower-case letters, and the subsubsections with arabic numbers (the default), we would simply execute the instructions
\renewcommand{\thesection}{\Alph{section}.}
\renewcommand{\thesubsection}{\thesection\alph{subsection}.}

\renewcommand{\thetable}{\Roman{table}}
\renewcommand{\thepage}{\roman{page}}

\textsuperscript{74} \textsuperscript{74}Note the explicit periods and the inclusion of the section “number” in the definition of the subsection “number”.ел
Table A.14: Structure to produce full-width text at top of double-column text.

```
\documentclass[\twocolumn]{article}
  \begin{document}

  \twocolumn
  [ \begin{center}
    \begin{large}\f Title\end{large}\hspace{12pt} % Bold title; extra space
    \begin{large}\f Author\end{large}\hspace{12pt} % Bold author; extra space
  \end{center}
  \begin{quotation}
    \noindent ... Text of abstract.
  \end{quotation}

  \end{center}
  ...

  \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 \TeX executes the command \texttt{\the\ldots}, where \ldots stands for the name of the appropriate counter. Redefining this command changes the translation \TeX applies to the counter when generating the “number”.

18. The way to create a full-width heading and abstract above (and on the same page as) a two-column presentation of text, which exploits an optional argument to the command \texttt{\twocolumn} (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 \TeX, 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, \TeX 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

\texttt{\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 \TeX source file.\footnote{The file \texttt{article.cls} in the version of \TeX 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 $\texttt{\textbackslashtablename}$ and can be changed in the way described for $\texttt{\textbackslashfigurename}$ 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 $\texttt{\textbackslashchaptername}$ and can be changed in the way described for $\texttt{\textbackslashfigurename}$ 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 $\texttt{\textbackslashcontentsname}$ and the word ‘Index’ is supplied by the command $\texttt{\textbackslashindexname}$. Each can be changed in the way described for $\texttt{\textbackslashfigurename}$ 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 $\texttt{\textbackslash\}}$ marks the end of each row. Thus, for example, the \LaTeX code

\begin{verbatim}
\begin{array}{cccc}
1 & 2 & 3 & 4 \\
5 & 6 & 7 & 8 \\
9 & 10 & 11 & 12
\end{array}
\end{verbatim}

will produce the display

\[
\begin{pmatrix}
1 & 2 & 3 & 4 \\
5 & 6 & 7 & 8 \\
9 & 10 & 11 & 12
\end{pmatrix}
\]

The entries in each cell can, of course, be much more elaborate than the simple choices made here.

24. The \LaTeX package \texttt{needspace}, which facilitates avoiding awkward page breaks conditionally. For example, when a section title occurs close to the bottom of a page, \LaTeX may place the title and only a single line of the text before turning the page. Sensible styles would dictate that the section be started on a fresh page. The command $\texttt{\textbackslash newpage}$ 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 $\texttt{\textbackslash needspace}$ will depend on what penalties are in effect but will usually be close enough to be acceptable; further, $\texttt{\textbackslash needspace}$ will leave a ragged bottom even if $\texttt{\textbackslash flushbottom}$ 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 $\texttt{\textbackslash flushbottom}$ 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 \textit{no} multiplication sign after the number.
25. The Windows batch command `cleanTEX` defined by the batch file

```
REM delete all TEX intermediate files
```

and the Unix shell script defined by the file

```
#!/bin/bash
# delete all TEX intermediate files
rm -f *.sav *.out *.ind
```

which facilitate removing all \TeX/\LaTeX auxiliary files\footnote{Files with file types `.aux`, `.log`, `.dvi`, `.toc`, `.ilg`, `.idx`, `.bak`, `.sav`, `.out`, and `.ind.} 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

```
REM Delete .bak files
del /s *.bak
del /s *.sav
```

and the UNIX shell script `cleanbak` defined by the the file

```
#!/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\footnote{Files with file types `.bak` and `.sav` and files for which the last character in the name is `~`.} 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 \url{www.tug.org} will yield valuable results. Further, when \LaTeX is installed in UNIX and UNIX-based operating systems (which includes Mac computers), the Shell command `man`, e.g. `man hyperref`, may bring up a

\footnote{Files with file types `.aux`, `.log`, `.dvi`, `.toc`, `.ilg`, `.idx`, `.bak`, `.sav`, `.out`, and `.ind.}
\footnote{Files with file types `.bak` and `.sav` and files for which the last character in the name is `~`.}
Numerous books have also been written for a variety of audiences and with objectives ranging from general discussions to very specific focus on particular tasks. Among the more common books are the following:


*Math Into *\documentclass{article} (Third Edition)*, George Grätzer (Birkäuser, Boston) [ISBN 0-8176-4131-9 or 978-0201433111, 2000]*


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 *\documentclass{article}* 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 *\documentclass{article}* 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 (*\documentclass{article}* environment);
- a right-justified date (using *\documentclass{article}* and either *\documentclass{article}* or the *\documentclass{article}flushright* environment);
- a left-justified block containing an inside address (*\documentclass{article}* 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

\[
\frac{dA}{dt} = -k_A A ; \quad \frac{dB}{dt} = k_A A - k_B B ; \quad \frac{dC}{dt} = k_B B
\]

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}

\begin{flushleft}
\textbf{TO:} Insert name of recipient here.
\textbf{FROM:} Insert name of sender here.
\textbf{SUBJECT:} Insert topic here.
\end{flushleft}

\textbf{MESSAGE:}

\textit{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
\texttt{memo\_template.tex}, so you have it available as a starting point for the multitude of memos you
will subsequently write.

A.6. In connection with a grant you received from the XYZ Foundation, you need periodically to submit
a progress report. The format and content of that report are dictated by the Foundation, and
the report is created by filling in the information indicated in italics in the template on page 499.
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>
</tr>
</thead>
<tbody>
<tr>
<td>Original grant</td>
<td>$xxx,xxx</td>
</tr>
<tr>
<td>Total expenses reported last time</td>
<td>$xxx,xxx</td>
</tr>
<tr>
<td>Available funds at start of current period</td>
<td>$xxx,xxx</td>
</tr>
<tr>
<td>Expenses in current period:</td>
<td></td>
</tr>
<tr>
<td>Salaries</td>
<td>$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>
<tr>
<td>Total for current period</td>
<td>$xxx,xxx</td>
</tr>
</tbody>
</table>

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

del nameonly.txt

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
# rdfileeps  (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 -l
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
opks-sum at this URL provide a description of the single- and double-precision components in the package. (Netlib is a large repository of programs maintained at Oak Ridge National Laboratory.)

**Mac\TeX**

Mac\TeX{} is a shareware implementation of the \TeX/\La\TeX{} system for Macintosh computers. Information is available at the URL \url{www.tug.org/mactex}.

**MAXIMA**

Originally called MACSYMA, this first of the computer algebra systems was developed at MIT and supported from its origin in 1968 until 1982 by MIT, NASA, ONR, and DOE. In 1982, MACSYMA became a commercial product that was further developed and remained available until about 1999. The 1982 MIT version remained available as DOE MACSYMA but was released in 1999 to a group that continues to develop and maintain the program, changing its name to MAXIMA, whose website is at the URL \url{http://maxima.sourceforge.net/}. MAXIMA is freely available for a wide variety of platforms.

