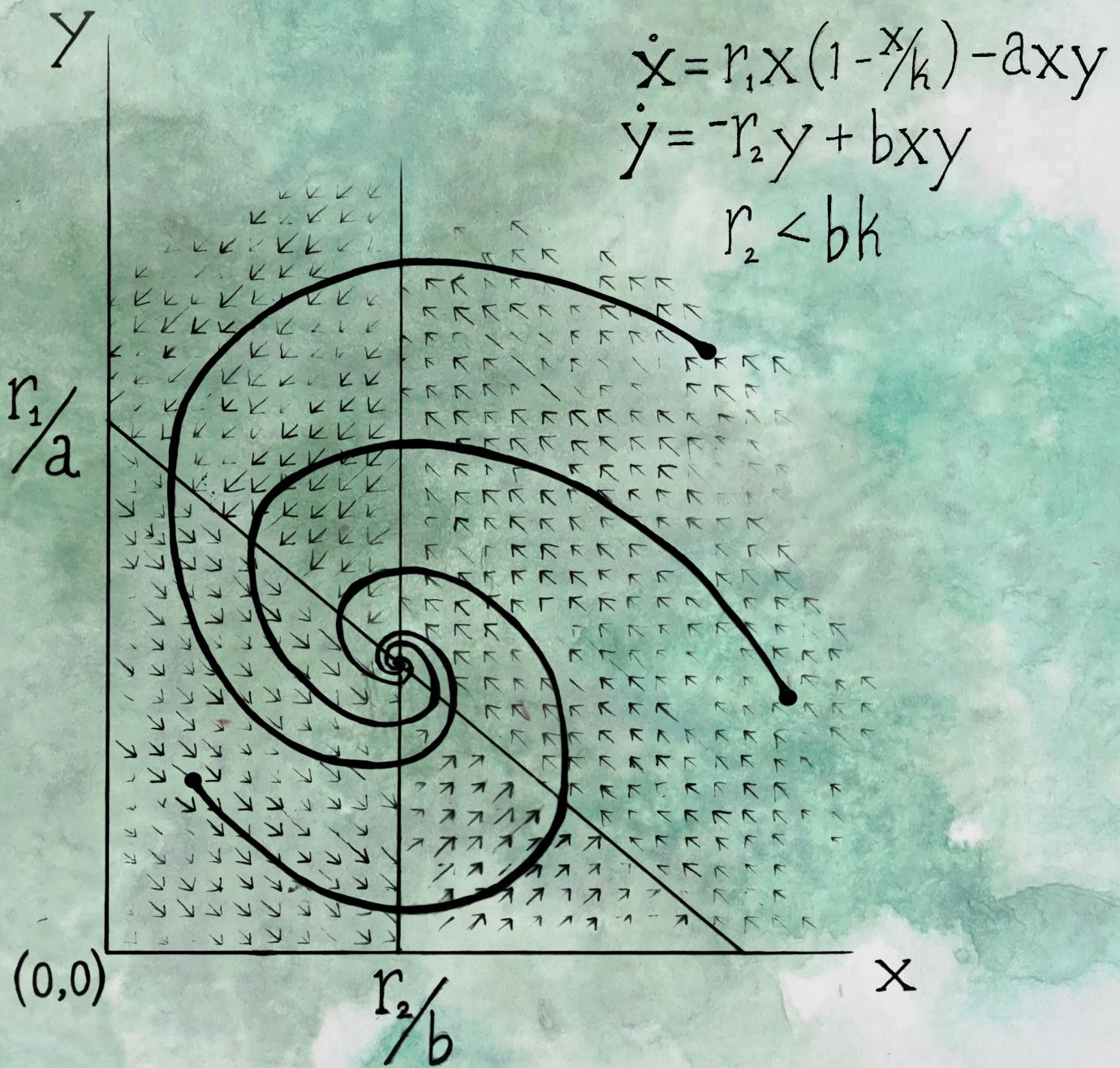


Differential Equations

A Toolbox for Modeling the World

Kurt Bryan



DIFFERENTIAL EQUATIONS
A TOOLBOX FOR MODELING THE WORLD
Version 1.11

Kurt Bryan

Department of Mathematics
Rose-Hulman Institute of Technology

Version 1.11, published July, 2022.

Copyright © 2022 Kurt Bryan

PUBLISHED BY SIMIODE, CHARDON OH USA

SIMIODE.ORG

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, excepted as permitted by law.

The right of Kurt Bryan to be identified as the author of this work has been asserted in accordance with law.

The publisher and author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of fitness for a particular purpose. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice herein may not be suitable for every situation. The fact that an organization or website is referred to in this work as a citation or potential source of further information does not mean that the author or the publisher endorses the information the organization or website may provide or recommendations it may make. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read.

This text was produced using the Legrand Orange Book template (LaTeX Template Version 2.4, (26/09/2018), with modifications) by Mathias Legrand (legrand.mathias@gmail.com), in accordance with the license: CC BY-NC-SA 3.0 (<http://creativecommons.org/licenses/by-nc-sa/3.0/>)

ISBN: 978-1-63877-937-7

This work was supported in part by the National Science Foundation through NSF:DUE-IUSE Grant # 1940532

Cover design and artwork by Ayla Walter, <http://www.aylawalter.com/>.

First printing, May 2021. Version 1.11 printed July, 2022.

For Frances.

Contents

Foreword	i
Preface	iii
1 Why Study Differential Equations?	1
1.1 The 2008 Olympic 100-Meter Dash	1
1.1.1 Usain Bolt's Olympic Victory	1
1.1.2 Modeling a Sprint	2
1.1.3 The Hill-Keller Differential Equation	3
1.2 Intracochlear Drug Delivery	5
1.2.1 The Challenge of Hearing Loss	5
1.2.2 A Compartmental Model for the Cochlea	6
1.2.3 The Differential Equation	7
1.3 Population Growth and Fishery Management	8
1.3.1 The Need to Manage Fish Harvesting	8
1.3.2 Modeling Fish Population	9
1.3.3 Modeling Harvesting	11
1.3.4 Parameter Estimation and Harvesting	11
1.4 Where Do We Go from Here?	12
1.4.1 A Toolbox for Describing the World	12
1.4.2 Some Terminology	13
1.4.3 You Already Know How to Solve Some Differential Equations	14
1.4.4 Exercises	16
1.5 The Blessing of Dimensionality	16
1.5.1 Definition of Dimension	16
1.5.2 The Algebra of Dimension	17
1.5.3 Derivatives, Integrals, Elementary Functions	18

1.5.4	Unit-Free Equations and Bending the Rules	19
1.5.5	Using Dimension to Find Plausible Models	20
1.5.6	Other Dimensions	20
1.5.7	Exercises	21
1.6	Modeling Projects	22
1.6.1	Project: Hang Time	22
1.6.2	Project: Money Matters	23
1.6.3	Project: Ant Tunneling	24
2	First-Order Equations	27
2.1	First-Order Linear Equations	27
2.1.1	Example: Solving the Hill-Keller Equation as a Linear ODE	28
2.1.2	A General Procedure for Solving Linear ODEs	30
2.1.3	Some Common First-Order Linear Models	31
2.1.4	Exercises	35
2.2	Separable Equations	38
2.2.1	Application: Falling Objects	39
2.2.2	Separation of Variables: A First Example	40
2.2.3	The General Procedure for Separation of Variables	41
2.2.4	Example: Solving the Falling Object ODE	42
2.2.5	Example: Solving the Logistic Equation	44
2.2.6	Exercises	45
2.3	Qualitative and Graphical Insights	48
2.3.1	Direction Fields	48
2.3.2	Autonomous Equations	50
2.3.3	Phase Portraits	51
2.3.4	Fixed Points and Stability	54
2.3.5	Determining the Stability of Fixed Points	55
2.3.6	Bifurcations	57
2.3.7	Exercises	60
2.4	The Existence and Uniqueness of Solutions	61
2.4.1	Some Inspiration from Calculus 1	61
2.4.2	What Are Solutions to ODEs?	62
2.4.3	The Existence-Uniqueness Theorem for ODEs	64
2.4.4	Exercises	66
2.5	Modeling Projects	67
2.5.1	Project: Money Matters 2	67
2.5.2	Project: Chemical Kinetics	70
2.5.3	Project: A Shot in the Water	74
3	Numerical Methods for ODEs	77
3.1	The Need for Numerics	77
3.1.1	Logistic Example: Time-Varying Parameters	77
3.1.2	Euler's Method	78
3.1.3	Evaluate, Extrapolate, Repeat as Necessary	79
3.1.4	The Accuracy of Euler's Method	81
3.1.5	Exercises	84

3.2	Improvements to Euler's Method	86
3.2.1	Improving Euler's Method	86
3.2.2	The Improved Euler Method	88
3.2.3	Exercises	90
3.3	Modern Numerical Methods	91
3.3.1	The RK4 Algorithm	92
3.3.2	Adaptive Step Sizing and Error Control	93
3.3.3	Exercises	99
3.4	Parameter Estimation	100
3.4.1	Hill-Keller Revisited	100
3.4.2	Least-Squares Estimation	102
3.4.3	Hill-Keller Again	104
3.4.4	Least Squares For ODE Parameter Estimation	107
3.4.5	A Cautionary Example	109
3.4.6	Exercises	110
3.5	Modeling Projects	116
3.5.1	Project: Sublimation of Carbon Dioxide	116
3.5.2	Project: Fish Harvesting Revisited	118
3.5.3	Project: The Mathematics of Marriage	120
3.5.4	Project: Shuttlecocks and the Akaike Information Criterion	124
4	Second-Order Equations	129
4.1	Vibration and the Harmonic Oscillator	129
4.1.1	The 2010 Chilean Earthquake	129
4.1.2	The Harmonic Oscillator	130
4.1.3	Initial Conditions	132
4.1.4	More Applications of Spring-Mass Models	132
4.1.5	Exercises	136
4.2	The Harmonic Oscillator	139
4.2.1	Solving the Harmonic Oscillator ODE: Examples	139
4.2.2	Solving Second-Order Linear ODEs: The General Case	142
4.2.3	The Underdamped and Undamped Cases	145
4.2.4	The General Underdamped Case	148
4.2.5	The Critically Damped Case	150
4.2.6	The Existence and Uniqueness of Solutions	152
4.2.7	Summary and a Physical Perspective	153
4.2.8	Exercises	153
4.3	The Forced Harmonic Oscillator	159
4.3.1	Solving the Forced Harmonic Oscillator Equation	161
4.3.2	Finding a Particular Solution: Undetermined Coefficients	163
4.3.3	When the Guess Fails	169
4.3.4	Exercises	171
4.4	Resonance	174
4.4.1	An Example of Resonance	174
4.4.2	Periodic Forcing	175
4.4.3	Exercises	186

4.5 Scaling and Nondimensionalization for ODEs	189
4.5.1 Motivation: Nonlinear Springs	189
4.5.2 Characteristic Variable Scales	190
4.5.3 Nondimensionalization: Logistic Equation Example	193
4.5.4 Nondimensionalization: Harvested Logistic Equation Example	195
4.5.5 The General Outline for Nondimensional Rescaling	197
4.5.6 Back to the Hard Spring	198
4.5.7 Exercises	201
4.6 Modeling Projects	205
4.6.1 Project: Earthquake Modeling	205
4.6.2 Project: Stay Tuned—RLC Circuits and Radios	207
4.6.3 Project: Parameter Estimation with Second-Order ODEs	208
4.6.4 Project: Bike Shock Absorber	210
4.6.5 Project: The Pendulum	211
4.6.6 Project: The Pendulum 2	213
5 The Laplace Transform	217
5.1 Discontinuous Forcing Functions	217
5.1.1 Motivation: Pharmacokinetics	217
5.1.2 Complication: Discontinuous Forcing	218
5.1.3 Complication: Impulsive Forcing	219
5.1.4 Discontinuous Forcing and Transform Methods	220
5.1.5 Exercises	220
5.2 The Laplace Transform	222
5.2.1 Definition of the Laplace Transform	222
5.2.2 What Kinds of Functions Can Be Transformed?	224
5.2.3 Laplace Transforms of Elementary Functions	225
5.2.4 Solving Differential Equations Using Laplace Transforms	228
5.2.5 The First Shifting Theorem	232
5.2.6 The Inverse Laplace Transform	233
5.2.7 The Initial and Final Value Theorems	236
5.2.8 Section Summary and Remarks	238
5.2.9 Exercises	238
5.3 Nonhomogeneous Problems and Discontinuous Forcing Functions	242
5.3.1 Some Nonhomogeneous Examples	242
5.3.2 Discontinuous Forcing	243
5.3.3 The Second Shifting Theorem	245
5.3.4 Some More Models and Examples	249
5.3.5 Summary and Remarks	252
5.3.6 Exercises	252
5.4 The Dirac Delta Function	255
5.4.1 Motivational Examples	255
5.4.2 Definition of the Dirac Delta Function	258
5.4.3 Three Models: Money, Masses, and Medication	263
5.4.4 The Laplace Transform of the Dirac Delta Function	264
5.4.5 Solving ODEs with Dirac Delta Functions	264
5.4.6 Summary and a Few Remarks	267
5.4.7 Laplace Transform Table	267
5.4.8 Exercises	267

5.5	Input-Output, Transfer Functions, and Convolution	270
5.5.1	A System Identification Problem	270
5.5.2	Input-Output Systems	270
5.5.3	Convolution	272
5.5.4	The Impulse Response and Convolution	276
5.5.5	System Identification with Impulsive Input	278
5.5.6	Exercises	280
5.6	A Taste of Control Theory	283
5.6.1	The Need for Control	283
5.6.2	Modeling an Incubator	283
5.6.3	Open-Loop Control	285
5.6.4	Closed-Loop Control	288
5.6.5	Proportional-Integral Control	292
5.6.6	Proportional-Integral-Derivative Control	294
5.6.7	Disturbances	295
5.6.8	Summary and Comments	298
5.6.9	Exercises	298
5.7	Modeling Projects	300
5.7.1	Project: Drug Dosage	300
5.7.2	Project: Machine Replacement	301
5.7.3	Project: Vibration Isolation Table Shakedown	304
5.7.4	Project: Segway Scooters and The Inverted Pendulum	306
6	Linear Systems of Differential Equations	311
6.1	Systems of Differential Equations	311
6.1.1	Motivation: More Pharmacokinetics	311
6.1.2	Existence and Uniqueness	316
6.1.3	Exercises	317
6.2	Linear Constant-Coefficient Homogeneous Systems of ODEs	320
6.2.1	Matrix-Vector Formulation	320
6.2.2	Solving the Homogeneous Case	320
6.2.3	Complex Eigenvalues	324
6.2.4	Defective Matrices	327
6.2.5	Exercises	330
6.3	Linear Constant-Coefficient Nonhomogeneous Systems of ODEs	332
6.3.1	Solving Linear Systems of ODEs with Laplace Transforms	332
6.3.2	Undetermined Coefficients for Systems of ODEs	334
6.3.3	The Significance of Eigenvalues	338
6.3.4	Exercises	338
6.4	The Matrix Exponential	340
6.4.1	Inspiration	341
6.4.2	Definition of the Matrix Exponential	341
6.4.3	Properties of the Matrix Exponential	343
6.4.4	Solving ODEs with the Matrix Exponential	343
6.4.5	Computing The Matrix Exponential: The Diagonal Case	346
6.4.6	Computing The Matrix Exponential: The Diagonalizable Case	347
6.4.7	Computing The Matrix Exponential: Putzer's Algorithm	348
6.4.8	Final Remarks	351
6.4.9	Exercises	351