**MAPLE®**

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: \url{www.maplesoft.com}

**MATHEMATICA®**

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: \url{www.wolfram.com}

**MATLAB®**

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: \url{www.mathworks.com}

**MiK\TeX**

MiK\TeX{} is an implementation of the \TeX/\La\TeX{} system for Windows computers. Information is available at the URL \url{miktex.org}. MiK\TeX{} is also available from CTAN, the \TeX/\La\TeX{} distribution network accessible from links at \url{www.tug.org}.

**MUDPACK**

The freely-available package of FORTRAN PDE solvers called MUDPACK, written by John C. Adams and others, originated at the National Center for Atmospheric Research (NCAR). The main web page for information about this FORTRAN package, and many others, is \url{www2.cisl.ucar.edu/research-software/software}. MUDPACK can be downloaded from the webpage \url{github.com/NCAR/NCAR-Classic-Libraries-for-Geophysics}. While this package is no longer under active development, it remains available and useful.

**Numerical Algorithms Library**

The Numerical Algorithms Library (NAG library) is a large commercially available library of C and Fortran subroutines/subprograms implementing a wide assortment of numerical and statistical algorithms.

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: \url{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 T\TeX/L\TeX\A. Information about licensing and downloads can be found at the URL www.winedt.com. Winedt is also available from CTAN, the \TeX/L\TeX\A 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

$\textsc{head}$, 11, 33, 39, 44, 125, 223, 244, 268, 286, 323, 324, 372, 455

$\textsc{idlhead}$, 11, 39, 244, 324

$\textsc{mathemhead}$, 137

$\textsc{nrhead}$, 11

A

aborting execution
in IDL, 16
in MATHEMATICA, 92
accuracy
in evaluating integrals, 281–283, 286–289
in finding roots, 322–323
in solving ODEs, 220–221, 230, 235
with conservation laws, 221, 235
in solving PDEs, 373–375, 387–388, 413–414, 431–433
acroread, 460
action statements, 151, 154
Adams/Bashforth/Moulton (ABM) methods, 222–223
adaptive methods
for integrals, 283
for ODEs, 221–222
adding text to graphics, see placing text in graphics
algorithms
definition, 151
for $a^2 + b^2 = c^2$, 177
for adding, 177
for bowling, 176
for controlling loops, 160
for counting, 160, 161
for exchange, 160
for finding extremes, 161, 162
for integration, 278–286
for Laplace’s equation, 162
for ODEs, 214–223
for parking lot, 175
for root finding, 320–323
for searching, 162, 176
for sorting, 177
for summing, 161
animation
in IDL, 66–68, 398–400, 404–405
in MATHEMATICA, 133
arithmetic operators
in IDL, 23
in MATHEMATICA, 95
array indices
in IDL, 19, 168
in MATHEMATICA, 100
in pseudocode, 153
arrays
for data storage, 164
in IDL, 19–25
in MATHEMATICA, 99
in pseudocode, 153
scalar, see scalar fields
vector, see vector fields
ASCII character codes, 5
ASCII text files, 8
aspell, 487–488
assigning values to variables
in IDL, 18
in MATHEMATICA, 96–97
in pseudocode, 154

B

backsubstitution, 339
backward difference, 368
Bessel functions, 306, 346
binary files, 8
binary number system, 4–5
Biot-Savart law, 271
bisection, see root finding
bit, 4–5
black body radiation, 144, 304
blocking statements, see also control structures
in IDL, 31–32, 38, 66
in MATHEMATICA, 123
in pseudocode, 156, 166

Boolean values, see logical expressions
Boolean variables, see logical expressions

boundary conditions
  Dirichlet, 353, 367, 369–370
  mixed, 353, 367, 369–370
  Neumann, 353, 367
  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, 208–211, 350, 353

byte, 5–6

C
  C, 1
  cardioid, 43, 74, 302
  carrying capacity, 192, 203–204
  Cartesian coordinates
    in MATHEMATICA, 149
  case sensitivity
    in IDL, 16
    in MATHEMATICA, 92
    in pseudocode, 152
  cells
    in a MATHEMATICA notebook, 101
  center of mass
    definition of, 258
    of cardioid, 302
    of semicircular plate, 259, 267
  central difference, 368
  central processing unit, 3
  chaos
    Lorenz system, 253
  character codes, 4–5
  character strings
    in IDL, 17, 18
    in MATHEMATICA, 94
  characteristic equation, 310
  charging capacitor, 82
  classification
    of boundary conditions, 353–354
    of integrals, 257
    of ODEs, 187
    of PDEs, 353
  closing notebook
    in MATHEMATICA, 93
  color
    in IDL, see also IDL keywords: color, 54, 67, 72, 74–76
    combining statements on one line
    in IDL, 16
    in MATHEMATICA, 92
    in pseudocode, 156
  command files
    in IDL, see IDL batch files
    in MATHEMATICA, see MATHEMATICA command files
  command line editing
    in IDL, 16
    in MATHEMATICA, 101
  commands
    in IDL, see IDL commands
    in MATHEMATICA, see MATHEMATICA commands
    in pseudocode, see pseudocommands
  comments, see also documentation
    in IDL, 35, 38
    in LATEX, 488
    in MATHEMATICA, 125
    in pseudocode, 157, 160
  common storage
  compiler directives, 153
    in IDL, see IDL executive commands
    in pseudocode, see pseudocode directives
  compiling source file, see Local Guide, 152
  complex numbers
    in IDL, 17
    in MATHEMATICA, 105–107
  Compton scattering, 84
  conditions, see logical conditions
  conservation of mass, 362–363
  constants
    in IDL, see IDL special symbols
    in MATHEMATICA, see MATHEMATICA constants
  contact information, 503–506
  continuing statement on second line
    in IDL, 16
    in MATHEMATICA, 92
    in pseudocode, 156
  control characters, 4
  control structures
    in IDL, see IDL control structures
    in MATHEMATICA, see MATHEMATICA control structures
    in pseudocode, see pseudocode control structures
  convergence criteria, 164, 220–221, 281–283, 322–323
  convergence error, 370–371, 396
  convert .eps/.ps to .pdf, 482–483
  convert .pdf to .eps/.ps, 483
  Cornu spiral, 262, 293–294
  coupled oscillators, 194–195, 207–208, 248
  CPU, see central processing unit
  Cramer’s rule, 338–339
  cross product
INDEX

data analysis, 261

data files
freelfall.dat, 348, 349
irradidl.dat, 169, 175
ldrum.dat, 66
pdensidl.dat, 170
quadpoleidl.dat, 174
radio.dat, 33, 35
rcdata.dat, 44
ruchardt.dat, 351
wavegdidl.dat, 173

data types
floating point, 6–7
integer, 6–7
string, 6–7

data typing
explicit, 153
implicit, 153
in IDL, 16
in pseudocode, 152–153, 155–156

default directory, see also current directory, 9
in IDL, 28–30, 35–37, 44, 48, 76, 238
in MATHEMATICA, 124, 131, 132, 136–137

defining ODEs
for IDL, 226–227
for MATHEMATICA, 115, 201, 203, 204, 207, 208, 246

deleing functions
in MATHEMATICA, 98

deleing variables
in IDL, 21
in MATHEMATICA, 97, 98, 109

demos
in IDL, 16
in MATHEMATICA, 91, 93
diffraction patterns, see also interference patterns, 168, 306, 346, 347
of square aperture, 48–54
diffusion equation
in 1D, 358–359
in IDL, 402–408
in 2D, 391–393
in 3D, 359
dimensionless casting
for plotting, 84
of integrals, 259–261, 263, 271, 301, 305
of ODEs, 118–119, 188–193, 195–197, 203–204, 309, 311
direct algorithm, 370
discretization
of Laplace’s equation, 163
of ODEs, 368–370
of PDEs, 354, 375
discretization error, 370–371, 374, 396
displaying data on screen
in IDL, 20–22
in MATHEMATICA, 92
in pseudocode, 154
dividing statement between lines, see continuing statement on second line
documentation, see also comments
in IDL pro-files, 38
in MATHEMATICA notebooks, 134–135
dot product
in IDL, 25, 80
in MATHEMATICA, 125–126, 150
double factorial, 12
drawing space curves
with IDL, 69–71
with MATHEMATICA, 138–139
dvipdfm, 460
dvips, 459, 477
dynamic allocation of memory, 153

E
elastic swing, 254
electric field, 262
in wave guide, 62–65, 171–173
of dipole, 180
of quadrupole, 173–174
of relativistic particle, 84
electrostatic potential, 262
of charged disks, 82, 83, 301
of charged rings, 262–263, 295, 299–300, 306
of finite line charge, 268–270
of spherically symmetric distribution, 302
elemental equation in finite element analysis, 380
elliptic integrals, 148, 261, 305
e m dash, 468
e m space, 465
e n dash, 468
e n space, 465
energy levels
in quantum oscillator, see harmonic oscillator
in quantum well, 211, 313, 329–331, 335–337, 346
equation of continuity, 362–363
equilibrium points, 308, 316–317
equivalent resistance, 348
error function, 262, 264, 286, 290, 292, 298–299, 304
error messages
in \LaTeX, 485–486
Euler’s method, 214–216, 223–225
improved, see improved Euler method
evaluating integrals
with IDL, 286–296
with MATHEMATICA, 112, 265–277, 296–300
exchange algorithm, 160
executing program, see Local Guide
  exiting
    from IDL, 16
    from MATHEMATICA, 93
expressing infinity
  in MATHEMATICA, 266
expressions
  in MATHEMATICA, 94–96

F

Fabry-Perot interferometer, 83
factorial, 12, 102
Fibonacci numbers, 81
field mapping, 86, 87
finding eigenvalues/eigenvectors
  with IDL, 39, 81
  with MATHEMATICA, 110–111, 145, 146
finding roots, see root finding
finite difference approximations
  for derivatives, 177
  for PDEs, 354
finite difference methods
  error in, 370–371
  in 1D, 368–375
  in 2D, 388–414
  with IDL, 371–375, 396–414
finite element approximations
  for PDEs, 354
finite element methods
  assembling equations, 375
  in 1D, 375–388
  in 2D, 414–433
  incorporating boundary conditions, 375
  with IDL, 384–388, 423–433
fitting data, see curve fitting, least squares fitting
floating point number
  double precision, 7
  single precision, 6
fluid dynamics, 361
fonts, see also Greek letters
  in IDL, 51, 52, 68–69
formatting data files, 164–165
formatting output
  in IDL, 29–30
  in pseudocode, 154
FORTRAN, 1
forward difference, 368
four-slit interference pattern, 84
Fourier series, 265, 275–277, 302
Fraunhofer diffraction, 168–169
Fresnel integrals, see also Cornu spiral, 262, 306
functions
  in IDL, see IDL functions
  in MATHEMATICA, see MATHEMATICA functions
typical available, 154

G

Galerkin method, 375, 377–380
gamma function, 12–13
Gauss-Jordan elimination, 375
Gauss-Kronrod rule, see quadrature
Gaussian distribution, 261, 304
Gaussian elimination, 339–340, 375
Gaussian integration, see quadrature
gedit, 1
ghostview, 460
global positioning, 348
global truncation error, see truncation error, global
global variables
  in IDL, see common storage
golden ratio, 345
gravitational force, 196
Greek letters, see also fonts
  in IDL, 51, 68–69
  in \LaTeX, 468–469
  in MATHEMATICA, 94

H

hard copy
  from IDL, 47–48
  from MATHEMATICA, 129–132
hazards of integer arithmetic, 17
Helmholtz coil, 270–274
Helmholtz equation, 360
help, see on-line help
hexadecimal number system, 5
high-pass filter, 44–47
high-resolution graphing
  in IDL, 26–27, 40–65
  in MATHEMATICA, 126–129
Hooke’s law, 309
hyperlinks, 481–482

I

IDL, 1
idl, 15
IDL batch files, 35–36
IDL command files, see IDL batch files
IDL commands, see also IDL functions, IDL procedures, IDL programs
  axis, 43, 74, 331
bindgen, 22
broyden, 343
INDEX

bytarr, 20, 66
byte, 18, 23
call_function, 36
cd, 76–77
cindgen, 22
close, 28–30, 33, 35, 44, 66
common, 38, 229–233, 235, 238, 295–296, 326–328
complex, 18, 23
complexarr, 20
contour, 49, 52–53, 58, 65
cramer, 341
dblarr, 20
dblcomplexarr, 20
dcindgen, 22
dcomplex, 18
delvar, 21
demo, 16
device, 48, 53, 74, 76
dindgen, 22
doc_library, 35, 38
double, 18
eigenql, 40
eigenvec, 40
elmhes, 40
erase, 70, 71
errplot, 45
exit, 16
extract_slice, 61
findgen, 22, 26, 36, 41, 43, 44, 46, 58, 68–70, 73, 74, 76, 231, 233, 239, 287–289, 292, 294, 296, 326–328, 397
fix, 18
float, 18, 74, 287–289, 325, 326
flow3, 65, 88
for matrix multiplication, 24
fx_root, 323–325
fx_roots, 323, 324, 326
general structure of, 19
help, 20–21
hqr, 40
indgen, 22
intarr, 20, 21
invert, 25, 26
isosurface, 60
journal, 30–31, 47
lindgen, 22
loadct, 74
lonarr, 20
long, 18
lsode, 227–228, 231–233
lubisect, 325, 330
ludc, 80, 341, 430
ludiffeq, 23, 228–230, 232, 238, 239, 241, 243
ludiffeq, 45, 228–229, 234–235, 237
lueigen, 39–40
luplot, 36–38
luqsimp, 290–296
lusol, 80, 341, 430
max, 70, 328
min, 70
movie, 67
newton, 324–325, 328, 343
obj_new, 61
openr, 29–30, 33, 35, 44, 66
openw, 28–29
oplot, 33–35, 41, 46, 76, 224, 230, 233, 238, 241, 243, 236, 327, 331, 373
plot_field, 64
plot_i0, 45
plot_i1, 45
plot_o0, 45
plotradio, 35
plotrc, 47
plots, 47, 70
polar_contour, 55–56
polar_surface, 55–56
polyshade, 60
print, 20–26, 29–32, 38, 39, 49, 50, 70, 224, 230, 235, 287, 288, 290, 324, 325, 328, 330, 331, 374, 387, 397, 401, 403, 431
printf, 28–30
project_vol, 61
ptr_new, 58, 60
qromb, 291
qsimp, 291
read, 79, 371, 384, 397, 399, 402, 404, 408, 424
readf, 29–30, 33, 35, 44, 66, 78
restore, 28, 78
rk4, 227–228, 230–231, 233
save, 28, 78
scale3, 60, 70, 88
set_plot, 48, 74, 76
shade_surf, 49, 53–54, 58, 66, 168
show3, 49, 55
showfont, 69
sindgen, 22
size, 21, 241, 243, 292, 294, 296
slicer3, 58, 60
sort, 80
stop, 403
strarr, 20
string, 18, 431
surface, 49–52, 56, 58, 167
svdc, 341
svsol, 341
total, 80, 287
transpose, 24–26, 40, 288, 326
triql, 40
trired, 40
trisol, 341, 371
tvrd, 66
tvsc1, 60, 67
vel, 63–64
veloject, 63, 65
voxel_proj, 61
wait, 399, 404, 405, 412
wdelete, 72–73
window, 66, 72–74
wset, 72–73
wshow, 72–73
x_interanimate, 67–68
xmanager, 169
xobjview, 61
xplot3d, 71
xsurface, 49, 54
xvolume, 61
xyouts, 33–35, 47, 241–243