6.5 Modeling Projects	353
6.5.1 Project: LSD Compartment Model	353
6.5.2 Project: Homelessness	354
6.5.3 Project: Tuned Mass Dampers	356
7 Nonlinear Systems of Differential Equations	361
7.1 Autonomous Nonlinear Systems and Direction Fields	361
7.1.1 Some Nonlinear ODE Models	362
7.1.2 Direction Fields	365
7.1.3 A Nonlinear Direction Field Example	367
7.1.4 Direction Fields in Higher Dimensions	369
7.1.5 Exercises	370
7.2 Direction Fields and Phase Portraits for Linear Systems	371
7.2.1 Direction Fields for Homogeneous Linear Systems	371
7.2.2 Application to the LSD Model	376
7.2.3 The Equation $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}$	379
7.2.4 Direction Fields for Larger Systems of ODEs	379
7.2.5 Exercises	380
7.3 Autonomous Nonlinear Systems and Phase Portraits	382
7.3.1 Sketching Phase Portraits for Nonlinear Systems	382
7.3.2 Linearizing ODEs at Equilibrium Points	387
7.3.3 Exercises	392
7.4 Analyzing Systems with Unspecified Parameters	394
7.4.1 Sketching Phase Portraits with Unspecified Parameters	395
7.4.2 Linearizing the Competing Species Model with General Parameters	397
7.4.3 Conclusions for Competing Species	400
7.4.4 Higher-Dimensional Systems	400
7.4.5 Exercises	401
7.5 Numerical Methods for Systems of First Order ODE's	404
7.5.1 Extending Basic Numerical Methods to Systems	404
7.5.2 Stiff Systems of ODEs	408
7.5.3 Implicit Numerical ODE Solvers	415
7.5.4 Exercises	418
7.6 Additional Techniques for Systems of First Order ODEs	424
7.6.1 First Integrals and Conservative Systems	424
7.6.2 Lyapunov Functions	428
7.6.3 Linearization and the Routh-Hurwitz Theorem	435
7.6.4 Exercises	438
7.7 Modeling Projects	444
7.7.1 Project: Homelessness Revisited	444
7.7.2 Project: Predator-Prey Model	445
7.7.3 Project: Parameter Estimation for Competing Yeast Species	447
8 An Introduction to Partial Differential Equations	449
8.1 Conservation of Stuff and the Continuity Equation	449
8.1.1 Industrial Furnaces and Metal Production	449
8.1.2 Conservation of Stuff	451
8.1.3 The Continuity Equation	452

8.1.4	The Heat Equation	454
8.1.5	Some Solutions to the Heat Equation: Separation of Variables and Linearity	458
8.1.6	Exercises	461
8.2	Fourier Series	464
8.2.1	An Example	464
8.2.2	Approximating Functions	465
8.2.3	The Fourier Cosine Expansion	467
8.2.4	The Fourier Sine Expansion	473
8.2.5	More on Fourier Series Convergence	474
8.2.6	Exercises	478
8.3	Solving the Heat Equation	483
8.3.1	Homogeneous Dirichlet Conditions	483
8.3.2	Insulating Boundary Conditions	484
8.3.3	Other Boundary Conditions	485
8.3.4	Diffusion	490
8.3.5	Solving the Nonhomogeneous Heat or Diffusion Equation	494
8.3.6	Exercises	497
8.4	The Advection and Wave Equations	501
8.4.1	The Advection Equation	501
8.4.2	Solution to the Advection Equation	502
8.4.3	The Wave Equation	504
8.4.4	Solution to the Wave Equation	506
8.4.5	The Wave Equation on the Real Line	510
8.4.6	Exercises	515
8.5	Modeling Projects	518
8.5.1	Project: It's a Blast (Furnace)!	518
8.5.2	Project: Finding Polluters	523
8.5.3	Project: Strung Out	525
8.5.4	Project: Frequency Analysis of Signals	529
8.5.5	Project: It's All Relative	535
A	Appendix Complex Numbers	541
A.1	Motivation and Definition	541
A.2	Arithmetic with Complex Numbers	542
A.3	Exponentiation of Complex Numbers	544
A.4	The Fundamental Theorem of Algebra	545
A.5	Partial Fraction Decompositions over the Complex Numbers	548
A.6	Additional Exercises	550
B	Appendix Matrix Algebra	551
B.1	Linear System of Equations	551
B.2	Matrix Algebra	554
B.3	Eigenvalues and Eigenvectors	559
B.4	The Eigenvalues for a General Two by Two Matrix	562
B.5	Diagonalization	564
B.6	Additional Exercises	566

C	Appendix Circuits	569
C.1	Current, Voltage, and Resistance	569
C.2	Capacitors	571
C.3	Inductors	574
C.4	RLC Circuits	576
C.5	Complex-Valued Solutions and Periodic Forcing	578
C.6	Impedance in Electrical Circuits	580
	Bibliography	583
	Index	591
	Back Cover	598



FOREWORD

It is with great pleasure that we introduce Dr. Kurt Bryans online text *Differential Equations: A Toolbox for Modeling the World*, to help you teach and learn differential equations in a rich modeling context. Kurt is a Full Professor of Mathematics at Rose-Hulman Institute of Technology, Terre Haute IN USA. Rose-Hulman is a “full caring” institution in which excellent scholarship and opportunities for student growth and learning are fostered by talented faculty who reach out and engage students in a supportive manner. Indeed, Rose-Hulman received the 2021 Award for an Exemplary Program for Achievement in a Mathematics Department by the American Mathematical Society. Professor Bryan has been a part of that exceptional Rose-Hulman team since he joined the faculty in 1993 (he is now an Emeritus Professor of Mathematics, having retired in 2022). That is where we first met, for I was senior member of the faculty at Rose-Hulman at the time.

When Kurt arrived on campus, he was given the usual first-year faculty teaching load. However, I approached him with the idea of co-teaching one of my sections (above his other load!) in which we would feature modeling to motivate learning. I proposed that we create modeling activities for students as we went along, just-in-time material production to supplement the text. We did this and had a rich and exhausting ten-week term experience with our students.

In our course-end evaluations it was universally clear that the students valued (and liked!) the modeling approach we offered them. We tried to have new models all the time, but for some classes we just could not produce a modeling setting, mostly through fatigue(!) We resorted to either lecture or small group work but proffered the usual technique-based approach to differential equations instruction. The students said essentially, We liked when you came to class with a model and were on with excitement and a sense of adventure. We knew when you had not prepared and were off. We preferred you being on and using modeling. It was at that point that we knew we were on to something and now, at a different point in our careers, we have joined together to bring you Kurts excellent text in which he is surely on by offering rich modeling throughout the text and in support of learning differential equations.

This text is part of the sustainability effort to maintain the Community of Practice at SIMIODE and is offered through a modest low-cost price of \$39US for this online version. All supporting materials at SIMIODE at www.simiode.org are Open Education Resources which are FREELY downloadable, fully modifiable, and customizable for educational use in the most generous Creative Commons license.

The production of this text is supported in part by the National Science Foundation through NSF:DUE: IUSE Grant # 1940532 with much appreciation from the SIMIODE community.

We are grateful for excellent copy editing and proofreading by Drs. Underwood Dudley and Sheila Miller, as well as an insightful set of comments from Dr. Glenn Ledder. Other readers of the preliminary version made suggestions and we are grateful to them as well. While we may not have

been able to incorporate all suggestions and recommendations in this edition, we have plans to enrich the text with many of these suggestions for our second edition. It is wonderful to work with a talented team and Kurt has been the lead driver and creator. We appreciate his boundless energy and enthusiasm in this project. We know you will like what he offers and be very comfortable in his style. Enjoy learning the mathematics of differential equations in context through modeling!

Dr. Brian Winkel, Director SIMIODE

PREFACE

Motivation

This book is a distillation of my 29 years of experience teaching introductory courses in ordinary differential equations (ODEs) to STEM majors at the Rose-Hulman Institute of Technology. My approach to teaching this material was strongly influenced by Brian Winkel, who took me under his wing during my first year at Rose. The very first class I was slated to teach in the fall was differential equations, to a group of sophomore engineers. Brian was teaching the same class and offered me the “opportunity” to co-teach his section with him. I’d planned a first lesson that involved the usual definitions and solution techniques. Brian saw what I had in mind and said “Let me show you how I do it . . . ”

The approach he showed me involves introducing a practical problem that students can understand and relate to, but cannot solve. This motivates and drives the mathematics we develop in class. The applications and models throughout the course become a scaffold for what we’re learning, touchstones that we return to again and again as we develop more and more sophisticated techniques. In the end we can make conclusions about the original problem that we could not have made without the mathematics. The goal is not to replace ODEs with modeling, but to augment the material with modeling that motivates and illuminates the mathematics, and highlights the common mathematical structure of many physical situations.

Outline

The order of topics is fairly standard: first-order ODEs, numerical methods, second-order ODES, the Laplace transform, the linear and nonlinear systems of ODES. However, in each chapter we kick off the study of new topics with one or more models to which we can apply the mathematics we are learning.

There’s more material in the book than can be done in a semester, though, as I have included some topics that are not usually taught in an introductory ODE course. This includes elementary dimensional analysis, scaling and nondimensionalizing ODEs, a bit more on modern numerical ODE solvers, parameter estimation, applications of the Laplace transform to control theory, the matrix exponential, and more qualitative analysis for nonlinear systems of ODEs than is usually done. Many of these additional topics I found useful when working in industry and government labs. However, I placed this material at the end of each chapter, in such a manner that it can be omitted without disrupting the flow of the text. But I do hope that some of it can be included in your course, as time and student interests permit. Much of this material, along with the associated Modeling Projects, would be great for independent study and student projects outside of the classroom. I have also included brief appendices that cover the essential facets of complex numbers and matrix algebra, and an appendix that does a bit more with circuits than would normally be covered in an ODE textbook.

Exercises, Activities, and Technology

I’ve sprinkled over 220 inline Reading Exercises throughout the text. These are short, straightforward exercises designed to keep the reader engaged. In some cases they help move the exposition forward, but are never essential to pursuing the material that comes after the exercise. There are also exercises at the end of each section, about 230 in all, ranging from routine computation and

solution techniques, to further analysis of the models presented in the text. Finally, at the end of each chapter there are three to six substantial Modeling Projects, a total of twenty-six such projects. Many of these are adaptations of projects available at the SIMIODE website, many are completely new for this text.

Solutions for all Reading Exercises and many of the section-end exercises are available to students at the book website [8], and a complete set of solutions is available to instructors. Solutions to the Modeling Projects are also available to instructors (though the Modeling Projects involve some creativity and flexibility, so there may not be “a” solution).

It’s inescapable that exercises and modeling projects that involve data or more sophisticated physical situations will require the use of technology. There is a selection of Maple, Mathematica, Matlab, and Sage code available at the book website at [8], to assist in this type of analysis. The data sets used in the text are available at the website too. I do not flag exercises and projects that require technology (many do not), but leave it up to the instructor’s or student’s judgement.

Acknowledgements

I would like to thank talented artist Ayla Walter (<http://www.aylawalter.com/>) for her beautiful cover design.

I would like to acknowledge and thank those who authored SIMIODE projects that I have adapted for this textbook, either as projects or examples in the exposition: Jue Wang, for the material on modeling intracochlear drug delivery; Wandi Ding, for material on modeling fisheries and fish harvesting; Karen Bliss, for material on modeling chemical kinetics and reaction rates; Erdi Karo and Tracy Weyand, for material on modeling certain sociological aspects of marriage; Sheila Miller, for material related to SIR disease models; and Mary Vanderschoot, for models related to the homelessness problem.

Finally, I would like to thank Brian Winkel, not only for his numerous contributions to this book, but his tireless promotion of modeling with ODEs and his mentorship, without which this book would not exist.

Kurt Bryan

1. Why Study Differential Equations?

To begin, we offer mathematical models of three quite different physical situations. Remarkably, all can be described by similar mathematics. These three examples and many others appear throughout the text and will help illustrate and motivate the mathematics to come.

1.1 The 2008 Olympic 100-Meter Dash

The material in this section is based on the SIMIODE project “Dash It All!” [32].

1.1.1 Usain Bolt’s Olympic Victory

Table 1.1 contains data from the 100-meter dash final at the Olympic Games in Beijing in 2008 [10]. The times belong to the gold medal winner Usain Bolt and represent a world record of 9.69 seconds, which he lowered in 2009 to 9.58 seconds. The data are in the form of (time, distance) pairs, where distance is measured in meters, horizontally along the track from the starting line, and time is measured in seconds elapsed from the firing of the starting gun. The initial (0.165, 0) data point indicates that Bolt had a reaction time of 0.165 seconds after the gun was fired before he started running and crossed the starting line. Between the 50 and 80 meter mark Bolt averaged 12.2 meters per second, an astonishing 27.3 miles per hour. After the 80 meter mark he actually eased up and looked back at the other runners; see [3].

Time (seconds)	0.165	1.85	2.87	3.78	4.65	5.50
Position (meters)	0	10	20	30	40	50

Time (seconds)	6.32	7.14	7.96	8.79	9.69
Position (meters)	60	70	80	90	100

Table 1.1: Race splits (seconds) every 10 meters for Usain Bolt’s 2008 Olympic gold medal final [10].

Our goals are to use this data to:

- (1) Develop a mathematical model of sprinting, a quantitative description that explains the data in Table 1.1. This description should be based on accepted physics, mathematics, and reasonable assumptions.
- (2) Test or validate this model by using it to make predictions about how fast Bolt or comparable sprinters could run other distances, and determine conditions under which the model is accurate.

Although we have data only for Bolt, we want our model to be more generally applicable. For the moment let us focus on sprinting, and assume that the runner applies maximum effort throughout the race.

1.1.2 Modeling a Sprint

We now consider a classic mathematical model for sprinting, the Hill-Keller model [65, 71]. The model is grounded in Newton's second law of motion, $F = ma$, where m denotes the mass of an object, a the acceleration of the object, and F the net force acting on the object. This is a fundamental law of physics, at least non-relativistic physics. (Bolt is fast, but not that fast.) In general, force and acceleration are three-dimensional vectors, but in our model they may be treated as scalars for reasons described below.

In order to explain the data in Table 1.1, we need to predict Bolt's position on the track as a function of time. We will thus introduce a variable t to denote time, in seconds, from the start of the race and x to denote position on the straight track, in meters, with $x = 0$ as the starting line and positive values of x in the direction of the finish line.

Mathematical models involve making assumptions and simplifications. In our case, matters are simplified by focusing only on the sprinter's horizontal motion along the track. Any other motion or forces, for example, vertical or side-to-side, will be ignored. As such, the sprinter's position, velocity, and acceleration are parallel to the track and may be considered scalar or one-dimensional quantities. We limit our attention to this component of motion. Our model will initially focus on the sprinter's velocity as a function of time, which can be described by some function $v(t)$, where v will be measured in meters per second. If $v(t)$ can be determined then we can then integrate $v(t)$ to predict the runner's position at any time, and in Bolt's case, compare this to the data.

Remark 1.1.1 In this text we will use various notations for the derivative of a function f . In most cases f will be a function of a single independent variable t , and t will denote time. We will write df/dt , f' , or \dot{f} ; this last notation is common in physics and is used only when the independent variable is time. In each case we may or may not explicitly write the independent variable, that is, we may write $f'(t)$ or just f' . Second derivatives are denoted by d^2f/dt^2 , f'' , or \ddot{f} .

We now apply Newton's second law of motion $F = ma$ to a sprinting race. Here m will be the sprinter's mass, which is not known and, happily, won't be needed. The variable a denotes the sprinter's acceleration, which will be a function of time, and F is the net force on the sprinter.

Reading Exercise 1.1.1 Express the sprinter's acceleration $a(t)$ in terms of velocity $v(t)$.

Reading Exercise 1.1.2 List in plain English all of the horizontal forces you can think of that might be relevant to the sprinter's progress down the track. Which forces aid progress down the track? Which impede progress?

The heart of the Hill-Keller model is an examination of the net horizontal force F on the runner, which is split into the sum of a propulsive force F_p that aids the runner and a resistive force F_r that impedes the runner's motion.

To quantify the propulsive force F_p , assume that F_p depends on the runner's level of exertion and is at a constant and maximum value throughout the race. That is, in a short race like a sprint,

the runner exerts a maximum propulsive force for the duration of the race and this force does not depend on the runner's current velocity. Moreover, in the Hill-Keller model this maximum propulsive force is treated as being proportional to the runner's mass m , under the assumption that if sprinter A is twice as massive as sprinter B then sprinter A should be capable of exerting twice the propulsive force of sprinter B. Under this assumption

$$F_p = mP, \quad (1.1)$$

where P is some constant that depends on the runner's maximum propulsive effort. Reading Exercise 1.1.3 gives some ideas on how we might interpret P physically.

Reading Exercise 1.1.3

- (a) Given that F_p is a force and m is a mass, argue that P should have units of acceleration. Hint: consider Newton's second law. (See Section 1.5 for a more extensive discussion of the value of understanding the units or physical dimensions involved when modeling.)
- (b) Suppose the runner is standing still, with no forces acting on the runner, who then suddenly applies maximum propulsive effort according to (1.1). Given that also $F = ma$, what physical interpretation can you give to P at this instant?