IDL contact information, 503

IDL control structures
begin/end, 31–32, 38, 39, 66, 327, 328, 399, 404, 405, 426, 428–430
if/then/else, 32–33, 241, 243, 397, 399, 403, 405, 426, 428, 429
while/do, 32

IDL data types, 16–18

IDL development environment, 15

IDL executive commands
.compile, 36, 230, 293, 294, 329
.reset_session, 78

IDL functions, see also IDL commands, IDL procedures, IDL programs, 19, 35–38
2oscil, 245
abs, 237
apollo, 245
built-in, 23
cos, 43, 55, 63, 68, 70, 73, 74, 168, 173, 293, 296, 328, 330
damp2hm, 245
damposc, 232, 233, 245
decay, 229, 245
determ, 23
drvmosc, 245
drvmpec, 245
exp, 41, 46, 55, 57, 170, 287–290, 326, 328
findgen, 371, 385, 403, 408
fix, 371, 384, 397, 399, 402, 408, 424
float, 371, 384, 397, 399, 402, 404, 408, 424
gaussint, 290
henon, 245
indgen, 22
fltarr, 428
intarr, 30, 426
inverse, 23
largamp, 245
lineshape, 36
lineshape1, 38
lorenz, 245
lubisect, 324, 352
lugen_grid, 49–50, 55, 57, 63, 68, 90, 168, 170, 173, 174
lugen_list, 168–170, 173, 174, 182–183
lugen_list_event, 169, 184–185
luread_list, 175, 186–187
max, 374, 387, 401
min, 374, 387, 401
moment, 291
onedshm, 245
planet, 235, 245
projectile, 326
qmlevels, 329
qmlower, 329
qmshm, 241, 245
qmupper, 329
reverse, 22
rings, 296
rossler, 245
sin, 41, 50, 63, 68, 70, 73, 76, 168, 173, 293, 326, 329, 330
size, 230, 239
spiral1c, 293
spirals, 293
sqrt, 46, 57, 170, 287–290, 326, 330
stdwaves, 238, 245
transpose, 22
trisol, 372
turnpts, 325
twodshm, 245
user-defined, 36
vandpol, 245

IDL keywords, 19, 42, 45, 52
atol, 227
ax, 51, 71
az, 51, 71
c_labels, 53
c linestyle, 53, 65
char_size, 27, 33, 35, 41, 43, 48, 51–53, 56, 58, 69, 74, 224, 230, 237, 331
close, 48, 74, 76
color, 63, 74, 76
decomposed, 53, 74
double, 328
drvmpec, 245
drvmpec, 326, 328
filename, 28, 48, 74, 76, 78
format, 29–30, 431
free, 66, 72
get_lun, 29
grid, 56
inches, 74
init, 228, 230, 232, 235, 237
isotropic, 56, 58, 73
INDEX

length, 63, 64
levels, 53, 65
linestyle, 41, 42, 46, 47, 224, 230, 233, 237
maxsteps, 228
n, 64
nlevels, 52, 56, 58
nsteps, 64
nvecs, 63
nx, 49–50, 55, 57, 63, 68, 168, 170, 173, 174
ny, 49–50, 55, 57, 63, 68, 168, 170, 173, 174
nz, 57, 170, 174
overplot, 65
polar, 43, 74
polygons, 61
prompt, 79, 371, 384, 397, 399, 402, 404, 408, 424
psym, 47
rtol, 227
set, 67
showload, 67
subtitle, 43, 74, 233
system_variables, 78
t0, 228, 230, 232, 235, 237
t3d, 60, 70, 71
tf, 228, 230, 232, 235, 237
plot, 373
ticklen, 27, 41, 69, 326, 328, 373, 387
title, 27, 41, 43, 45, 51, 52, 63, 66, 69, 72, 224, 230, 233, 237, 327, 373, 387, 402, 407
tol, 228, 232
tolf, 328
tolx, 328
verbose, 28
xaxis, 43, 74
xlog, 45
xoffset, 74
xpos, 72
xrange, 49–50, 55, 57, 60, 63, 68, 70, 168, 170, 173, 174, 327
xsize, 66, 72–74
xstyle, 34, 35, 41, 43, 74, 331
xtitle, 33, 41, 45, 51–53, 58, 63, 65, 69, 224, 230, 233, 237, 327, 328
yaxis, 43, 74, 331
ylog, 45
yoffset, 74
ypos, 72
yrange, 45, 49–50, 55, 57, 60, 63, 68, 70, 76, 168, 170, 173, 174, 241, 243, 326, 327, 402, 407
ysize, 66, 72–74
ystyle, 34, 35, 41, 43, 45, 74, 331
ytitle, 33, 41, 45, 51–53, 58, 63, 65, 69, 224, 230, 237, 327, 328, 331
zrange, 52, 57, 60, 66, 170, 174
ztitle, 51, 52
IDL linestyle codes, 42
IDL pro-files, see IDL batch files, IDL functions, IDL procedures
IDL procedures, see also IDL commands, IDL functions, IDL programs, 19, 35–36
ludc, 385
lusol, 385
user-defined, 36
IDL programs, see also IDL commands, IDL functions, IDL procedures
fdm1d.pro, 372, 441
fdmdiffus1d.pro, 404, 445–446
fdmlap2d.pro, 409, 446–447
fdmlap2d.plot.pro, 447
fdmwaved.pro, 398, 444–445
fem1d.pro, 386, 442–443
fem2d.pro, 430, 451–453
IDL special symbols
13 (simplex Roman; default), 51, 52, 69, 327, 328
14 (simplex Greek), 51, 52, 327, 328
16 (complex Roman), 69
18 (complex Italic), 69
1C (new line), 51, 52
1D (exponent level), 331
1E (exponent level), 69
1N (normal level), 69, 331
!p, 168, 170
"..." (enclose strings), 18
'...' (enclose strings), 18
* (multiplication), 23
* (wild card), 21, 33
+ (addition), 23
- (subtraction), 23
/ (division), 23
: (range), 22
; (comments), 35
++;-- (documentation), 38
= (assignment), 18
? (help), 34
@ (execute command file), 36, 38, 373, 386, 387, 401
$ (continue statement), 16, 27, 38, 41, 43, 48, 50, 51, 53, 56, 57, 60–61, 65, 66, 374
& (separate statements), 32, 371–373, 387
^ (exponentiation), 23
Inf (∞), 50, 79
NaN (not a number), 79, 330
IDL symbol codes, 46
IDL system variables, 71–72, 74
!p.background, 71
!p.charsize, 71, 327
!p.color, 72
!p.linestyle, 72
!p.multi, 44, 71, 233, 327, 373, 402, 407
indices in \LaTeX, 479
indicial equation, 211
inert commands
  in \textsc{Mathematica}, see suppressing
evaluation in \textsc{Mathematica}
inhomogeneities in linear equations, 338
initial value problems, 187, 353
initialization files
  in IDL, 77
  in \textsc{Mathematica}, 135
integer
  signed, 6
  unsigned, 6
interference patterns, see also diffraction
  patterns
Fabry-Perot, 83
  from four slits, 84
internal storage of expressions
  in \textsc{Mathematica}, 95–96
interpolation functions, see shape functions
interpolation points, 279, 280
interpreting source code, see also \textit{Local Guide}, 152
interrupt
  in IDL, 16
  in \textsc{Mathematica}, 92
ispell, 487–488
iterative algorithm, 371
\textbf{K}
Kaprekar Constant, 89
keywords
  in IDL, see IDL keywords
Kirchoff’s laws, 348
Knuth, Donald, 455
Kronecker delta, 264
\textbf{L}
Lagrange equations
  for 2D harmonic oscillator, 146
  for compound pendulum, 147
Laguerre polynomials, 144
Laguerre’s method, see root finding
Lamport, Leslie, 455
\texttt{LAPACK}, 1
  contact information, 503
Laplace transforms, 13–14, 199–200, 249
  definition, 13, 114, 199
  for ODEs, 199
  in \textsc{Mathematica}, 114, 148
Laplace’s equation, 350, 359, 360
  in 2D, 394–396
  in IDL, 408–414
  in IDL, 167–168
  in pseudocode, 162–164
\LaTeX, 1
  accents in, 491
  counters in, 492
  dashes in, 467
  math mode in, 468–469
  quotation marks in, 468
  special characters in, 492
\texttt{latex}, 458, 459, 477
\LaTeX commands
  \verb|\backslash| (italic correction), 464
  \verb|\backslash| (new line), 466, 471
  \verb|\neg| (negative thin space), 469
  \verb|\thin| (thin space), 469
  \verb|\hyphen| (allow hyphen), 468
  \verb|\medium| (medium space), 469
  \verb|\thick| (thick space), 469
  \addcontentsline, 478–479
  \addtocounter, 478–479
  \Alph, 470, 493
  \alph, 470, 493
  \arabic, 470
  \author, 488
  \bf, 488
  \bfseries, 488
  \caption, 471, 472, 474, 477
  \centerline, 474
  \chapter, 468
  \chaptername, 495
  \char, 492
  \clearpage, 465
  \contentsname, 495
  \date, 488
  \definecolor, 481, 485
  \documentclass, 456, 458, 461, 482, 488
  \documentstyle, 456
  \fbox, 492
  \figurename, 494
  \footnote, 468, 488
  \footnotesize, 465
  \frac, 488
  \framebox, 492
\hspace, 465, 466
\hspace*, 465
\Huge, 465
\huge, 465, 488
\IfFileExists, 485
\ifthenelse, 484–485
\includegraphics, 472–475, 482
\index, 479
\indexname, 495
\input, 491
\int, 488
\item, 469–470, 493
\itsshape, 464
\label, 468, 471, 472, 474, 477, 488
\labelitemi, 470
\labelitemii, 470
\labelitemiii, 470
\labelitemiv, 470

\LARGE, 465
\Large, 465, 488
\large, 465
\LaTeX, 488
\listoffigures, 478–479
\listoftables, 478–479
\makebox, 492
\makeindex, 479–480
\maketitle, 488
\mathbf, 465
\mathcal, 491
\mathit, 469
\mathrm, 469
\mathtt, 469
\mbox, 468, 469, 492
\mdseries, 464
\Needspace, 468
\maketitle, 488
\newboolean, 484–485
\newcommand, 489
\newcounter, 493
\newlength, 466
\newpage, 465
\noindent, 488
\enumerate, 493
\footnote, 493
\page, 493
\secnumdepth, 478

\setboolean, 484–485
\setcounter, 493
\setlength, 461, 465, 488
\settodds, 465
\settoheight, 465
\settowidth, 466
\small, 465
\section, 468
\subsection, 468
\subsubsection, 468
\tableofcontents, 478–479
\textbf, 464
\textit, 464
\textmd, 464
\textnormal, 464
\textsf, 464
\textsl, 464
\texttt, 464
\textup, 464
\the, 493–494
\thenumii, 470
\thenumiii, 470
\thenumiv, 470
\thispagemargin, 488, 491
\tiny, 465
\today, 489, 490
\ttfamily, 464
\twocolumn, 494
\typein, 485
\typeout, 485
\upshape, 464
\usepackage, 467, 473, 479
\verbatim, 493
\verb, 464, 488, 492
\verbatiminput, 492
\vspace, 465, 472
\verbs, 465, 494

\LaTeX contact information, 455, 505
\LaTeX counters

enumi, 470, 493
enumii, 470
enumiii, 470
enumiv, 470
figure, 493
footnote, 493
page, 493
secnumdepth, 478
section, 493
subsection, 493
INDEX

subsection, 493
table, 493
tocdepth, 478
\TeX declarations, see \TeX commands
\TeX document classes, 457
article, 461, 468
book, 461, 468
report, 461
slides, 461
\TeX environments, 466–467
\(...\) (in-line equation), 468
\[ ... \] (displayed equation), 468
$...$ (in-line equation), 468
array, 495
center, 466, 471, 473, 477, 488
comment, 492
displaymath, 468
document, 457, 458, 488
enumerate, 469–470, 492, 493
eqnarray, 468
equation, 468, 488
figure, 472–475, 477
flushleft, 488
items, 469–470
minipage, 492
picture, 472, 478
quotation, 466
ref, 466
table, 470–472
tabular, 470–472
tikzpicture, 475–478
\TeX length parameters, 462
\baselineskip, 465, 467
\baselineskip, 467
\bottomfraction, 493
\evensidemargin, 461, 480
\floatpagefraction, 493
\itemsep, 470
\marginpar, 480
\marginparwidth, 480
\oddsidemargin, 461, 488
\parindent, 461, 488
\parsep, 470
\parskip, 461, 470, 488
\textfraction, 493
\textwidth, 461, 480, 488
\topfraction, 493
\topmargin, 461, 488
\TeX Manual, 455
\TeX output
from MATHEMATICA, 132–133
\TeX packages, 467
alltt, 491
color, 481, 485
graphix, 472–475, 481–482
hyperref, 481–482, 485
ifthen, 484–485
imakeidx, 479–480
makeidx, 479–480
needspace, 495
notebook2e, 133
pict2e, 495
picture, 472, 478
quote, 466
quote, 466
ref, 470–472
verbatim, 491–492
\TeX page previewers, see also xdvi, yap, 486–487
\TeX preamble, 461–463
\TeX sample, 457–460, 488–489
\TeX special symbols
' (closing double quotation), 468
--- (em dash), 467
-- (en dash), 467
- (dash), 467
- (minus sign), 468
~ (hard space), 467
" (opening double quotation), 468
\TeX templates, 462
\TeX type size, 464
\TeX type styles, 463–464
lawn mower, 344
least squares fitting, 348, 349, 351
Legendre polynomials, 142, 286, 303, 345
Lennard-Jones potential, 348
licensing of software, 14
line labels
in MATHEMATICA, 92
linking object code, see Local Guide, 152
lists
in MATHEMATICA, 94–96, 99–100, 207
logical conditions, see logical expressions
logical expressions
in IDL, 19, 32–33
in \TeX, 484–485
in MATHEMATICA, 123–124
in pseudocode, 155–156
logical operators, see logical expressions
logical unit number (LUN)
in IDL, 29
logistic growth, 192, 203–204
long integer
signed, 6
unsigned, 6
loops
in IDL, 31–33
in MATHEMATICA, 103, 108, 122–123
in pseudocode, 157–159
Lorentz distribution, 82, 301, 305
Lorentz lineshape, 37
lp, see UNIX commands
LSODE, 1
LSODE contact information, 503
LU decomposition, 340, 375
INDEX