The quantity P has been measured for world-class sprinters like Bolt and is approximately 11.0 meters per second squared (see [98]). This is the value we'll use for P , at least for now. In doing so we are using meters, kilograms, and seconds for our analysis (SI units), so t will be measured in seconds and $v(t)$ in meters per second.

The resistive force F_r should, of course, oppose the runner's motion. In general we expect that the faster the runner moves, the greater the resistive force. Let's start with the simplest model that captures this idea: F_r should be proportional to the runner's velocity v , and in the opposite direction to v . Where do these resistive forces come from? In the Hill-Keller model these resistive forces are considered to be predominantly internal to the runner, a sort of friction of joints and muscles that opposes rapid motion, rather than external factors like air resistance. As such, like the propulsive force, the resistive force is modeled as proportional to the mass of the runner, under the reasoning that a runner who is twice as large has twice the internal resistance to motion. In summary, the resistive force F_r is proportional to both the runner's velocity v and the runner's mass m . That is, F_r is jointly proportional to v and m , and opposed to v . A simple model that captures this is

$$F_r = -kmv(t) \quad (1.2)$$

where k is a positive constant. The explicit minus sign on the right in (1.2) with the specification that $k > 0$ assures that F_r is opposed to v .

Modeling Tip 1.1.1 In mathematical modeling one encounters many physical constants, e.g., m and k as in (1.2). When a constant must be of one sign, positive or negative, it is common to take the constant as positive and explicitly add a minus sign if necessary. This helps to alert the reader that the constant in question is of one particular sign.

The value of k in (1.2) is not known, but can be estimated from data. For now think of k as a known but unspecified positive constant.

1.1.3 The Hill-Keller Differential Equation

We now have all the pieces necessary to construct a model that will (eventually) lead to an explanation of the data in Table 1.1, and provide broader insight into how a sprinter progresses down the track.

Reading Exercise 1.1.4 The total force F on the runner is $F = F_p + F_r$. Combine (1.1), (1.2), and $F = ma$ with the result of Reading Exercise 1.1.1 to show that the function $v(t)$ must satisfy

$$v'(t) = P - kv(t). \quad (1.3)$$

Note that m drops out.

The relation between $v'(t)$ and $v(t)$ in (1.3) must hold for all times during the race at which the runner is exerting maximum propulsive effort. The function $v(t)$ is the unknown we seek. Equation (1.3) is an **ordinary differential equation (ODE)**, that is, an equation involving an unknown function of one independent variable and that function's derivatives. The goal is to determine the unknown function $v(t)$ by using (1.3).

Reading Exercise 1.1.5 Based on your intuition, sketch what the graph of $v(t)$ would look like over the course of a race that lasts about 10 seconds.

Reading Exercise 1.1.6 Unfortunately, there are infinitely many solutions to (1.3). Verify that each of the following choices for $v(t)$ satisfies equation (1.3) (that is, $v'(t)$ is identically equal to $P - kv(t)$ as a function of t).

- (a) $v(t) = P/k$
- (b) $v(t) = P/k - Pe^{-kt}/k$
- (c) $v(t) = P/k - Ce^{-kt}$, where C is any constant.

In Reading Exercise 1.1.6 you may notice that parts (a) and (b) are special cases ($C = 0$ and $C = P/k$, respectively) of part (c). Indeed, the differential equation (1.3) has infinitely many solutions, all of the form in part (c) for some choice of the constant C . In this case $v(t) = P/k - Ce^{-kt}$ is called a **general solution** to the ODE (1.3). This means that all solutions to (1.3) can be expressed in the form $v(t) = P/k - Ce^{-kt}$ for some constant C . Given that there are infinitely many solutions, which one is relevant to the present case? It seems we need a bit more information, something that you may have noticed in Reading Exercise 1.1.5.

Reading Exercise 1.1.7 What piece of information is missing? Hint: look at the first few entries in Table 1.1. What was Bolt's velocity at the start of the race?

You should conclude from Reading Exercise 1.1.7 that $v(t)$ satisfies $v(0.165) = 0$. This is the **initial condition** for $v(t)$. After $t = 0.165$ the function $v(t)$ satisfies (1.3). Equation (1.3), together with the initial condition $v(0.165) = 0$ constitutes an **initial value problem**. It turns out that there is a unique (one, and only one) solution $v(t)$ to this initial value problem, and this is what we want to find. A proof of this fact is presented in more advanced differential equations texts, but further remarks will be made on this matter in Section 2.4. Once $v(t)$ is known we can integrate with respect to t to find Bolt's position as a function of time, and then adjust k to match this model to the data in Table 1.1. But it will first be helpful to develop some techniques for solving differential equations, and for determining the optimal value for k . Although we will initially use $P = 11$ meters per second squared, we may wish to adjust this value as well, to better fit the data in Table 1.1.

Reading Exercise 1.1.8 You might be tempted to state $v(0) = 0$ for the initial condition, which is certainly true since Bolt should not have been in motion when the gun was fired. But does (1.3) hold for all $t > 0$? Hint: What is P in (1.1) for $0 < t < 0.165$?

At this point we will leave the Hill-Keller model and return to it after mastering some techniques for analyzing and solving differential equations. The Hill-Keller model rests upon basic physics, specifically Newton's second law of motion. The next section illustrates another common modeling technique.

Reading Exercise 1.1.9 Review the derivation of (1.3) and list every assumption we made in constructing this model.

1.2 Intracochlear Drug Delivery

The material in this section is based on the SIMIODE project “Intracochlear Drug Delivery” [115].

1.2.1 The Challenge of Hearing Loss

Over 5% of the world’s population—or 466 million people—have disabling hearing loss [129]. The World Health Organization estimates that by 2050 over 900 million people, or about one in every ten people, will suffer from such hearing loss. Treating this hearing loss will be a significant challenge.

One aspect of this challenge is that the inner ear is surrounded by dense temporal bone and protected by the blood-cochlea barrier, as illustrated in Figure 1.1. The cochlea, with the shape of a snail, is the part of the inner ear involved in hearing. It is lined by sensory hair cells and is filled with fluid (about 0.2 milliliters, or 200 microliters). The cochlea is a particularly difficult target for drug therapy aimed at treating hearing loss. Oral medications and injections are typically blocked by the blood-cochlea barrier, and thus ineffective in reaching or delivering precise doses of drugs to the cochlea.

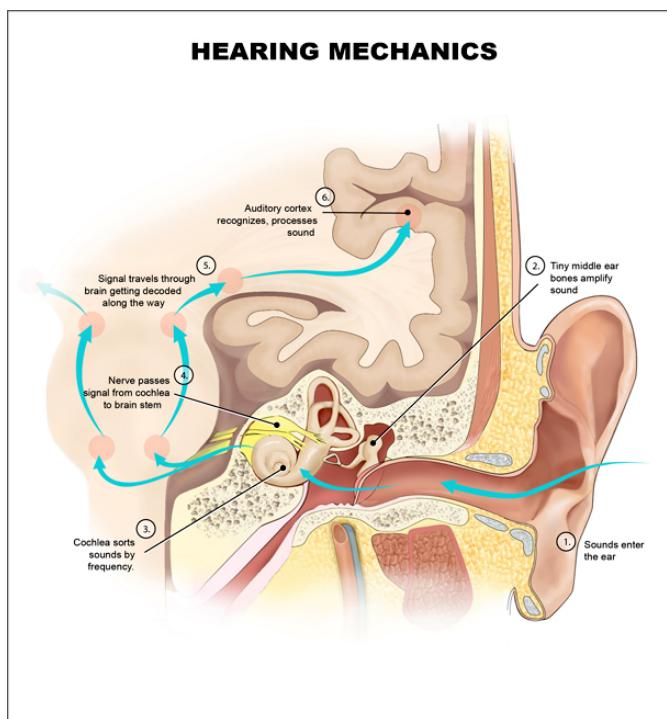


Figure 1.1: Anatomy of the human ear (from Wikimedia Commons, [40]).

As an alternative to systemic administration, localized drug delivery methods have emerged. One approach is the use of reciprocating perfusion systems based on microfluidic technologies [108]. This approach releases drugs directly to the inner ear, in order to support regeneration of the sensory hair cells and auditory nerves inside the cochlea, and enables precise targeting of drug concentrations within the therapeutic window for extended delivery. These implantable microfluidic devices are connected with a small tube to the cochlea. A battery-powered micropump pulses precise quantities of a drug from a small reservoir into the cochlea in a push-pull mode, i.e., infusing and withdrawing cochlear fluid in a cyclic manner nearly simultaneously so that the fluid volume inside the cochlea stays constant.

In order to avoid damage to hearing structures, limits on the maximum rate at which fluid can be pumped into the cochlea place stringent requirements on the system. It is challenging to design

reliable systems that are capable of maintaining control over drug concentrations for long-term drug release. To address the difficulties in drug delivery and achieve safety and efficiency, we need an effective quantitative model of the situation. In particular, we need to know how the concentration of the drug being administered varies over time inside the cochlea, and how the concentration depends on the physical parameters involved.

1.2.2 A Compartmental Model for the Cochlea

As a first approximation, consider Figure 1.2. This is an example of a **compartmental model**, a model which consists of one or more compartments with conduits into and out of each. The model in Figure 1.2 is a single-compartment model. Our task is to account for the rate at which some substance moves into and out of the compartment, and how the amount of the substance in the compartment changes with time. In this case the compartment represents the cochlea.

The input to the cochlea is through a single conduit in a push-pull or input-output operation. That is, a tiny amount of drug-containing fluid is introduced through the pipe into the cochlea; the drug then diffuses throughout the cochlea, and then a short time later the same amount of fluid is withdrawn.

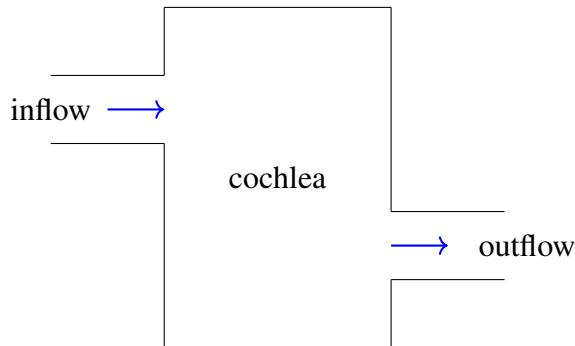


Figure 1.2: A compartmental model for cochlear drug delivery.

However, we will model the situation as if the drug-containing fluid is introduced through one input (the inflow pipe in Figure 1.2) and withdrawn through another (the outflow pipe) continually, at the same volumetric rate. The justification for this is as follows: we assume that when a tiny amount of drug-containing fluid is introduced into the cochlea, the drug diffuses rapidly and uniformly throughout the volume of the cochlea, so the concentration of the drug in the cochlea is always spatially constant. As a result, during the withdrawal phase of the push-pull cycle, the fluid removed has a constant concentration of the drug. The amount of fluid introduced and withdrawn during each cycle is very small, so the total volume of fluid in the cochlea remains essentially constant. On a sufficiently large time scale (say, several push-pull time cycles) the process may be viewed as the introduction of drug-containing fluid through one pipe; the drug is then considered to instantaneously diffuse to constant concentration, with fluid being withdrawn from another pipe at the same volumetric rate as the inflow. We also assume that the drug is not metabolized or otherwise destroyed (or created) in the cochlea.

Thus the only way into the cochlea for the drug is through the inflow pipe, the only way out is through the outflow pipe, and the drug is neither created nor destroyed in the cochlea. If during a time interval Δt an amount A_1 micrograms (μg) enters through the inflow pipe and an amount A_2 μg exits through the outflow pipe, then the amount of drug in the cochlea has changed (increased or decreased) by an amount $(A_1 - A_2)$ μg . Divide by Δt to obtain $(A_1 - A_2)/\Delta t$ as the average rate at which the amount of drug in the cochlea is changing during this time interval. The quantity $A_1/\Delta t$ is the average rate at which the drug enters through the inflow, and $-A_2/\Delta t$ is average rate the drug

leaves through the outflow. The equation $(A_1 - A_2)/\Delta t = A_1/\Delta t - A_2/\Delta t$ states that

The average rate of change of the amount of drug contained in the cochlea equals the average inflow rate minus the average outflow rate.

When $\Delta t \rightarrow 0$ the average rates of change become instantaneous rates of change, and we have

$$\begin{aligned} \text{instantaneous rate of change of drug in the cochlea} &= \text{instantaneous rate drug enters} \\ &\quad - \text{instantaneous rate drug exits}. \end{aligned} \tag{1.4}$$

Equation (1.4) is the basis of our mathematical model.

Reading Exercise 1.2.1 Let $u(t)$ denote the amount (in μg) of drug in the cochlea at time t , with t measured in minutes. What familiar mathematical quantity denotes the instantaneous rate of change of the amount of drug with respect to time? What units does this instantaneous rate of change have here?

Reading Exercise 1.2.1 quantifies the left side of (1.4). To quantify the right side of (1.4) we also need to know the instantaneous rate at which the drug is entering the cochlea through the inflow pipe, and the instantaneous rate at which the drug is leaving. Suppose that fluid is entering the cochlea through the inflow pipe in Figure 1.2 at a volumetric rate of r microliters per minute ($\mu\text{L}/\text{min}$). This fluid contains the drug at a constant concentration of c_1 micrograms per microliter ($\mu\text{g}/\mu\text{L}$).

Reading Exercise 1.2.2 At what rate is the drug entering the cochlea through the inflow pipe? Your answer should be in units of micrograms per minute ($\mu\text{g}/\text{min}$, physical dimensions of mass per unit time). Hint: the answer depends on r and c_1 .

Modeling Tip 1.2.1 Always keep track of the units or physical dimensions of the quantities of interest. They should always make sense, and in particular one should only ever have to add, subtract, or equate quantities with like dimension, e.g., add a mass and a mass. If you ever find yourself adding a mass and a meter, you messed up. In many equations there will also be dimensionless quantities, constants like π, e , or other real or complex numbers. The same logic applies to these quantities. You can only add or subtract a dimensionless quantity to another dimensionless quantity. This topic will be explored in more detail in Section 1.5.

Computing the rate at which the drug is leaving the cochlea is similar to computing the inflow rate. Suppose that at time t there are $u(t)$ μg of the drug in the cochlea. Suppose the volume of the cochlea is $V \mu\text{L}$. We have assumed that the drug is uniformly distributed throughout the cochlea, and hence will have a concentration of $\frac{u(t)}{V} \frac{\mu\text{g}}{\mu\text{L}}$ at any time. That is, each μL of fluid in the cochlea contains $\frac{u(t)}{V} \frac{\mu\text{g}}{\mu\text{L}} \times 1 (\mu\text{L}) = \frac{u(t)}{V} \mu\text{g}$ of the drug. Each minute $r \mu\text{L}$ of this fluid exits the cochlea.

Reading Exercise 1.2.3 At what rate is the drug exiting the cochlea through the outflow pipe? Your answer should be in units of micrograms per minute ($\mu\text{g}/\text{min}$, which has physical dimensions of mass per unit time). Hint: the answer depends on $u(t), V$, and r .

1.2.3 The Differential Equation

Let's now put it all together. Based on (1.4) and Reading Exercises 1.2.1-1.2.3 we have

$$\underbrace{u'(t)}_{\text{rate of change of } u(t)} = \underbrace{rc_1}_{\text{rate in}} - \underbrace{\frac{r}{V}u(t)}_{\text{rate out}}. \tag{1.5}$$

Each term in (1.5) has units of μg per minute.

It's worth noting that this type of reasoning, "rate of change equals rate in minus rate out," appears frequently in modeling. It rests on a **conservation law**, a principle that requires that a substance is neither created nor destroyed in a given situation. In this case the drug is neither created nor destroyed in the cochlea. In situations where the drug or other substance is created or destroyed (which we'll encounter) (1.5) must be modified to account for this.

Reading Exercise 1.2.4 A patient is implanted with a reciprocating perfusion device to treat hearing loss. Suppose that the drug reservoir is primed with a drug solution at a concentration of $1.2 \mu\text{g}/\mu\text{L}$ (micrograms per microliter). The drug solution is infused to the patient's cochlea at a steady rate of $1 \mu\text{L}$ of the drug solution every 30 minutes. Simultaneously the well-mixed fluid in the cochlea is withdrawn at the same rate. The fluid volume inside the cochlea stays constant at $200 \mu\text{L}$. What does (1.5) become in this case? If time t is measured in minutes with $t = 0$ corresponding to the moment drug delivery begins, what is the appropriate initial condition?

Reading Exercise 1.2.5 Verify that the function

$$u(t) = c_1 V (1 - e^{-rt/V}) \quad (1.6)$$

satisfies (1.5) with initial condition $u(0) = 0$. That is, if $u(t)$ is as defined in (1.6), then $u'(t)$ equals $rc_1 - ru(t)/V$. With the parameters r, V , and c_1 of Reading Exercise 1.2.4, use (1.6) to determine how much of the drug (μg) will be in the cochlea after one week, and after two weeks. What is the concentration (in μg per μL) of the drug in the cochlea at each of these times? Plot the amount and concentration of the drug in the cochlea over time. What do you observe? Does this seem reasonable?

Remark 1.2.1 At this point you should take note of an important fact, a theme that will recur over and over throughout this book and is well-illustrated by the previous two sections. The model of Usain Bolt's Olympic victory as described by the ODE (1.3) and the intracochlear drug delivery ODE (1.5) are more than just similar: for all practical purposes *they are exactly the same differential equation*. Each is of the general form

$$x'(t) = a + bx(t)$$

for constants a, b , and an unknown function $x(t)$. In the Hill-Keller ODE we have $a = P$ and $b = -k$, with function $x(t) = v(t)$. In the intracochlear model we have $a = rc_1$ and $b = -r/V$, with $x(t) = u(t)$. As a result, the analysis we perform on any one equation will be valid for the other, and this allows us to make some general conclusions that apply to these very different physical systems. These models illustrate the power of ODEs (and more generally, mathematics) to highlight and capture the commonality of situations that, on the surface, seem very different.

1.3 Population Growth and Fishery Management

The material in this section is based on the SIMIODE project "Fishery Harvesting" [42].

1.3.1 The Need to Manage Fish Harvesting

Fish are a valuable source of protein, and many people depend to a large extent on fish and other seafood for nourishment. However, overfishing has driven many species of fish to near extinction [63, 64]. This applies in particular to the Mediterranean Sea, the Baltic Sea, and the North Atlantic Ocean. The collapse of the Newfoundland or Baltic sea cod fisheries should be taken as a pointed warning that the fishing industry needs more careful controls [75]. With appropriate stock assessment data, mathematical models can be used to derive possible management strategies, which may aid the supervision and enduring success of this industry.

U.S. stocks of Atlantic cod came close to commercial collapse in the mid-1990s. This precipitous decline is illustrated in Figure 1.3. The 2012 assessments of Gulf of Maine and Georges Bank cod indicated that both stocks are seriously overfished and are not recovering as quickly as expected. Based on these assessments, quotas for fishing for both stocks were significantly reduced in 2013 to help ensure that overfishing does not occur and that these stocks rebuild. The Gulf of Maine cod quota was cut by 80%, and the Georges Bank cod quota was cut by 61%. National Oceanic and Atmospheric Administration (NOAA) Fisheries and the New England Fishery Management Council continue to work on management measures that will further protect cod stocks and provide opportunities for fishermen to target other healthy fish stocks instead of cod [2]. A quantitative model of how harvesting affects the fish population is an essential part of any program to manage this industry sustainably.

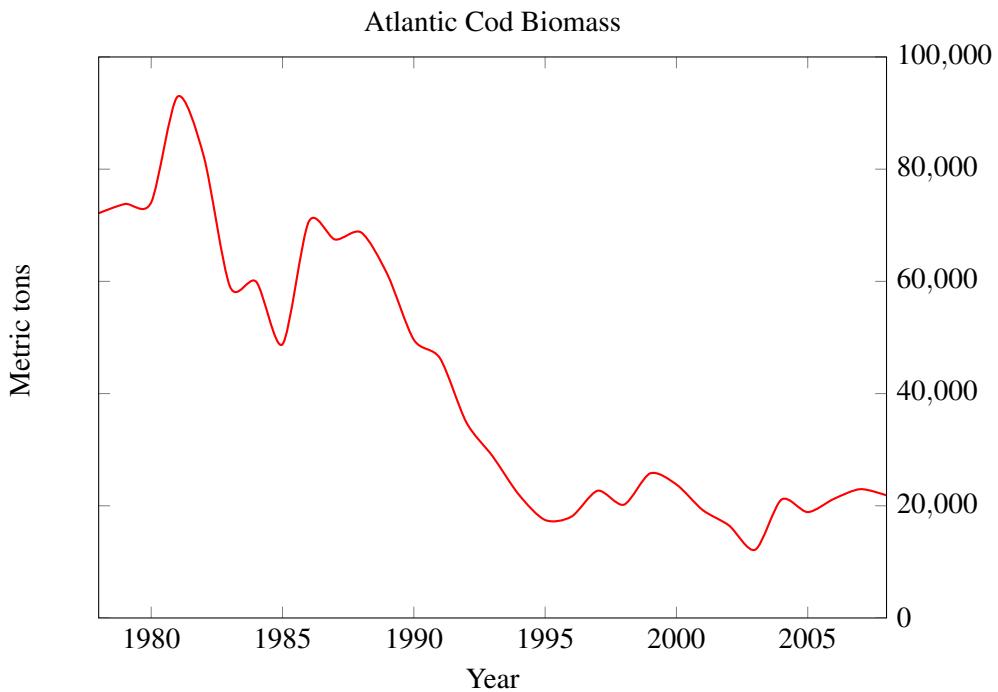


Figure 1.3: Atlantic Cod Biomass, 1978-2008.

1.3.2 Modeling Fish Population

Let us consider the cod population to be confined to a closed, finite region of the ocean. We will use $u(t)$ to denote the cod population in this region at time t ; units for u and t will be specified below. One of the simplest models for the population of an organism in a given environment, whether these organisms are bacteria, fish, or humans, is equation

$$u'(t) = ru(t). \quad (1.7)$$

This model is based on the assumption that at any time the population produces new individuals at a rate proportional to the number of individuals present at that time. Here r is a positive constant of proportionality, and is called the **intrinsic growth rate**. Equation (1.7) is a differential equation with solution

$$u(t) = u_0 e^{rt}, \quad (1.8)$$

where u_0 is the population at time $t = 0$. The drawback to the model (1.7) is that the solution (1.8) exhibits **exponential growth**, without limit, so the population tends to infinity. We need a better model.

Reading Exercise 1.3.1 Verify that $u(t)$ as defined by (1.8) satisfies (1.7) with $u(0) = u_0$. How long does it take the population to double from its initial population u_0 ? That is, at which time t is $u(t) = 2u_0$ satisfied? Hint: the answer depends on r . How long does it take for the population to quadruple? How long does it take to increase eight-fold over the initial population?

The difficulty with (1.7) is that it models the growth rate r as constant, regardless of how large the population becomes. In reality as the population increases, limits on space, food, and other resources (to say nothing of disease and predation) should slow the population growth. One common approach to capture this idea is to alter the growth rate r so that it decreases as the population increases. We thus assume that r tapers to zero at some maximum sustainable value for the population. This maximum sustainable population is commonly called the **carrying capacity** of the environment. Let us use K to denote this population value. Of course K is positive.

To incorporate these ideas into the model, write (1.7) in the form $u'(t)/u(t) = r$; this emphasizes that in our first model new individuals are produced at a constant rate of r individuals per unit time per individual in the population. But in the new model this rate should drop to zero when $u = K$, the maximum sustainable population. A simple modification to (1.7) that accomplishes this is

$$\frac{u'(t)}{u(t)} = r(1 - u(t)/K). \quad (1.9)$$

The right side in (1.9) is a modification to account for the limited resources available to the population that limits growth.

Reading Exercise 1.3.2

- (a) If $u(t) \approx 0$ (but $u(t)$ is still positive) at some time t , what does the right side of (1.9) equal? What is the growth rate $u'(t)/u(t)$ of the population?
- (b) If $u(t) = K$ at some time t (the population is at the carrying capacity), what does the right side of (1.9) equal? What is the growth rate $u'(t)/u(t)$ of the population?
- (c) If $u(t) > K$ at some time t (the population is above the carrying capacity), show that $u'(t)/u(t) < 0$. What is the population $u(t)$ doing at this time?

With this modification the intrinsic growth rate r in (1.9) might be interpreted as a maximum growth rate, the growth rate that the organism is capable of when $u(t) \approx 0$ and environmental limitations have not come into play. Equation (1.9) is conventionally written in the form

$$u'(t) = ru(t)(1 - u(t)/K), \quad (1.10)$$

obtained by multiplying both sides of (1.9) by $u(t)$; (1.10) is called the **logistic equation**.

Reading Exercise 1.3.3 What units on r are necessary for (1.10) (or (1.7)) to be dimensionally consistent, if u measures the population in units of organisms and t is measured in days? What units are necessary for K ?

Reading Exercise 1.3.4 As you will compute in the next chapter, the solution $u(t)$ to (1.10) with initial condition $u(0) = u_0$ is

$$u(t) = \frac{K}{1 + e^{-rt}(K/u_0 - 1)}. \quad (1.11)$$

Take $K = 10$, $r = 1$, and $u_0 = 2$ in (1.11). Plot the solution $u(t)$ for $0 \leq t \leq 10$. Is it consistent with the modeling assumptions that were made? Try increasing or decreasing the value of r ; how does this affect the behavior of the solution? What happens if you take $u_0 > 10$?

1.3.3 Modeling Harvesting

Let's now consider the case in which the population quantified by $u(t)$ is harvested at some rate, that is, a certain portion of the population is taken out of the environment per unit time. To be specific, let's focus on the Atlantic cod population. We will assume that the cod are harvested by humans at a rate that is proportional to the number of cod present. The reasoning is that if fisherman put a certain amount of effort into catching fish for a certain period of time (e.g., the number of boats in the water), then the number of fish caught should be proportional to the number of fish present.

Let us call this constant of proportionality h , and so assume that the rate at which fish are harvested (fish per unit time) is $hu(t)$. Since the rate at which the fish are reproducing is quantified by the right side of (1.10) (fish per unit time) and humans are harvesting fish at rate $hu(t)$, the rate at which $u(t)$ is changing is the difference between these quantities, $ru(t)(1 - u(t)/K) - hu(t)$. This yields an ODE

$$u'(t) = ru(t)(1 - u(t)/K) - hu(t), \quad (1.12)$$

which is called the **logistic equation with harvesting**. See [37] for more discussion of this model.

Reading Exercise 1.3.5 Before considering the solution to the differential equation (1.12), what do you expect of the behavior of the fish population when $h > 0$? Will harvesting increase or decrease the long-term population? What might happen if the harvesting constant h is very large?

Reading Exercise 1.3.6 The solution to (1.12) (which we will deduce in the next chapter) is

$$u(t) = \frac{(1 - \frac{h}{r})K}{1 + e^{-(r-h)t}(\frac{K}{u_0}(1 - \frac{h}{r}) - 1)}. \quad (1.13)$$

When $h = 0$ this is the same as (1.11).

- (a) Take $K = 10, r = 1, u_0 = 2$, and $h = 0.1$ in (1.13). Plot the solution for $0 \leq t \leq 10$, and compare to the solution (1.11) with these same choices for K, r , and u_0 . Are the results consistent with the modeling assumptions that were made?
- (b) Repeat part (a) but increase h to 0.5. What happens?
- (c) What is $\lim_{t \rightarrow \infty} u(t)$ in (1.13)? How large can h be before the population cannot survive?

1.3.4 Parameter Estimation and Harvesting

The estimated Atlantic cod biomass (in metric tons) and harvest rate h in Georges Bank from 1978 to 2008 are given in Table 1.2; see [130]. Note the fish population is estimated not in individuals, but in total mass. Let's assume that these quantities are proportional to each other, so that biomass can be used as a proxy for population and the logistic model with harvesting derived above should still hold, if we instead think of $u(t)$ in terms of mass, rather than individuals.

Reading Exercise 1.3.7 Although the harvest rate h in Table 1.2 varies, let us model this as a constant for the moment. The average value for h_t in Table 1.2 is $h \approx 0.200$ over the time period listed. If we treat 1978 as time $t = 0$ with t measured in years, then the initial condition is $u(0) = 72,148$, with $u(t)$ in units of metric tons. Plot the data in Table 1.2. With $h = 0.2$ and $u_0 = 72148$, can you find values for r and K that provide a reasonable fit to the data when you plot $u(t)$ from (1.13)? Hint: try something around $K = 10^5$, and r just a bit larger than 0.2. You may find that a different value for u_0 (or even h) gives better results.

The process of adjusting unspecified parameters in a model to fit data is known as **parameter estimation**. In Section 3.4 we'll look at more methodical and effective ways to estimate these parameters.

Year	u_t	h_t	Year	u_t	h_t	Year	u_t	h_t
1978	72,148	0.18847	1988	68,702	0.23154	1998	20,196	0.18953
1979	73,793	0.14974	1989	61,191	0.20860	1999	25,776	0.17011
1980	74,082	0.21921	1990	49,599	0.33565	2000	23,796	0.15660
1981	92,912	0.17678	1991	46,266	0.29534	2001	19,240	0.28179
1982	82,323	0.28203	1992	34,877	0.33185	2002	16,495	0.25287
1983	59,073	0.34528	1993	28,827	0.35039	2003	12,167	0.25542
1984	59,920	0.20655	1994	21,980	0.28270	2004	21,104	0.08103
1985	48,789	0.33819	1995	17,463	0.19928	2005	18,871	0.08740
1986	70,638	0.14724	1996	18,057	0.18781	2006	21,241	0.08195
1987	67,462	0.19757	1997	22,681	0.19357	2007	22,962	0.10518
						2008	21,848	unknown

Table 1.2: Annual (1978–2008) values of Atlantic cod biomass in metric tons, u_t , and harvest rate, $h(t)$, in Georges Bank, from [130].

Reading Exercise 1.3.8 What is the long-term behavior of the cod population with the parameters you found in Reading Exercise 1.3.7? What does harvesting do to the cod population? According to your model, can the cod population survive under these conditions? Using the data in Table 1.2, if we increase the constant harvest rate, h , to be 0.4, how will the population of Atlantic Cod change over time?

Acknowledgement

We used a simplified version of the models from W. Ding, G.E. Herrera, H.R. Joshi, S. Lenhart, and M.G. Neubert [43, 69].

1.4 Where Do We Go from Here?

1.4.1 A Toolbox for Describing the World

We've modeled a world-class sprinter, a microfluidic pump, and the population of a species occupying a swath of ocean a thousand miles wide. These phenomena evolve on vastly different scales in time and space, yet all can be described by an equation of the form

$$u'(t) = f(t, u(t)). \quad (1.14)$$

This is quite remarkable and provides testimony to the importance and ubiquity of differential equations as a tool for describing the world.

In each case the function $u(t)$ in (1.14) is considered as an unknown to be found, while the function $f(t, u)$ defines the precise ODE that arises from the physical model. In the Hill-Keller model $f(t, u) = P - ku$ (though there we used v instead of u), while in the intracochlear drug delivery model we had $f(t, u) = rc_1 - ru/V$, and in the fish harvesting model we had $f(t, u) = ru(1 - u/K) - hu$. In each case we used an additional piece of information: an initial condition of the form $u(t_0) = u_0$.

For the ODEs encountered so far, we presented an explicit closed-form or analytical solution with which to experiment. In the remainder of this text we'll look at how one can methodically find such solutions to ODEs like (1.14) and many others. In cases where an analytical solution cannot be found, we will explore other techniques for gleaned information about solutions. The models developed in this chapter, their extensions, and additional models we'll develop later will serve as templates to illuminate the techniques presented in the coming chapters.

1.4.2 Some Terminology

In discussing how to solve or otherwise analyze differential equations, the approach we take will depend greatly on the structure of the differential equation, so it's helpful to make a few definitions.

Scalar ODEs, Systems, and PDEs

The focus of this text is **ordinary differential equations**. In the scalar case this means there is an unknown function $u(t)$ of a single independent variable t . Equations (1.3), (1.5), and (1.12) are examples. But one may also consider systems of ordinary differential equations, for example,

$$\begin{aligned} u'(t) &= u(t) - u(t)v(t), \\ v'(t) &= -2v(t) + 3u(t)v(t) \end{aligned}$$

in which two (or more) unknown functions of a single independent variable t appear. The focus of the first five chapters will be scalar ODEs. In Chapters 6 and 7 we will develop techniques for analyzing system of ODEs.