**M**

- machine code, 152
- Macintosh OS, 1
- MacTeX
  - contact information, 504
- magnetic field, 262
  - of circular loops, 41, 48, 85, 177, 270–274, 306
  - of Helmholtz coil, 270–274
  - of solenoid, 84
- magnetic vector potential, 262
- MAPLE, 1
- MAPLE contact information, 504
- MATHEMATICA, 1
- mathematica, 92
- MATHEMATICA command files, 124–126
  - crossdot, 125, 150, 206, 217
  - testmultiply, 125
- MATHEMATICA commands, see also MATHEMATICA functions, 94
  - Animate, 133
  - Apart, 104, 140, 148
  - Append, 137
  - Black, 127, 128
  - Boole, 123
  - Cancel, 118
  - Clear, 97, 98
  - CoefficientArrays, 209
  - CoefficientList, 142
  - Collect, 117, 142, 202, 208, 213
  - ComplexExpand, 209–213
  - Conjugate, 105
  - ContourPlot, 129
  - Cross, 126
  - Curl, 120, 149, 150
  - D, 112–116, 126, 142, 145, 246, 317
  - Dashing, 127, 128
  - Delete, 137
  - Det, 110, 210
  - DiagonalMatrix, 100
  - Directory, 136
  - Display, 131
  - Div, 120, 149, 150
  - Dot, 126
  - DSolve, 115–119, 200–206
  - Eigensystem, 111
  - Eigenvalues, 111
  - Eigenvectors, 111
  - Element, 210, 275–276
  - Evaluate, 246
  - Exit, 102
  - Export, 131, 132
  - ExpToTrig, 106, 107, 208
  - Factor, 103, 117, 140, 320
  - Remove, 109
- FileNames, 137
- FilePrint, 126
- FindRoot, 273, 332–336, 341, 343
- FullForm, 104, 112, 113, 141
- Get, 124, 125, 268, 271
- Grad, 120, 149, 150
- Graphics, 138, 337
- Head, 103, 113
- HoldForm, 113
- IdentityMatrix, 100
- Insert, 137
- Integrate, 112, 114, 265–277, 296, 298
- Inverse, 110
- InverseLaplaceTransform, 114, 148
- LaplaceTransform, 114
- Laplacian, 120, 149, 150
- Limit, 114, 203, 206, 266
- Line, 337
- LinearSolve, 109, 341
- ListPlot, 299, 300, 334, 337
- lucross, 125–126, 271
- ludot, 125–126, 268, 271
- Map, 117, 336
- MatrixForm, 100, 110, 146, 210
- N, 102, 109, 267, 273, 274, 296, 298, 316, 317, 331, 341
- Names, 97
- NDSolve, 246–247
- NIntegrate, 296–300
- Normal, 113, 114, 209, 269, 270, 273, 275, 319, 320
- NSolve, 109, 331–337, 341
- Numerator, 273
- Off, 136
- On, 136
- Options, 94, 129, 138
- ParametricPlot, 128
- ParametricPlot3D, 139
- Part, 103–105, 141
- Plot, 126–128, 131, 133, 138, 204, 206, 246, 272, 277, 318
- Plot3D, 129, 133
- Prepend, 137
- Print, 103, 108, 124–125
- Protect, 98
- Put, 132
- PutAppend, 132
- Range, 99, 299, 300
- ReleaseHold, 114
- Remove, 97, 98
- Replace, see /. 104
- ReplacePart, 105, 137, 141
- Rest, 336
- Root, 315
- Series, 112, 114, 202, 267, 269, 270, 272, 274, 275, 319, 320
- SetDirectory, 136–137
INDEX

SetOptions, 137
Show, 138, 337
Solve, 108, 109, 117, 201, 210, 212, 213, 266, 273, 315–319, 331, 341, 343
StyleForm, 127, 128, 204, 206
Sum, 277
Table, 99, 100, 212, 213, 247, 299, 300, 334, 336, 337
TeXForm, 132
Text, 127, 138
Thickness, 127, 128, 204, 206, 272, 299, 300, 318, 334, 337
Tr, 110
Transpose, 110
TrigExpand, 107, 108
TrigReduce, 107, 108
TrigToExp, 107
Union, 207
Unprotect, 98
ZeroMatrix, 146
MATHEMATICA constants, 94
E (2.71828...), 94
False, 94
Im (√−1), 94, 105, 106, 207
Infinity (∞), 94, 112, 266
Pi (π), 94, 98, 99, 102, 129, 210, 298, 299, 334
True, 94, 202, 203
MATHEMATICA contact information, 504
MATHEMATICA control structures
For, 103, 108, 122–123, 299, 300, 334–337
If, 112, 123–124
While, 123
MATHEMATICA domains
Integers, 210, 275
MATHEMATICA functions, see also
MATHEMATICA commands, 94–95, 97–98
Abs, 106
ArcTan, 106
Arg, 106
ConditionalExpression, 112, 332
Cos, 95, 97, 105, 107, 126, 133, 139, 208, 268, 271, 275–277, 299, 318, 332–334
Cosh, 105
Cot, 336
Exp, 106, 112, 114, 126, 203, 298, 299, 318, 334
Factorial, 102
Im, 105, 106, 112
LegendreP, 143
Log, 270
Re, 105, 106
Sin, 98–100, 105, 107, 113, 114, 126, 129, 131, 133, 138, 139, 208, 265, 268, 271, 275–277, 318, 335
Sinh, 105
MATHEMATICA heads
Plus, 95, 96, 103, 104, 113
Power, 96, 103, 104, 113
Rational, 112, 113
SeriesData, 112
Times, 95, 96, 103, 104, 113
MATHEMATICA logical expressions, see logical expressions in MATHEMATICA
MATHEMATICA manuals, 91
MATHEMATICA notebooks, 134–135
MATHEMATICA options, 94
AccuracyGoal, 247, 297, 333
AnimationRate, 133
AspectRatio, 128, 129
Assumptions, 94, 112, 267, 269, 270
AxesLabel, 127, 128, 131, 204, 206, 246, 272, 277, 318, 334, 337
AxesOrigin, 337
Cartesian, 119–121
ColorFunction, 129
Contours, 129
ContourShading, 129
Cylindrical, 119–121
DisplayFunction, 139
Epilog, 127
FontFamily, 127, 128, 204, 206
FontSize, 127, 128, 204, 206
FontWeight, 127, 128, 204, 206
FormatType, 131
GaussPoints, 297
GridLines, 299, 300, 337
MaxIterations, 333
MaxRecursion, 297
MaxSteps, 247
MaxStepSize, 247
Method, 297, 298
PlotJoined, 299, 300, 334, 337
PlotLabel, 127, 128, 204, 206, 299, 300
PlotPoints, 127–129
PlotRange, 94, 127, 128, 133, 204, 206, 272, 277, 300, 318, 337
PlotStyle, 127, 128, 139, 204, 206, 246, 272, 277, 299, 300, 318, 334, 337
PrecisionGoal, 247, 297
Spherical, 119–121
TargetFunctions, 94, 106–107
TextStyle, 94
Thickness, 139
WorkingPrecision, 247, 297, 333
MATHEMATICA packages, 121–122
LinearAlgebra
MatrixManipulation, 146
MATHEMATICA special symbols
! (factorial), 102
′ (differentiation), 115, 118, 201–204, 207, 208, 212
INDEX