Ordinary differential equations stand in contrast to **partial differential equations** (or **PDE's**) in which the unknown function depends on two or more independent variables. An example is the **wave equation**

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0,$$

where $u(x, t)$ depends on two independent variables, x and t . One can even have systems of partial differential equations.

Order and Linearity

Equation (1.14) is an example of a **first-order** differential equation, that is, an equation involving an unknown function $u(t)$, in which the highest derivative of u that appears is the first derivative. We'll refer to (1.14), in which $u'(t)$ is given explicitly in terms of t and $u(t)$, as the **standard form** for a first-order ODE. More generally, the **order** of an ODE is the highest derivative of the unknown function that appears. Second-order differential equations (involving $u''(t)$) are also common, and form much of the mathematical basis for significant engineering applications. Occasionally higher-order equations make an appearance.

The distinction between linear and nonlinear ODEs is the last we need to make for the moment. An n th-order ODE for $u(t)$ is **linear** if it can be written as

$$a_n(t)u^{(n)}(t) + a_{n-1}(t)u^{(n-1)}(t) + \cdots + a_1(t)u'(t) + a_0(t)u(t) = b(t) \quad (1.15)$$

for some given functions $a_k(t)$, $b(t)$, $0 \leq k \leq n$, where $u^{(k)}$ denotes the k th derivative of $u(t)$. Equations that are not linear are **nonlinear**. Linear equations have a lot of structure and are often comparatively easy to analyze or solve. Nonlinear equations are everything else and their analysis can be a bit like the Wild West, although there are techniques of general utility.

For convenience, let's amalgamate the essential terminology into the following definition.

Definition 1.4.1 — Basic ODE Terminology. Let $u(t)$ be a function of a single independent variable t . A **scalar ordinary differential equation** for u is an equation of the form

$$F(t, u(t), u'(t), \dots, u^{(n)}(t)) = 0,$$

where F is a function of $n+1$ variables that relates the independent variable t , the function $u(t)$, and the derivatives $u'(t), \dots, u^{(n)}(t)$. The integer n indicates the highest derivative of u that appears in the equation and is called the **order** of the differential equation. If the ODE can

be written in the form of (1.15) then the differential equation is **linear**.

There are other important adjectives to describe ODEs, such as constant-coefficient, autonomous, separable, homogeneous, and more. Their definitions will be presented later, in the context where they naturally arise. Each type of equation demands different techniques for analysis.

1.4.3 You Already Know How to Solve Some Differential Equations

You are already familiar with the solution procedure for certain types of differential equations, specifically, those that can be solved with one or more straightforward applications of antiderivation.

First-Order Equations

Consider a first-order differential equation of the form

$$u'(t) = g(t), \quad (1.16)$$

where $g(t)$ is a specified function and $u(t)$ is the unknown to be found. Comparison of (1.16) to (1.14) shows that the right side in (1.16) does not involve the unknown function u . In this case we can antiderivative both sides of (1.16) with respect to t and find

$$u(t) = \int g(t) dt + C, \quad (1.17)$$

where C is an arbitrary constant of integration. Equation (1.17) is a **general solution** to the ODE (1.16): all functions that satisfy (1.16) can be expressed in the form of (1.17) for some choice of C . If an initial condition $u(t_0) = u_0$ is given then the constant C can be adjusted so that $u(t)$ in (1.17) satisfies this initial condition.

■ **Example 1.1** Let us find a general solution to the ordinary differential equation

$$u'(t) = 3t^2$$

and use this to find a solution with the initial condition $u(1) = 3$.

Antiderivative both sides of the ODE yields general solution $u(t) = t^3 + C$. The initial condition $u(1) = 3$ requires $u = 3$ when $t = 1$, so substitute this into this general solution to find $3 = 1 + C$ and solve for $C = 2$. The solution with the required initial condition is $u(t) = t^3 + 2$. ■

Reading Exercise 1.4.1 Verify that $u(t) = t^3 + 2$ does, in fact, satisfy $u'(t) = 3t^2$ with $u(1) = 3$. If the initial condition is changed to $u(t_0) = u_0$ for unspecified values of t_0 and u_0 , what would the solution $u(t)$ be (it depends on t_0 and u_0)? Can $u'(t) = 3t^2$ be solved with any initial data t_0 and u_0 ?

Reading Exercise 1.4.2 Find a general solution to $u'(t) = e^{2t}$. Use this general solution to find a particular solution with $u(0) = 8$.

Second- and Higher-Order Equations

At this time certain second-order differential equations are also within reach, specifically, second-order equations of the form

$$u''(t) = g(t). \quad (1.18)$$

To solve this ODE, first integrate both sides of (1.18) with respect to t to find

$$u'(t) = \int g(t) dt + C_1, \quad (1.19)$$

where C_1 is an arbitrary constant of integration. Let $G(t) = \int g(t) dt$ be any antiderivative for $g(t)$, so (1.19) becomes $u'(t) = G(t) + C_1$. Integrating both sides of (1.19) with respect to t then yields

$$u(t) = \int G(t) dt + C_1 t + C_2, \quad (1.20)$$

where C_2 is a second arbitrary constant of integration. Equation (1.20) is a general solution to (1.18). Finding the constants C_1 and C_2 requires two additional pieces of information about the solution. They typically (but not always) take the form of initial conditions such as $u(t_0) = u_0$ and $u'(t_0) = u'_0$ for some initial time t_0 and constants u_0 and u'_0 .

■ **Example 1.2** Let us find a general solution to the ordinary differential equation

$$u''(t) = e^t$$

and then find a particular solution with the initial conditions $u(1) = 2, u'(1) = 3$.

Antidifferentiating both sides of the ODE with respect to t yields

$$u'(t) = e^t + C_1.$$

Antidifferentiate again to find

$$u(t) = e^t + C_1 t + C_2.$$

This is a general solution to the ODE. The condition that $u'(1) = 3$ requires that $e + C_1 = 3$, so $C_1 = 3 - e$. The condition $u(1) = 2$ then requires $e + (3 - e) + C_2 = 2$, so $C_2 = -1$. The solution with the required initial condition is

$$u(t) = e^t + (3 - e)t - 1.$$

■

Reading Exercise 1.4.3 Verify that $u(t) = e^t + (3 - e)t - 1$ does in fact satisfy $u''(t) = e^t$ with $u(1) = 2$ and $u'(1) = 3$. If the initial conditions are $u(t_0) = u_0$ and $u'(t_0) = u'_0$ for given constants t_0, u_0 , and u'_0 , what would the solution be? (It depends on t_0, u_0 and u'_0 .) Can $u''(t) = e^t$ be solved with any initial conditions of this form?

Reading Exercise 1.4.4 Find a general solution to $u''(t) = \sin(t)$. Use this general solution to find a particular solution with initial data $u(0) = 2$ and $u'(0) = 4$.

It should be clear that this process can be extended to solve any differential equation of the form $u^{(n)}(t) = g(t)$, where $u^{(n)}(t)$ denotes the n th derivative of u . Integrate n times, and in the process pick up n constants of integration that require n additional pieces of information to find. This information is often in the form of initial data that specifies values for $u(t_0), u'(t_0), \dots, u^{(n-1)}(t_0)$.

Remark 1.4.1 You may have noticed that we usually refer to finding “a” general solution to an ODE rather than “the” general solution. The reason is that a general solution to an ODE may assume various superficially different forms, especially later in the text. For instance, in Example 1.2 a general solution $u(t) = e^t + C_1 t + C_2$ was found. But $u(t) = e^t + 17C_1 t - 3C_2$ could also be considered a general solution, since all solutions to $u''(t) = e^t$ are still of this form, and C_1 and C_2 can still be adjusted to obtain any initial conditions.

1.4.4 Exercises

Exercise 1.4.1 For each ODE and initial condition below, find a general solution to the ODE, and then find the specific solution with the given initial condition by using the technique of Example 1.1 or Example 1.2 as appropriate.

- (a) $u'(t) = t, u(0) = 3$
- (b) $u'(t) = \cos(t), u(0) = 0$
- (c) $u'(t) = e^t, u(0) = 4$
- (d) $u'(t) = 1/t, u(2) = 0$
- (e) $u'(t) = \cos(t), u(0) = 1$
- (f) $u'(t) = t \cos(t), u(0) = 1$
- (g) $u'(t) = 1/(t^2 + 1), u(1) = 2$
- (h) $v'(t) = g, v(0) = v_0$, where g and v_0 are some constants
- (i) $h'(t) = t^n, h(0) = 0$, where n is a positive integer
- (j) $u''(t) = t, u(0) = 1, u'(0) = 3$
- (k) $u''(t) = \sin(t), u(0) = 1, u'(0) = 0$
- (l) $y''(t) = -g, y(0) = 10, y'(0) = 0$, where g is a constant
- (m) $x''(t) = 5 - e^{-2t}, x(0) = 0, x'(0) = 0$

Exercise 1.4.2 The model for drug delivery to the cochlea in Section 1.2 is a special case of a more general compartmental model in which one has a tank of volume V (in our model, the tank was the cochlea) with an input pipe and an output pipe. These types of problems are often called **salt tank models**, since the input and output pipes are assumed to carry salt, dissolved in water. In our model the drug played the role of the salt. The general situation is still accurately depicted by Figure 1.2.

Consider a tank of volume $V = 100$ liters, into which a pipe delivers a salt solution at a rate of 5 liters per minute; this input salt solution has a concentration of 50 grams of salt per liter. The solution in the tank is well-stirred and always of uniform concentration. An output pipe carries away this well-stirred solution, also at 5 liters per minute. Let $u(t)$ denote the amount (grams) of salt in the tank at time t . If the tank starts with no salt at time $t = 0$, use the reasoning that led to (1.5) to formulate an appropriate ODE and initial condition for u . Use (1.6) to write out the solution. Plot the solution $u(t)$ for $0 \leq t \leq 200$ minutes. What limit does the amount of salt in the tank approach? What limit does the concentration approach? Does this make sense, in light of the incoming fluid concentration?

1.5 The Blessing of Dimensionality

1.5.1 Definition of Dimension

The subject of differential equations involves a lot of fundamental physical quantities such as distances, velocities, electric charge, mass, etc. Most of these quantities have units or physical dimensions. For example, mass, length, and time are fundamental dimensions. Other dimensions we'll encounter later are electric charge and temperature. These are the basic building blocks for the dimension of all other quantities in this text. In this section we'll look at how the consideration of the basic dimensions of physical quantities can aid mathematical modeling and setting up ODEs, and provide a sanity check for our work. This is the subject of **dimensional analysis**.

Mass, Length, and Time

To illustrate the notion of dimension, a variable r in a given problem may have the dimension of length, in which case we will write

$$[r] = L.$$

The notation $[r]$ indicates the physical dimension of the quantity r and L is the notation for the physical dimension of length. Note that L here is not the actual length of whatever r quantifies; L just stands for the dimension length. We will use T to denote the dimension of time and M to denote the dimension of mass. The dimension of many other common quantities can be derived from these. Further examples:

- If A denotes an area then $[A] = L^2$. If V denotes a volume then $[V] = L^3$.
- If v is a velocity then $[v] = LT^{-1}$, length (or distance) per time.
- If a is an acceleration then $[a] = LT^{-2}$.
- If ρ is a density (mass per volume) then $[\rho] = ML^{-3}$.

The dimension of a physical quantity will generally be expressed as $M^aL^bT^c$ where a, b , and c may be positive, negative, or zero. In many but not all cases a, b , and c will be integers.

Dimension Versus Units

The dimension of a physical variable is not quite the same as the units used to measure the variable. Thus length is a fundamental physical dimension, but it can be measured using many systems of units, e.g., meters, feet, or inches. If I ever get sloppy and refer to the dimension of a velocity as “meters per second” feel free to write me a stern email. However, specifying the units of a given quantity does allow us to determine its dimension.

Reading Exercise 1.5.1 Air is being pumped into a balloon at a fixed rate q liters per second. What is $[q]$? What is the dimension of the rate at which the balloon surface area is changing in time? What is the dimension of the rate at which the radius of the balloon is increasing?

1.5.2 The Algebra of Dimension

Add, Subtract, Multiply, Divide

In addition to mass, length and time, we'll later encounter charge (denoted by Q) and temperature (denoted by Θ .) It is a fundamental property of our mathematical framework for describing the world that both sides of any equation or inequality involving physical variables must have the same physical dimensions. It is nonsense to ask if the length of a string equals the mass of an apple or whether a certain time interval is longer than a stick. Similarly, it only makes sense to add or subtract physical quantities that have the same dimensions. You can add years to your life span, but you cannot add years to the mass of an apple. We can, however, take products and quotients of dimensionally dissimilar quantities, for example, dividing a length by a time interval to obtain a velocity. If a variable x has dimension $[x] = M^{a_1}L^{a_2}T^{a_3}$ and variable y has dimension $[y] = M^{b_1}L^{b_2}T^{b_3}$ then

$$[xy] = [x][y] = M^{a_1+b_1}L^{a_2+b_2}T^{a_3+b_3} \quad \text{and} \quad [x/y] = [x]/[y] = M^{a_1-b_1}L^{a_2-b_2}T^{a_3-b_3}. \quad (1.21)$$

Reading Exercise 1.5.2 If v has dimension $[v] = LT^{-1}$ (a velocity, perhaps) and $[\Delta t] = T$, what is the dimension of $v\Delta t$? What is a physical interpretation of this situation?

Dimensionless Constants

In many formulas certain **dimensionless** mathematical constants appear. For example, the formula for the area of a circle is $A = \pi r^2$. Here $[r] = L$, $[A] = r^2$, and π is a dimensionless constant. We

write $[\pi] = M^0 L^0 T^0$ to denote this, or just $[\pi] = 1$. (Careful: put the square brackets around π or else you're claiming $\pi = 1$.) In accordance with (1.21) it follows that

$$[A] = [\pi r^2] = [\pi][r^2] = M^0 L^0 T^0 M^0 L^2 T^0 = M^0 L^2 T^0 = L^2.$$

It's also worth noting that the angular measure radian is dimensionless. The definition of the radian involves the ratio of two lengths (the radius of a circle and the arc length of that portion of the circle subtended by the angle); the ratio of these two lengths is dimensionless.

Deducing Dimension from Common Formulas

It's frequently possible to determine the dimension of certain quantities by looking at familiar formulas that involve them. For example, what is the dimension of force? If you remember $F = ma$ and know that $[m] = M$ and $[a] = LT^{-2}$ then

$$[F] = [m][a] = MLT^{-2}.$$

What is the dimension of kinetic energy? You may recall the formula $E = \frac{1}{2}mv^2$ for the kinetic energy E of a mass m moving at speed v . Since $[m] = M$, $[v] = LT^{-1}$, and $1/2$ is dimensionless, we find

$$[E] = [1/2][m][v]^2 = ML^2T^{-2}.$$

Reading Exercise 1.5.3 Newton's universal law of gravitation specifies that the force F between two point masses m_1 and m_2 separated by a distance r is

$$F = \frac{Gm_1 m_2}{r^2}.$$

Use this to determine $[G]$.

1.5.3 Derivatives, Integrals, Elementary Functions

Differentiation With Respect to Time

Suppose that $y = f(t)$ is a function with input argument t , a time, and f outputs a physical variable y with dimension $[y] = M^a L^b T^c$ for some constants a, b, c . What is the dimension of the derivative $f'(t)$? This is a common situation. The derivative is defined as

$$f'(t) = \lim_{\Delta t \rightarrow 0} \frac{f(t + \Delta t) - f(t)}{\Delta t}.$$

The numerator $f(t + \Delta t) - f(t)$ has dimension $[f(t + \Delta t) - f(t)] = M^a L^b T^c$ and Δt has dimension $[\Delta t] = T$. As a result, the difference quotient $(f(t + \Delta t) - f(t))/\Delta t$ has dimension $M^a L^b T^{c-1}$ and $f'(t)$, as the limit of such quantities, also has this dimension. That is,

$$[f'(t)] = M^a L^b T^{c-1}.$$

For example, if $f(t)$ is a function that outputs a position (displacement from the origin) as a function of time t then $[f] = L$, while $[t] = T$. Then $[f'(t)] = LT^{-1}$, which has the dimension of velocity.

Reading Exercise 1.5.4 An object has kinetic energy $E(t)$ that varies with time. What is the dimension of $E'(t)$? What is the dimension of $E''(t)$?

Integration With Respect to Time

If $f(t)$ is a function that accepts time t as an input argument and outputs a variable y with dimension $[y] = M^a L^b T^c$ then

$$\left[\int_p^q f(t) dt \right] = M^a L^b T^{c+1}.$$

This isn't surprising, given that computing the integral typically involves computing an antiderivative, but this can also be deduced using the definition of the integral. Recall from basic calculus that the integral is defined as the limit of a Riemann sum,

$$\int_p^q f(t) dt = \lim \sum_k f(t_k) \Delta t_k.$$

We needn't go into the details of the nature of this limit here. It suffices to note that each term $f(t_k) \Delta t_k$ has dimension $[f(t_k)][\Delta t_k] = M^a L^b T^{c+1}$ and hence so does the sum, and also the limit.

Reading Exercise 1.5.5 Water flows into a tank at a variable rate of $r(t)$ liters per second. What is the dimension of $r(t)$? What is the dimension of

$$\int_a^b r(t) dt?$$

What is the physical interpretation of this integral?

Elementary Functions

Consider expressions like $\sin(z), \cos(z), e^z$, each an elementary transcendental function of a variable z . These types of expressions require that the input argument z be dimensionless. The reason is that these expressions have Taylor series, typically of the form

$$a_0 + a_1 z + a_2 z^2 + \dots$$

(the a_k are dimensionless), and so each of $1, z, z^2, \dots$ must have the same dimension if they are to be added. This only occurs if z is dimensionless, and in this case the sum consists of dimensionless quantities, and hence is itself dimensionless. As an example, consider the expression $\cos(\omega t)$. Here ωt should be dimensionless; if t is time ($[t] = T$) then it must be the case that $[\omega] = T^{-1}$, reciprocal time. This is the case— ω is always some kind of radial frequency, with the dimension of reciprocal time.

1.5.4 Unit-Free Equations and Bending the Rules

Unit-Free Equations

In general when we write down fundamental laws of physics or differential equations the equations should be independent of any particular system of units. As an example, consider the formula $d = \frac{1}{2}gt^2$ for the distance an object falls in t time units under the influence of gravitational acceleration g , with no other forces acting on the object. This equation holds in any system of units. In the SI system however, the equation can be written approximately as $d = 4.9t^2$, while in English units it becomes $d = 16t^2$. These expressions hard-code the units into the equation. Equations that do not depend on the system of units are said to be **unit-free** and are usually a more desirable way to express the situation.

As another example, consider the Hill-Keller model (1.3), $v' = P - kv$. This model remains unchanged when the system of units changes. If we use $P = 11$ meters per second squared, however, the ODE $v' = 11 - kv$ is specific to SI units and won't be $v' = P - kv$ in, say, English units.

Bending the Rules

We won't always strictly adhere to these rules, so long as no confusion results. For example, we may wish to express the position of a particle moving along the x axis as a function of time t as $x = \cos(\omega t)$. Here ω has dimension T^{-1} and $[t] = T$, so the argument of the cosine function is dimensionless, in accordance with the discussion above. But then $\cos(\omega t)$ is a dimensionless quantity, while x should have dimension L . Writing $x = \cos(\omega t)$ requires choosing a unit for length; it would be more precise to say $x = A \cos(\omega t)$ where $[A] = L$ and $A = 1$ in whatever units we choose to measure length. In principle this is what we'll do, we just won't remark on it, except in cases where it might cause confusion.

1.5.5 Using Dimension to Find Plausible Models

The fact that physical quantities come with a dimension can be an incredibly powerful tool for figuring out things that we have no right to know. As an example, consider a black hole, a roughly spherical region in space-time left behind by the collapse of a massive star. How does the radius r of the black hole depend on its mass m ? Since $[r] = L$ and $[m] = M$, there must be other variables involved. Black holes are black because light cannot escape them, so maybe the speed of light c also plays a role; $[c] = LT^{-1}$. But light can't escape because of the intense gravitational field, so presumably the gravitational constant G is important. In Reading Exercise 1.5.3 you showed that $[G] = M^{-1}L^3T^{-2}$.

Let's put these observations together. Suppose a formula of the form

$$r = KG^\alpha c^\beta m^\gamma \quad (1.22)$$

holds for some constants α, β , and γ (that need not be integers) and dimensionless constant K . What choices for α, β, γ lead to a dimensionally consistent formula? To find out, note that

$$[KG^\alpha c^\beta m^\gamma] = [K][G]^\alpha [c]^\beta [m]^\gamma = (M^{-\alpha} L^{3\alpha} T^{-2\alpha})(L^\beta T^{-\beta})M^\gamma = M^{-\alpha+\gamma} L^{3\alpha+\beta} T^{-2\alpha-\beta},$$

after combining exponents. If the right side is to have the same dimension as r , namely $[r] = M^0 L^1 T^0$, then $-\alpha + \gamma = 0$ (to match the M exponents), $3\alpha + \beta = 1$ (to match the L exponents), and $-2\alpha - \beta = 0$ (to match the T exponents). The solution to these three equations in three unknowns is $\alpha = 1, \beta = -2, \gamma = 1$. From (1.22), a formula of the form

$$r = K \frac{Gm}{c^2}$$

is dimensionally consistent, and in fact the only dimensionally consistent formula involving G, m, c and r in the form (1.22). The dimensionless constant K can be anything.

With $K = 2$ the formula is correct. The radius r is called the **Schwarzschild radius** of the black hole. It is often the case that this kind of dimensional analysis leads to a formula of the correct form with one or more dimensionless constants that are simple, e.g., 2 or π or such. It seems rather amazing that we just derived an important result from physics that presumably requires an understanding of general relativity to truly understand, but we used nothing more than the dimensions of the variables involved.

1.5.6 Other Dimensions

Later in the text we will encounter other physical dimensions, specifically temperature, which has dimension denoted by Θ , and electric charge, which has dimension denoted by Q . These dimensions are independent from mass, length, and time. In certain specific instances it can be helpful to temporarily assign other dimensions. For example, in a problem involving money we could use V to denote the dimension value, that is, the worth of some quantity; it might be tempting

to use the symbol \$ but that is a unit, dollars. See the project “Money Matters” in Section 1.6 and [110]. In a population model we might use N to denote the dimension of population for some species. In a problem involving two or more species we might introduce a unique dimension for each. See Exercise 1.5.6.

1.5.7 Exercises

Exercise 1.5.1

- (a) What is the dimension of momentum?
- (b) What is the dimension of angular velocity?
- (c) What is the dimension of work (force times distance)?
- (d) What is the dimension of pressure?

Exercise 1.5.2 The energy of a photon with wavelength λ is $E = hc/\lambda$, where c is the speed of light and h is Planck’s constant. Find the dimension of Planck’s constant.

Exercise 1.5.3 What must be the dimension of the constant k in the Hill-Keller ODE (1.3)?

Exercise 1.5.4 Suppose $f(x)$ is a function that outputs a variable with dimension $[f] = M^a L^b T^c$ and that the input argument has dimension $[x] = M^\alpha L^\beta T^\gamma$. What is the dimension of $f'(x)$?

Exercise 1.5.5 Verify that the ODE (1.5) is dimensionally consistent. Then verify that the solution (1.6) is also dimensionally correct. In particular, is the argument of the exponential function dimensionless?

Exercise 1.5.6 Recall the fish harvesting model of Section 1.3, and in particular the ODE (1.10). The variable t in that equation is time, but u has no obvious dimension. Let us take $[u] = N$, where N denotes the dimension of “population.” (Although we could consider u as dimensionless since it simply counts how many fish are present, in other contexts we’ll encounter later it can be beneficial to think of $u(t)$ as having a specific dimension.) If $[u] = N$, then in the model leading to the ODE (1.10), what is the dimension of K ? What must be the dimension of r for the ODE to be dimensionally consistent?

Exercise 1.5.7 The orbital period P of an object in a circular orbit of radius r around a comparatively massive body like the earth, with mass m , is given by $P = 2\pi\sqrt{\frac{r^3}{Gm}}$. Verify that this formula is dimensionally correct.

Exercise 1.5.8 Find a plausible formula $v = G^a m^b r^c$ for the escape velocity v of a planet with mass m and radius r . Here G is the gravitational constant. Look up the correct formula and compare.

Exercise 1.5.9 Find a plausible formula for the period P of a pendulum as a function of its length ℓ , the mass m of the bob, and the earth's gravitational acceleration g , in the form $P = \ell^a m^b g^c$.

Exercise 1.5.10 Find a plausible formula for the speed of sound v in a gas as a function of its pressure P and density ρ , in the form $v = P^a \rho^b$. Note that pressure has the dimension of force per area.

Exercise 1.5.11 Find a plausible formula for the frequency f of a string's vibration in terms of its linear density λ , the tension τ in the string (which has the same units as force), and the length ℓ of the string, in the form $f = \lambda^a \tau^b \ell^c$.

Exercise 1.5.12 Dimensionless constants like 2 or π do not change value in physical formulas when the system of units is changed, and might therefore be considered very fundamental.

Four of the most important constants in physics are the speed of light c , Planck's constant \hbar , the charge e on an electron, and the Coulomb constant k_e in Coulomb's law. These constants have dimensions and approximate values (SI units) of

- $[c] = LT^{-1}$, value 299792458 meters per second
- $[\hbar] = ML^2T^{-1}$, value $1.054571817 \times 10^{-34}$ joule-seconds
- $[k_e] = ML^3T^{-2}Q^{-2}$, value 8.9875517923×10^9 kg-meters cubed per second squared per coulomb squared
- $[e] = Q$, value $e = 1.602176634 \times 10^{-19}$ coulomb

Show that the quantity $\alpha = \frac{e^2 k_e}{\hbar c}$ is dimensionless. This number is often called the *fine structure constant*. Compute its value using the data above (the value should be near 1/137.) What does this constant signify about our universe? Is it related to π or other fundamental mathematical constants? No one knows.

1.6 Modeling Projects

1.6.1 Project: Hang Time

This project is based on the SIMIODE Modeling Scenario "Hang Time" [122].

The phrase "hang time" is common in sports. An announcer in a football game may refer to the hang time for a punter's kick, or a basketball announcer may refer to the amount of time a player appears to hang in the air during a jump. In this modeling project we'll take a closer look at this phenomenon. Why do objects sometimes appear to hang in midair, even to the point that they seem to defy the law of gravity?

Consider an object of mass m in an idealized one-dimensional situation in which the object goes straight up and comes straight down. In particular, let's focus on a basketball player about to take a jump shot. The best professional basketball players have vertical jumps of 40 inches, possibly even higher (that's how high their hips or head go, above the standing position). Let's go with 1 meter, a bit over 39 inches, as the height a good professional player might jump.

We use t for time and $g = 9.81$ meters per second squared for gravitational acceleration. Let $y(t)$ denote the height of the player's hips during the jump, with $y = 0$ corresponding to the height of the player's hips when standing (so all vertical displacements are relative to this) with $y > 0$ corresponding to upward displacement. If $t = 0$ is the time the jump starts (ignore $t < 0$ when the player may crouch before jumping) then the appropriate initial condition is $y(0) = 0$.

Modeling Exercise 6.1.1 Newton's Second Law of Motion is $F = ma$, where a is the acceleration

of an object of mass m and F is the sum of all forces acting on the object. In this case a is the vertical acceleration of the player. Express a in terms of $y(t)$.

Modeling Exercise 6.1.2 If the only force acting on the player is gravity, what is F in $F = ma$? Hint: be careful with the sign—make sure F points downward.

Modeling Exercise 6.1.3 Put Modeling Exercises 6.1.1 and 6.1.2 together to find a second-order differential equation for $y(t)$. One initial condition is $y(0) = 0$. Take the other as $y'(0) = v_0$, where v_0 is some (as yet unknown) initial velocity the player gets from crouching and jumping.

Modeling Exercise 6.1.4 Find a general solution to the differential equation in Modeling Exercise 6.1.3 by integrating twice, as was done in Section 1.4. Then find the particular solution that satisfies the initial conditions. Check your work by making sure your solution satisfies the ODE and initial conditions. Hint: the solution should be quadratic in t and should involve v_0 .

Modeling Exercise 6.1.5 Suppose that $y(t)$ attains a maximum value of $y(t_1) = 1$ meter for some unknown time t_1 (when the player attains peak altitude). Show that this yields the equation

$$v_0 t_1 - \frac{1}{2} g t_1^2 = 1 \quad (1.23)$$

in SI (metric) units, so the 1 on the right side in (1.23) signifies 1 meter. Verify that all terms in (1.23) have the same dimension, namely length.

Modeling Exercise 6.1.6 What is $y'(t_1)$ equal to, if t_1 is the time at which the player attains maximum altitude? Use this to find a second equation relating the unknowns t_1 and v_0 . Then use this equation along (1.23) from Modeling Exercise 6.1.5 to find v_0 and t_1 . Work in SI units with $g = 9.81$ meters per second squared.

Modeling Exercise 6.1.7 How long does the entire jump last? What percentage of the total jump time is spent in the top 25 percent of the jump? How might this explain why the player seems to hang near the top of the jump?

1.6.2 Project: Money Matters

Almost everyone, at some point, joins the workforce, works for a period of time, then retires. It is of course essential to plan for retirement, and to save for that day. How much should you be saving throughout your career, and how should you invest it?

As an illustration, let's say you start work at age 22 and work until age 67. Let us use time $t = 0$ to indicate age 22, with t measured in years, so $t = 45$ years is retirement time. Like most people, as you progress in your career you earn promotions, responsibility, and more and more money. Suppose your pre-tax income at time t is given by

$$p(t) = 50000e^{t/45} \quad (1.24)$$

dollars per year (so you make \$50,000 per year starting at age 22, typical for college graduates in 2021 according to [53]). Note that the argument $t/45$ of the exponential function in (1.24) is dimensionless, since t has the dimension time, as does 45 (years). We can write $[p] = V$ with V as the dimension of value. For simplicity let's ignore inflation in this first analysis. Suppose you diligently save 10 percent of your income each year throughout your career, which is harder than it sounds. That is, you put away money at a rate of $0.1p(t)$ dollars per year. As a supremely conservative investor, you invest your money in a bank account that pays no interest. Let $S(t)$ denote the amount of money you've saved at time t .

Modeling Exercise 6.2.1 Explain why $S(t)$ obeys the ODE

$$S'(t) = 0.1p(t), \quad (1.25)$$

with $p(t)$ given by (1.24). Then find a general solution to this ODE. What is the dimension of S ? Of p ? What is the dimension of the constant 0.1?

Modeling Exercise 6.2.2 At age 22 you inherit \$100,000 tax-free from your grandparents, to start you on the path to retirement. It seems you're set for life. What initial condition for $S(0)$ is appropriate?

Modeling Exercise 6.2.3 Find the particular solution to (1.25) that satisfies the initial condition from Modeling Exercise 6.2.2.

Modeling Exercise 6.2.4 How much money will you have saved at retirement (what is $S(45)$)? If social security is defunct at this time and you live to age 90, how much money do you have on which to live each year? Will that support the lifestyle to which you will have become accustomed?

1.6.3 Project: Ant Tunneling

This project is based on the SIMIODE Modeling Scenario “Ant Tunnel Building” [120]; see also [118].

If a well-meaning relative ever purchased an ant colony for you when you were in grade school, or even if you just watched an ant hill on a summer day, you've noticed that ants are extremely industrious, and tireless tunnel builders. How long does it take an ant to build a single tunnel? This seems like an interesting modeling question. To address the issue we need to narrow the scope of the problem, simplify, and identify some terms and parameters.

To begin, let's define a few crucial variables. Consider a single ant digging a tunnel into a hillside as illustrated in Figure 1.4. (The ant drawing was provided by Isaac H. All.) Let x denote the current length, in feet, of the tunnel that the ant is digging. Let $T(x)$ be the time, in hours, it has taken the ant to build the tunnel of length x .

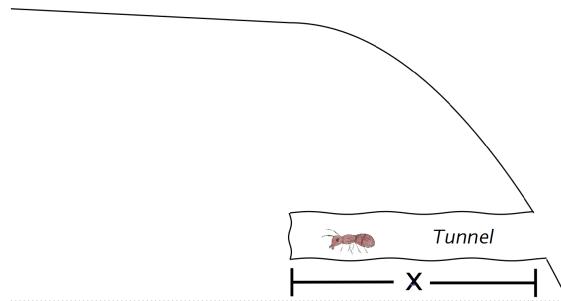


Figure 1.4: Illustration of an ant building a tunnel. Here x is the current length of the tunnel and $T(x)$ is the time it has taken the ant to build the tunnel of length x . Ant drawing provided by Isaac H. All.