(*...*) (comments), 125
* (multiplication), 94, 95, 97, 102, 110
+ (addition), 94, 95, 101
- (subtraction), 94, 95, 102
. (clear variable), 98
. (matrix multiplication), 110
/ (division), 94, 95, 101
// (convert form), 100, 102, 273, 316
:= (function definition), 94, 97–98, 113, 276, 336
= (assignment), 97, 99
? (help), 93, 97, 98
?? (more help), 93
[[...]] (element of list), 99, 100, 108, 116, 117, 201, 203, 206, 210, 212, 213, 266, 299, 318, 319, 335, 337
#1 (generic variable), 315
% (previous output), 100, 103, 105, 107–110, 112, 114, 213, 269, 273, 275, 317, 319, 334
%% (second previous output), 103, 112
%%% (third previous output), 103
%n (output on line n), 96, 103, 213
_ (pattern), 97–98, 113, 276, 299
>> (direct output to file), 132
>>> (append output to file), 132
^ (exponentiation), 95, 97, 98, 102, 103
Out[n] (output on line n), 96, 103
MATHEMATICA statements
termination of, 92
MATHEMATICA system variables, 94
$BaseDirectory, 135
$HomeDirectory, 94
$IgnoreEOF, 94
$MachinePrecision, 247, 297
$MessageGroups, 136
$Packages, 94, 121
$Path, 94, 137
$RecursionLimit, 94
$TextStyle, 94
$TopDirectory, 94
$UserBaseDirectory, 135
mathematical functions

in IDL, see also IDL functions, IDL procedures, IDL programs, 23
in MATHEMATICA, see MATHEMATICA functions

mathematical symbols

in IDL, 68–69
MATLAB, 1
MATLAB contact information, 504
matrices

in IDL, 20–25
in MATHEMATICA, 99, 100, 110–111
matrix multiplication

in IDL, 24
in MATHEMATICA, 110
max-min problems, 308
MAXIMA, 1
MAXIMA contact information, 504
Maxwell-Boltzmann distribution, 304
method of bisection, see root finding
method of Frobenius, 211–213
method of least squares, see least squares fitting
midpoint method, 219
midpoint rule, see quadrature
MiKTeX, 1
MiKTeX contact information, 504
ML (left mouse button), 9
MM (middle mouse button), 9
moment of inertia

definition of, 259
defined at, 302
of semicircular plate, 259, 268, 291
motion

of membrane, 66–68, 88
of string, 88
MR (right mouse button), 9
MUDPACK, 1
MUDPACK contact information, 504
Müller’s method, see root finding
multi-way selection

in pseudocode, 157
multistep methods

for ODEs, 222–223

N

NAG library contact information, 504
named common

in IDL, see common storage
natural frequencies

of spherical cavity, 346
of stiff bar, 345, 346
of three-mass system, 345
of two-mass system, 309–310, 317–318, 325–326
Navier-Stokes’ equations, 361
Newton’s method, see root finding
Newton’s second law, 257, 258, 309, 311
in fluids, 363–365
Newtonian fluid, 365
INDEX

node numbers
  global, 376
  local, 376
nodes, 353
non-viscous fluid flow, 365
normal distribution, see Gaussian distribution
normal modes
  of membrane, 48–54, 85
  of string, 88
normal stress
  in fluids, 363–365
numerical approximation
  for first derivative, 368
  for second derivative, 369
numerical integration, see quadrature
Numerical Recipes, 1, 290, 291
NUMERICAL RECIPES contact information, 505

O

object code, 152
object-image relations, 83
octal number system, 5
OCTAVE, 1
OCTAVE contact information, 505
on-line help
  in IDL, 34–35
  in MATHEMATICA, 93
ordinary differential equations
  with IDL, 223–245
  with MATHEMATICA, 115–119, 200–213, 246–247
orthogonal functions, 264–265
OzTEX contact information, 505

P

page previewer
  in I\TeX, 486–487
partial fractions, 104, 148, 248
passwords, see Local Guide
path (in directory tree), 8
path specification
  in IDL, 76
  in MATHEMATICA, 136
PDEs
  classification of, 360–361
PDF files, 459–460
pdfcrop, 482–483
pdflatex, 459, 476, 481–482
pdftops, 483
pendulum, simple, see simple pendulum
pivotal condensation, 340
pivoting, 339
placing text in graphics
  in IDL, see IDL commands, xouts

Q

Q of resonant circuit, 85
quadratic formula, 108, 314, 315, 343, 345
quadrature, 278–286
  adaptive algorithms, 283
  assessing error in, 281–283, 286–289
  Gauss-Kronrod, 298
  Gauss-Legendre, 286
  Gaussian, 283–286, 288–289, 304
  iterative algorithms, 283
  midpoint rule, 148, 278
INDEX 521

Newton-Cotes, 278–283
recursive trapezoidal rule, 280
Romberg integration, 281
roundoff error, see roundoff error
Simpson’s rule, 148, 279, 281, 285, 288, 303
trapezoidal rule, 278, 287
truncation error, see truncation error
with IDL, 286–296
with MATHEMATICA, 296–300
quality factor, see Q of resonant circuit
quantum energy levels, see energy levels
quantum probabilities, 264
in hydrogen atom, 56, 86, 169–170, 179, 301, 305
of harmonic oscillator, 264, 291, 298, 304
quantum scattering, 143
quantum wave function, see wave function

R
radius of gyration
definition of, 259
of cardioid, 302
of semicircular plate, 268
railroad problem, 347
range of projectiles, see trajectories
Rayleigh criterion, 347
reaction kinetics, 254
reading data
from files
in IDL, 28–31, 33–35, 44, 66, 175
in pseudocode, 167
from keyboard
in IDL, 78–79
in pseudocode, 154
recasting
algorithms, 176
control structures, 176
recursion relationship, 211
reflection at dielectric film, 83
relativity
mass/momentum/energy, 83
motion under constant force, 266–267
relaxation, 163–164
residual, 323
resonance
in LRC circuit, 85, 194, 250
restoring initial state
in IDL, 78
in MATHEMATICA, 101
returning to command level
in IDL, 16
in MATHEMATICA, 92
returning to operating system
from IDL, 16
from MATHEMATICA, 93
ReVTeX, 491
Richardson extrapolation, 280, 288
Ritz variational method, 375
Rodrigues’ formula, 142
Romberg integration, see quadrature
root finding
accuracy in, 322–323
by bisection, 320, 324
by Laguerre’s method, 322, 323
by Müller’s method, 322, 323
by Newton’s method, 321, 324, 332, 342–343, 345
by secant method, 322, 333
by trial and error, 342
in IDL, 323–331, 343
in MATHEMATICA, 315–320, 331–337, 343
roundoff error, 370–371, 374, 396
in numerical quadrature, 281
in ODEs, 220
rubber lengths in LATEX, 461
Ruchardt’s experiment, 351
Runge-Kutta methods, 218–220
running program, see Local Guide

S
saving files, see writing data to files
scalar fields
three-dimensional, 169–170
in IDL, 170
two-dimensional, 168–169
in IDL, 168–169
Schrödinger equation, 198–199, 211, 240–244, 313, 329–331, 360
search path
in IDL, 33, 35–37, 77
in MATHEMATICA, 137
secant method, see root finding
self-adjoint ODE, 367
separating statements on one line
in IDL, 16, 32
in MATHEMATICA, 92
in pseudocode, 156
series solutions to ODEs, 211–213
shape functions, 375, 377
shear stress
in fluids, 363–365
shooting method, 237
sigmoid curve, 192, 205
simple pendulum, 252, 260, 274–275, 305
Simpson’s rule, see quadrature
simultaneous equations
by direct methods, 340
by iterative methods, 340
linear, 338–341
nonlinear, 341–343
singular equations, 338
singular value decomposition, 340
solving Laplace’s equation, see Laplace’s equation
INDEX

solving linear equations, 338–341
  in IDL, 25–26
  in MATHEMATICA, 109
solving ODEs
  with IDL, 223–245
  with MATHEMATICA, 200–213, 246–247
sound waves in fluids, 366–367
source code, 152
spawning to operating system
  from IDL, 16
special symbols
  in IDL, see IDL special symbols
  in MATHEMATICA, see MATHEMATICA
    special symbols
speed of sound, 366
spherical coordinates
  in MATHEMATICA, 150
spurious roots, 307, 313, 330, 336
stable equilibrium, 308
standing waves
  in cube, 86, 180
  in string, 197–198, 208–211, 237–240, 251
Stark effect, 81, 146, 345
startup files, see initialization files
structure of data files, 164–165
subroutines
  in IDL, see IDL functions, IDL procedures
  in MATHEMATICA, see MATHEMATICA
    functions
suppressing evaluation
  in MATHEMATICA, 113–114
suppressing output
  in IDL, 20
  in MATHEMATICA, 92
T
  tables
    in \LaTeX, 470–472
Taylor series
  for solving ODEs, 218–221
  for solving simultaneous equations, 343
  in MATHEMATICA, 112, 142, 148
termination of statements
  in IDL, 16
  in MATHEMATICA, 92
  in pseudocode, 156
\LaTeX, 455, 463
\LaTeX contact information, 455, 505
\LaTeX output
  from MATHEMATICA, see \LaTeX output
texdoc, 475, 496
\TeXlive, 1
\TeXlive contact information, 505
text editors, see Local Guide
text in graphics, see placing text in graphics
TGIF, 1
TGIF contact information, 506
thermal equilibrium, 85
thin lens equation, 83
tikz, 475–478
tikz commands
  \definecolor, 476
  \draw, 476–477
  \foreach, 476, 477
  \node, 476, 477
  \x, 476, 477
trajectories, 257
double pendulum, 249
earth falling into sun, 301, 305
exponentially decaying force, 301
in central forces, 195–197, 234–237, 252, 253
in constant magnetic field, 69, 138, 178
in Coulomb field, 250
in crossed $\mathbf{E}$ and $\mathbf{B}$, 253
in viscous medium, 188–191, 201–202, 250,
simple harmonic oscillator, 300
under constant force, 266–267, 300
transformation rules
  in MATHEMATICA, 104, 108, 200, 210
transmission at dielectric film, 83
transpose
  in IDL, 22, 24–26
  in MATHEMATICA, 110, 125
trapezoidal rule, see quadrature
tree-structured directories, 8
tridiagonal systems, 340, 350
truncation error
  global, 221
  in midpoint formula, 282
  in numerical quadrature, 281
  in ODEs, 220, 224–225
  in Simpson’s rule, 282
  in trapezoidal rule, 282, 287
    per step, 221
truth table, 155
turning points, 264, 292, 304, 308, 316, 324–325,
  333, 347
two’s complement notation, 6
two-way selection
  in IDL, 32–33
  in MATHEMATICA, 123–124
  in pseudocode, 157
U
UNIX, 1
UNIX commands
  aspell, 487–488
cleanbak, 496
cleanTEX, 496
dvipdfm, 460
dvips, 459
ispell, 487–488
latex, 458, 459
lp, 460
INDEX

ls, 483
pdfcrop, 482–483
pdflatex, 459
pdftops, 483
ps2pdf, 460
texdoc, 475
xdvi, 460
UNIX shell scripts
  ceps2pdf, 483, 501
cleanbak, 496
cleanTEX, 496
cps2pdf, 483, 501
rdfileeps, 483, 501
rdfileps, 483, 501
UNIX tutorial, see Local Guide
unstable equilibrium, 308
user accounts, see Local Guide
user-defined commands, see command files
user-defined functions
  in IDL, see IDL functions
  in MATHEMATICA, see MATHEMATICA functions
user-defined procedures
  in IDL, see IDL procedures
V
Van der Pol oscillator, 252
variable names
  in IDL, 18
  in MATHEMATICA, 94
  in pseudocode, 152
vector arrays
  three-dimensional, 173–174
  two-dimensional, 171–173
vector calculus
  with MATHEMATICA, 119–121
vector fields
  three-dimensional
    in IDL, 174
  two-dimensional
    in IDL, 172–173
vector identities, 150
velocity field
  of rotating object, 180
vertices, 353
vim, 1
viscosity, 365
viscous damping, 189, 192
W
wave equation
  in 1D, 355–357
  in IDL, 397–402
  in 2D, 389–391
  in 3D, 357–358
wave function, 198–199, 240–244, 264, 302, 304
weighting functions, 378
Wheatstone bridge, 143
Windows, 1
Windows batch files
  ceps2pdf.bat, 483, 500
  cleanbak.bat, 496
  cleanTEX.bat, 496
  cps2pdf.bat, 483, 500
  rdfileeps.bat, 483, 500
  rdfileps.bat, 483, 500
Windows commands
cleanbak, 496
cleanTEX, 496
dir, 483
dvipdfm, 460
dvips, 459
latex, 458, 459
pdflatex, 459
pdftops, 483
print, 460
ps2pdf, 460
texdoc, 475
yap, 460
Winedt, 1
Winedt contact information, 506
writing data to files, 177–180
  in IDL, 28–31, 168–170, 172–174
  in pseudocode, 164–167
writing data to screen
  in IDL, 20–22
  in pseudocode, 154
X
xdvi, 460, 486–487
Xemacs, 1
Xemacs contact information, 506
Xv contact information, 506
Y
yap, 460, 486–487