Modeling Exercise 6.3.1 Write down several candidate functions for $T(x)$ and give one or two statements in each one's defense, and one or two statements against each.

Modeling Exercise 6.3.1 should convince you that jumping right to a defensible formula for $T(x)$ can be hard. So, instead of going after $T(x)$ directly, let us examine Figure 1.5, a depiction of the essential aspects of the situation. We'll make some assumptions that reflect the relevant geometry and physics, and might also make the model simpler to formulate. Specifically, when an ant digs a tunnel, the ant must extend the tunnel incrementally, by removing soil between coordinates x and $x + h$, carrying this soil back to the tunnel entrance, and then returning to remove

more soil. Let us consider how long it takes the ant to complete the extension of the tunnel from length x to length $x + h$.

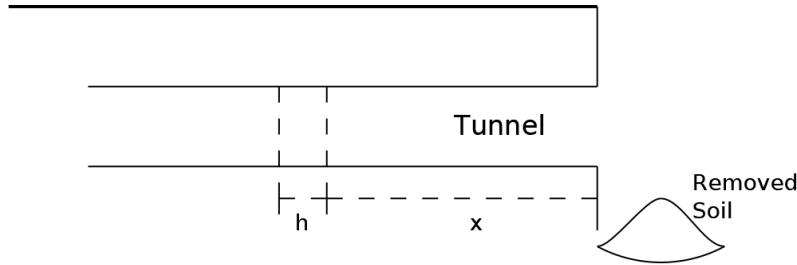


Figure 1.5: A useful diagram for modeling the time it takes to extend a small section of the ant tunnel from distance x to $x + h$.

Modeling Exercise 6.3.2 On the right side of (1.26), we seek an expression for how long it would require the ant to take a short length h of soil and then carry it a distance x to the mouth of the tunnel.

$$T(x+h) - T(x) = \underline{\hspace{2cm}} \quad (1.26)$$

Notice that $T(x+h) - T(x) \neq T(h)$, for $T(h)$ represents the time it takes to extend the tunnel a distance h from the mouth of the tunnel (at $x=0$), while $T(x+h) - T(x)$ includes this time plus the time it takes to carry the soil to the mouth of the tunnel.

Below are a few possibilities for the right side of (1.26). Defend or reject each and offer your reasons. Modify one or two to improve them. When trying to reject a model consider some extreme cases and see if the model makes sense, e.g., $h=0$ or $x=0$, or either h or x very large.

- (a) $T(x+h) - T(x) = x + h$
- (b) $T(x+h) - T(x) = x - h$
- (c) $T(x+h) - T(x) = x^h$
- (d) $T(x+h) - T(x) = x \cdot h$
- (e) $T(x+h) - T(x) = h^x$
- (f) $T(x+h) - T(x) = c$

Modeling Exercise 6.3.3 Convert your model difference equation (1.26) to a related differential equation with appropriate initial conditions. It may be helpful to consider the familiar expression $(T(x+h) - T(x))/h$ and what happens as h approaches 0.

Modeling Exercise 6.3.4 Solve the differential equation you create in Modeling Exercise 6.3.3 for $T(x)$. Hint: what initial condition $T(0)$ will you use?

Modeling Exercise 6.3.5 Use your solution from Modeling Exercise 6.3.4 to determine how much longer it takes to build a tunnel that is twice as long as an original tunnel of length L . What would some of the original function models you set forth in (a)-(f) have told you here?

Modeling Exercise 6.3.6 Suppose two ants dig from opposite sides of the sand hill, directly toward each other along the same straight line. How would this alter the total time for digging the tunnel?

Modeling Exercise 6.3.7 Of course, the same principles can be applied to model tunnel building for engineers. If we were considering Modeling Exercise 6.3.6 as related to engineering the construction of a long tunnel of length L , outline some of the issues we should be aware of when having two crews (one from each end of the tunnel) working on it.

2. First-Order Equations

In this chapter we'll look at techniques for solving or otherwise analyzing first order ODEs of the form $u'(t) = f(t, u(t))$. How we proceed will depend on the function f , which determines the nature of the differential equation (1.14). We will examine solution techniques for two common classes of first-order ODEs: those that are linear and those that are separable; some ODEs are both. However, many first-order equations fall into no particular category and have no analytical solution, at least not in terms of the familiar elementary functions. In these cases we may turn to qualitative methods that illuminate how solutions behave when an explicit solution cannot be obtained. These qualitative methods are of value even when explicit solutions do exist, and can give a lot of geometric insight into the nature of solutions. The ODEs developed in Chapter 1 and some new models we introduce in this chapter will serve as testbeds for these techniques. In Chapter 3 we will consider methods for numerically approximating solutions to ODEs when an analytical solution cannot be found.

2.1 First-Order Linear Equations

We begin with some definitions and terminology.

Definition 2.1.1 — Linear First-Order ODEs. A first-order ODE that can be written in the form

$$u'(t) = g(t) + h(t)u(t) \tag{2.1}$$

for some functions $g(t)$ and $h(t)$ is **linear**, and otherwise the ODE is **nonlinear**. If $g(t)$ is the zero function then the ODE is **homogeneous**, and otherwise the ODE is **nonhomogeneous**. If $h(t)$ is a constant function then (2.1) is a **constant-coefficient** ODE, and otherwise (2.1) is a **variable-coefficient** ODE.

Remark 2.1.1 At times we will write ODEs without the independent variable explicitly attached to the unknown function. That is, we will write $u' = f(t, u)$ instead of $u'(t) = f(t, u(t))$. Sometimes this makes it easier to see the structure of the equation.

■ **Example 2.1** Let us classify the ODEs (1.3), (1.5), (1.10), and (1.12) from the last chapter as either linear or nonlinear, and if the ODE is linear, classify it as homogeneous or nonhomogeneous, and constant- or variable-coefficient.

- The Hill-Keller ODE (1.3), $v'(t) = P - kv(t)$, is linear, for it is of the form (2.1), with v as the unknown function, $g(t) = P$, and $h(t) = -k$. It is also constant-coefficient and, if $P \neq 0$, it is nonhomogeneous.
- The model for intracochlear drug delivery (1.5), $u'(t) = rc_1 - ru(t)/V$, is linear, with $g(t) = rc_1$ and $h(t) = -r/V$. This is also a constant-coefficient nonhomogeneous ODE.
- The logistic equation (1.10) and the harvested logistic equation (1.12) are nonlinear equations. In each case the right side of the ODE is a quadratic function of u .

■

In Section 1.4.3 certain ODEs were solved by directly integrating both sides of the equation. Equations like (2.1) also require integration to solve, but the application is not direct: integrating both sides of (2.1) with respect to t leads to

$$u(t) = \int (g(t) + h(t)u(t)) dt + C,$$

but the right side above involves the very function $u(t)$ that we don't know, and so we can't evaluate the integral. Compare this to (1.16) from the last chapter; direct integration succeeds there precisely because $h(t)$ was the zero function, so we only had to integrate $g(t)$, which was known.

2.1.1 Example: Solving the Hill-Keller Equation as a Linear ODE

Before showing the general technique for solving linear first-order ODEs, let's consider the Hill-Keller equation (1.3) from Chapter 1, to demonstrate how the solution technique for linear equations works in a specific case. Recall that k in that equation is a constant.

■ **Example 2.2** We will find a general solution to the Hill-Keller ODE (1.3) and the particular solution that satisfies $v(0) = 0$. Recall that the appropriate initial condition for the data in Table 1.1 is really $v(0.165) = 0$, but let's use $v(0) = 0$ for our first example. The approach is to rearrange the ODE so that integrating actually leads somewhere.

The first step is to write the Hill-Keller ODE (1.3) as

$$v'(t) + kv(t) = P. \quad (2.2)$$

This is completely equivalent to the original version. The next step, which is not at all obvious, is to multiply both sides of (2.2) by the function e^{kt} to obtain an equivalent equation

$$e^{kt}(v'(t) + kv(t)) = Pe^{kt}. \quad (2.3)$$

This apparently pointless operation has prepared the ODE so that integrating both sides is possible; the reason for this step will be discussed below. The quantity e^{kt} is the **integrating factor** for this ODE. You can check using the product rule that the identity

$$\frac{d}{dt}(e^{kt}v(t)) = e^{kt}(v'(t) + kv(t)) \quad (2.4)$$

holds for any function $v(t)$, and that the right side of (2.4) is precisely the left side of (2.3). Using (2.4) to replace the left side of (2.3) yields

$$\frac{d}{dt}(e^{kt}v(t)) = Pe^{kt}. \quad (2.5)$$

The next step is to integrate both sides of (2.5) with respect to t , which is now possible even though we don't know $v(t)$. The integration undoes the derivative on the left in (2.5), and the antiderivative for Pe^{kt} on the right is easy to compute: it is Pe^{kt}/k . Equation (2.5) yields

$$e^{kt}v(t) + C_1 = \frac{P}{k}e^{kt} + C_2, \quad (2.6)$$

where C_1 and C_2 are arbitrary constants of integration. By *constant* we mean that they do not depend on the independent variable t . Notice that in (2.6) all derivatives have disappeared—finding $v(t)$ is now an algebra problem.

Reading Exercise 2.1.1 Differentiate both sides of (2.6) with respect to t and verify that this yields (2.3). Then verify that division by e^{kt} takes us right back to (2.2) and (1.3). This shows that all steps are reversible and so (1.3) and (2.6) are equivalent: any function $v(t)$ that satisfies one equation will satisfy the other.

The last step to computing a general solution to (1.3) is to solve (2.6) for $v(t)$. We can lump C_1 and C_2 in (2.6) together as $C_2 - C_1$ on the right side of the equation, and then note that the difference of two arbitrary constants is itself an arbitrary constant. Define $C = C_2 - C_1$ so that (2.6) becomes

$$e^{kt}v(t) = \frac{P}{k}e^{kt} + C. \quad (2.7)$$

Divide both sides of (2.7) by e^{kt} to find

$$v(t) = \frac{P}{k} + Ce^{-kt} \quad (2.8)$$

where C is an arbitrary constant. This is a general solution to the Hill-Keller ODE (1.3). In moving from (1.3) to (2.2) and on to (2.8), at each step we used simple algebra and antiderivatiation. Moreover, each step was reversible—you can go from (2.8) back to (1.3) using algebra and differentiation. This makes it clear that any solution to (1.3) must be of the form dictated by (2.8) for some choice of C , and conversely, (2.8) provides a solution to (1.3) for any C . That's why (2.8) is called a general solution.

To obtain the desired initial condition, choose C in (2.8) so that $v(0) = 0$. Substituting $t = 0$ into the right side of (2.8) and setting the result to zero leads to $P/k + C = 0$, with solution $C = -P/k$. In (2.8) this yields the particular solution with the required initial condition $v(0) = 0$, namely

$$v(t) = \frac{P}{k} - \frac{P}{k}e^{-kt}.$$

As a quick sanity check, note that P has the dimension of acceleration ($[P] = LT^{-2}$) and k has the dimension of reciprocal time ($[k] = T^{-1}$), so the quantity P/k has the dimension of velocity, as expected. Moreover, the argument $-kt$ of the exponential above is dimensionless. This provides a coarse check on the correctness of our computations. ■

Reading Exercise 2.1.2 Adjust C in (2.8) to obtain the initial condition $v(0.165) = 0$, as suggested by the data in Table 1.1. The constant C will depend on k .

Reading Exercise 2.1.3 Let $k = 1$ (units reciprocal seconds) in the solution from Reading Exercise 2.1.2, with $P = 11$ meters per second per second. Plot $v(t)$ as a function of t on the range $0.165 \leq t \leq 10$. Does this seem like a reasonable velocity profile for a sprinter? How does changing k affect the graph? What value of k might give reasonable agreement with Bolt's race data? Hint: his top speed was 12.2 meters per second. In the next chapter we'll look at more sophisticated ways to estimate k .

2.1.2 A General Procedure for Solving Linear ODEs

The procedure used to solve the Hill-Keller ODE works for more general linear equations. To solve the ODE (2.1) take the following steps.

1. **Rewrite the ODE:** Write the ODE in the form

$$u'(t) - h(t)u(t) = g(t). \quad (2.9)$$

The left side of (2.9) looks a little like a derivative that might come out of the product rule, and it is with a bit of help; see the next step.

2. **Multiply by the Integrating Factor:** Let $H(t)$ be any antiderivative for $h(t)$, so $H'(t) = h(t)$. Multiply both sides of (2.9) by $e^{-H(t)}$ to obtain

$$e^{-H(t)}(u'(t) - h(t)u(t)) = e^{-H(t)}g(t). \quad (2.10)$$

The quantity $e^{-H(t)}$ is called the **integrating factor** for this ODE. The left side of (2.10) is an exact derivative, since $\frac{d}{dt}(e^{-H(t)}u(t)) = e^{-H(t)}(u'(t) - h(t)u(t))$, so (2.10) can be written as

$$\frac{d}{dt}(e^{-H(t)}u(t)) = e^{-H(t)}g(t). \quad (2.11)$$

How in the world would anyone come up with the inspiration of multiplying the ODE by $e^{-H(t)}$? See Exercise 2.1.12.

3. **Integrate:** Integrate both sides of (2.11) with respect to t to obtain

$$e^{-H(t)}u(t) = \int e^{-H(t)}g(t)dt + C. \quad (2.12)$$

Here C is the difference of the two arbitrary constants obtained when we integrate both sides of (2.11) with respect to t , just as we did to obtain (2.7). The right side of (2.12) could be evaluated if we had specific choices for g and h , but for the moment the integral must be left unevaluated.

4. **Solve:** Multiply both sides of (2.12) by $e^{H(t)}$ to obtain a general solution

$$u(t) = e^{H(t)} \int e^{-H(t)}g(t)dt + Ce^{H(t)}. \quad (2.13)$$

Every solution to (2.1) is of the form (2.13) for some choice of the constant C . A warning: the $e^{-H(t)}$ in front of the integral in (2.13) is not constant and cannot be moved inside the integral.

5. **Obtain the Initial Condition:** If an initial condition is given, adjust C in (2.13) as required to obtain the initial condition.

Note that (2.13) provides a general solution to the ODE (2.1) for any choice of g and h . However, it only gives the solution explicitly if we can find an antiderivative H for h and then evaluate the integral in (2.13). In the rather common constant-coefficient case where $h(t) = A$, a constant, the choice $H(t) = At$ works, with integrating factor e^{At} .

■ **Example 2.3** Let us find a general solution to the variable-coefficient first-order ODE $u'(t) = u(t)/t + t^2$ and the solution with initial condition $u(1) = 2$. We will restrict our attention to the domain $t > 0$ (we'll discuss issues concerning the domain of the solution to an ODE later). First write the ODE as $u'(t) - u(t)/t = t^2$, so here $h(t) = 1/t$ and $g(t) = t^2$. An antiderivative for $1/t$ is $\ln|t|$, but because we are only interested in $t > 0$, we'll drop the absolute values. The integrating factor is then $e^{-\ln(t)} = 1/t$. Multiply both sides of the ODE by the integrating factor $1/t$ to obtain

$$\frac{u'(t)}{t} - \frac{u(t)}{t^2} = t.$$

The left side above is $\frac{d}{dt}(u(t)/t)$, so

$$\frac{d}{dt}\left(\frac{u(t)}{t}\right) = t.$$

Integrate both sides with respect to t and lump all constants of integration together on the right to find

$$\frac{u(t)}{t} = \frac{t^2}{2} + C.$$

Multiply both sides by t to obtain a general solution

$$u(t) = \frac{t^3}{2} + Ct.$$

To find a solution with $u(1) = 2$, substitute the initial time $t = 1$ into this general solution and find that $1/2 + C = 2$ is required, so $C = 3/2$. The solution with $u(1) = 2$ is thus $u(t) = t^3/2 + 3t/2$. ■

Reading Exercise 2.1.4 In Step 2 of the integrating factor method, it seems that $H(t)$ can be any antiderivative for $h(t)$. Since antiderivatives are determined only up to any additive constant, it should be possible to add any constant to H and still obtain a valid result. Verify this by redoing Example 2.3 but using the antiderivative $H(t) = \ln(t) + A$ where A is an arbitrary constant. Verify that A cancels out of the computation.

Reading Exercise 2.1.5 Verify that when $h(t)$ is the zero function the integrating factor procedure above (and in particular (2.13)) gives the same result for the solution of (2.9) as was obtained in (1.17) of Chapter 1.

Reading Exercise 2.1.6 Work through the integrating factor solution in the special case that $g(t)$ is the zero function, to show that a general solution to $u'(t) = h(t)u(t)$ is

$$u(t) = Ce^{H(t)},$$

where $H(t)$ is an antiderivative for $h(t)$.

2.1.3 Some Common First-Order Linear Models

In this section we present three additional situations commonly modeled with first-order ODEs. Each of the ODEs is linear and so can be solved using the integrating factor approach.

Newton's Law of Cooling

It sounds like something from a murder mystery novel: a detective estimates the time of death of a person by measuring the temperature of the corpse. By using the fact that a living person has a nominal temperature of 98.6 degrees Fahrenheit and knowing the temperature of the environment in which the corpse was found, the amount of time since the person died can be estimated by taking into account the rate at which the body cools. The reality of determining the time of death isn't quite so simple, but body temperature is one factor that a medical examiner can use for this purpose [13].

More generally, let's consider the problem of modeling how an object changes temperature in response to its environment. We'll make several important assumptions.

- The object is small enough to have a single temperature throughout, at least to some approximation. If not then the temperature is a function of position inside the object and we find ourselves in the realm of partial differential equations, more than we want to deal with right now.

- The environment in which the object exists has a temperature that does not vary with position, but may vary with time. This is called the **ambient** temperature.
- The object does not generate any internal heat, but changes temperature solely because it loses or gains heat energy to or from the environment.

Temperature is a fundamentally new physical dimension that cannot be expressed in terms of mass, length, or time. We use the symbol Θ to denote the dimension of temperature. Let t denote time and $u(t)$ denote the temperature of the object at time t (remember, u is constant throughout the object), so $[u] = \Theta$. Let A denote the ambient temperature. For the moment let's assume that A is constant in time. Perhaps the simplest model for how the object's temperature changes in response to the environment is **Newton's law of cooling**, which posits that:

The rate at which an object's temperature changes is proportional to the difference between the object's temperature and the ambient temperature.

That's Newton's law of cooling in English. The language of ODEs makes it possible to state it much more concisely.

Reading Exercise 2.1.7

- If $u(t)$ denotes the temperature of an object, what is a mathematical expression for the rate at which the object's temperature changes?
- What is a mathematical expression for the difference between the object's temperature and the ambient temperature?
- Write an equation to express that the two quantities in (a) and (b) are proportional to each other. Use k for your constant of proportionality.

Reading Exercise 2.1.7 leads to the differential equation

$$u'(t) = k(u(t) - A).$$

This formulation clearly requires k to be negative, for if $u(t) > A$ (the object is hotter than the environment) then $u'(t) < 0$ (the object is cooling), while $u(t) < A$ (the object is colder than the environment) should require $u'(t) > 0$ (the object is warming). In view of Modeling Tip 1.1.1, we shall instead require that k is positive and put an explicit minus sign on the right side of the ODE to obtain Newton's law of cooling in the form

$$u'(t) = -k(u(t) - A). \quad (2.14)$$

Reading Exercise 2.1.8 Is the Newton cooling ODE (2.14) linear or nonlinear? If linear, is it constant- or variable-coefficient? Homogeneous or nonhomogeneous?

Reading Exercise 2.1.9 Given that $[u] = \Theta$, what is $[u']$? What must the dimension $[k]$ of k be?

■ **Example 2.4** Suppose an object with temperature $u(t)$ has initial temperature $u(0) = 98.6^\circ\text{F}$ in an environment with ambient temperature $A = 72^\circ\text{F}$. Let t denote time in hours. After three hours the object has temperature 94°F . Let us find the temperature of the object as a function of time t , and determine that time t at which $u(t) = 80^\circ\text{F}$.

To begin let us note that, as you will demonstrate in Exercise 2.1.4, the solution to the Newton cooling ODE (2.14) with initial condition $u(0) = u_0$ is

$$u(t) = A + (u_0 - A)e^{-kt}. \quad (2.15)$$

With $u_0 = 98.6$ and $A = 72$, $u(t)$ in (2.15) becomes $u(t) = 72 + 26.6e^{-kt}$. Given the information that $u(3) = 94$, (2.15) yields $94 = 72 + 26.6e^{-3k}$. Solving for k gives $k \approx 0.0633$ (units are reciprocal hours). Then

$$u(t) = 72 + 26.6e^{-0.0633t},$$

approximately. The equation $u(t) = 80$ then becomes $72 + 26.6e^{-0.0633t} = 80$, which we can solve for t to find $t \approx 18.98$ hours. ■

Salt Tank Problems

The next example is a generalization of the Intracochlear Drug Delivery model that we considered in Section 1.2.2. In **salt tank models** the quantity “salt” refers to any substance that is neither created nor destroyed in the course of the problem. Salt tank problems are a type of compartmental model. (Such a model was the focus of Exercise 1.4.2.)

Consider the following scenario: a tank contains 200 liters of water in which 3 kg of salt is dissolved at time $t = 0$. A pipe carries water into the tank at a rate of 2 liters per second; the incoming water contains salt at a concentration of 0.1 kg per liter. The well-stirred mixture leaves the tank through an outflow pipe at a rate of 2 liters per second. The situation is illustrated in Figure 2.1. The goal is to determine the amount of salt (kg) dissolved in the tank water at any time t where $t > 0$, and to determine how this amount of salt behaves in the limit as $t \rightarrow \infty$.

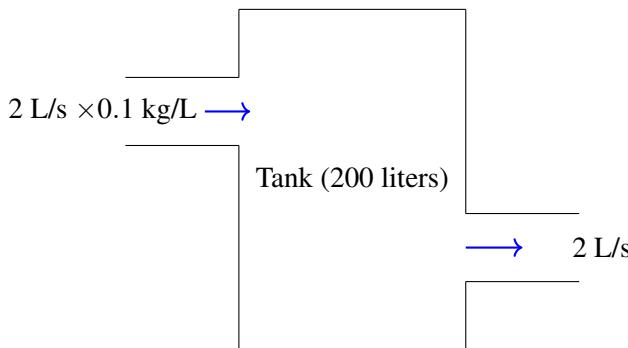


Figure 2.1: An illustration of a typical salt tank compartmental model with 200 liter tank, inflow pipe carrying 2 liters of fluid per second at a concentration of 0.1 kg of salt per liter, outflow pipe carrying out 2 liters per second of well-stirred solution.

The essential modeling here is similar to that of Section 1.2.2, and in particular (1.4). The conservation of salt implies that the rate at which the amount of salt in the tank is changing equals the rate salt enters minus the rate salt leaves, with all rates quantified as mass per time. We will use t for time in seconds and $x(t)$ for the amount of salt (kg) in the tank at time t .

To find the rate at which salt enters the tank at the inflow pipe, note that 2 liters of fluid enters each second and each liter contains 0.1 kg of salt. The rate at which salt enters is thus

$$2 \left(\frac{\text{liters}}{\text{second}} \right) \times 0.1 \left(\frac{\text{kg}}{\text{liter}} \right) = 0.2 \left(\frac{\text{kg}}{\text{second}} \right).$$

The rate at which salt is leaving the tank can be found by noting that since the tank is well-stirred, the concentration of salt in the tank is spatially uniform, and so each liter of tank fluid at time t contains $x(t)/200$ kg of salt. Precisely 2 liters of this solution exit the tank each second, thus the rate at which salt is leaving the tank is

$$2 \left(\frac{\text{liters}}{\text{second}} \right) \times \frac{x(t)}{200} \left(\frac{\text{kg}}{\text{liter}} \right) = \frac{x(t)}{100} \left(\frac{\text{kg}}{\text{second}} \right).$$

The rate at which the amount of salt in the tank is changing is the rate that salt enters minus the rate that salt exits, so from the analysis above we get

$$\frac{dx}{dt} = 0.2 - \frac{x(t)}{100}, \tag{2.16}$$

in which all terms have units of kilograms per second. At $t = 0$ the tank contains 3 kg of salt, so the initial condition is $x(0) = 3$.

Equation (2.16) is a linear first-order differential equation. In Exercise 2.1.5 at the end of this section, you will show that the solution is

$$x(t) = 20 - 17e^{-t/100}. \quad (2.17)$$

From (2.17) it follows that $\lim_{t \rightarrow \infty} x(t) = 20$ kg, since the exponential term $e^{-t/100}$ decays to zero.

RC Circuits

In the following example we use Q to denote the physical dimension of charge; Q is independent of the previously introduced dimensions of mass, length, time, and temperature. For a more detailed description and derivation of the equations that govern basic circuits, see Appendix C.

Consider a simple **RC series circuit** with voltage source $V(t)$, resistor R , and capacitor C , as illustrated in Figure 2.2. Let $I(t)$ denote the current in the circuit, with $I > 0$ indicating clockwise current, and let $q(t)$ denote the charge on the capacitor (with the positive side $+q$ as indicated, $-q$ on the other side). Then $[q] = Q$ and since the current I is the amount of charge passing through the wire per unit time, $[I] = QT^{-1}$. If we start at the negative side of the voltage source and step around the RC loop in the clockwise direction, we gain $V(t)$ volts over the source, then have a voltage drop $-RI(t)$ across the resistor of and a voltage drop across the capacitor of $-q(t)/C$, and return to the negative side of the voltage source. From Kirchhoff's voltage law it follows that

$$V(t) - q(t)/C - RI(t) = 0. \quad (2.18)$$

Since the charge entering the positive side of the capacitor flows in through the wire from the voltage source it follows that $q'(t) = I(t)$. Then (2.18) can be written as

$$Rq'(t) + q(t)/C = V(t). \quad (2.19)$$

This is a linear, first-order, nonhomogeneous ODE for $q(t)$, the charge on the capacitor. A typical initial condition might be of the form $q(0) = q_0$ (often $q(0) = 0$).

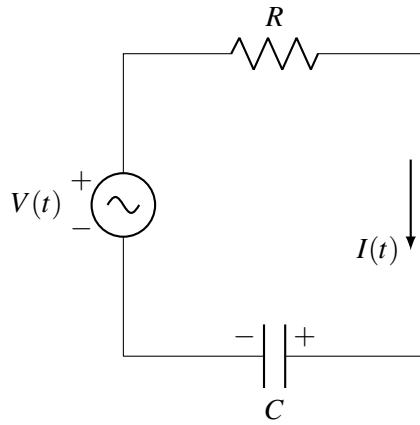


Figure 2.2: Single-loop RC series circuit with voltage source $V(t)$ and current $I(t)$ in the direction indicated.

Reading Exercise 2.1.10 As shown in Appendix C, voltage has the dimensions of work per charge, which is $[V] = ML^2T^{-2}Q^{-1}$. Use this with (2.19) to show that $[C] = M^{-1}L^{-2}T^2Q^2$ and $[R] = ML^2T^{-1}Q^{-2}$.

■ **Example 2.5** Consider the RC circuit of Figure 2.2 with $V(t) = 5$ volts, $R = 100$ ohms, and $C = 10^{-6}$ farad. At $t = 0$ the capacitor is uncharged. Let us find the charge $q(t)$ on the capacitor as a function of time t , as well as the current $I(t)$ in the circuit and the voltage V_R across the resistor.

With these values the ODE (2.19) becomes

$$100q'(t) + 10^6 q(t) = 5$$

with initial condition $q(0) = 0$. To find a general solution, the integrating factor approach is appropriate, and yields $q(t) = 1/200000 + Ce^{-10000t}$. The initial condition $q(0) = 0$ gives

$$q(t) = \frac{1}{200000} - \frac{e^{-10000t}}{200000}.$$

The capacitor thus charges to a limiting value of $1/200000$ farad in a few ten-thousands of a second. Thus current through the circuit is

$$I(t) = q'(t) = e^{-10000t}/20$$

amps; the current is initially $1/20$ amp and drops to zero as the capacitor charges. The potential across the resistor R can be computed using Ohm's law $V_R = IR$ and is

$$V_R = 5e^{-10000t}$$

volts. ■

2.1.4 Exercises

Exercise 2.1.1 Find a general solution to each linear ODE, and then find the specific solution with the given initial condition by using the integrating factor technique from Section 2.1.2.

- (a) $u'(t) = u(t) + 3$, $u(0) = 3$
- (b) $u'(t) = 2u(t) + 4$, $u(0) = 0$
- (c) $u'(t) = -3u(t) + 3$, $u(0) = 5$
- (d) $u'(t) = -3u(t) + 9t$, $u(0) = 5$
- (e) $u'(t) = u(t) + 2\sin(t)$, $u(0) = 1$
- (f) $u'(t) = -4u(t) + e^t$, $u(0) = 2$
- (g) $u'(t) = tu(t) + t$, $u(0) = 2$
- (h) $u'(t) = u(t)/t + 2$, $u(1) = 3$
- (i) $u'(t) = \sin(t)u(t) + \sin(t)$, $u(0) = 4$
- (j) $u'(t) = au(t) + b$, $u(0) = u_0$, where a, b , and u_0 are constants

Exercise 2.1.2 Solve (1.5) from the previous chapter with $u(0) = 0$, and so verify (1.6).

Exercise 2.1.3 Many physical processes, e.g., radioactive decay, are governed by an ODE of the form

$$u'(t) = -ku(t),$$

where k is some positive constant.

- (a) If t denotes time, what is the dimension of k ?
- (b) Find a general solution to this ODE with initial condition $u(0) = u_0$, and show that the solution approaches zero as $t \rightarrow \infty$ for any u_0 .

- (c) Show that the amount of time Δt necessary for the solution to decrease by half is given by $\Delta t = \ln(2)/k$. Hint: solve $u(t + \Delta t) = u(t)/2$, and note the result does not depend on t . The quantity Δt is called the *half-life* of whatever process is governed by this ODE. .

Exercise 2.1.4 Solve the Newton cooling ODE (2.14) with initial condition $u(0) = u_0$ by using the integrating factor approach, and so demonstrate that (2.15) is correct.

Exercise 2.1.5 Solve the compartmental salt tank ODE (2.16) with initial condition $x(0) = 3$, by using the integrating factor approach, and so demonstrate that (2.17) is correct.

Exercise 2.1.6 A body is found at a certain time and has a temperature of 92.6 degrees Fahrenheit in an environment with ambient temperature 70 degrees. Two hours later the body has a temperature of 90 degrees Fahrenheit. The flu was going around and it was believed the victim was on her way to the drugstore because her roommate said she had a temperature of 102.4 degrees Fahrenheit. If Newton's law of cooling holds, estimate when the person died, relative to the time the body was found. Hint: call the time the body is found $t = 0$ and solve $u'(t) = -k(u(t) - A)$ with $A = 70$ and initial condition $u(0) = 92.6$. Then use $u(2) = 90$ to find k , and from that figure out at what time $u(t)$ equaled 102.4.

Exercise 2.1.7 A tank contains 400 liters of pure water. At time $t = 0$ water containing 0.2 kg of salt per liter begins to flow into the tank at a rate of 4 liters per minute. The well-stirred mixture flows out of the tank at 4 liters per minute. Formulate an appropriate ODE for $x(t)$, the amount of salt in the tank at time t , with an initial condition. Solve the ODE using the integrating factor approach. What is the limiting amount of salt in the tank?

Exercise 2.1.8 A tank contains 400 liters of pure water at time $t = 0$ minutes, when water containing 0.2 kg of salt per liter begins to flow into the tank at a rate of 4 liters per minute. The well-stirred mixture flows out of the tank at 5 liters per minute (so the tank slowly empties.)

- Find the volume $V(t)$ of liquid in the tank as a function of time.
- Formulate an appropriate ODE for $x(t)$, the amount of salt in the tank at time t , with an initial condition. Hint: for the rate at which salt is exiting the tank at time t , use the fact that $x(t)$ kg of salt are uniformly distributed among $V(t)$ liters of fluid in the tank.
- Solve the ODE using the integrating factor approach and graph the solution. How much salt is in the tank at any time?
- When is the amount of salt in the tank maximized, and how much salt is in the tank at this time?

Exercise 2.1.9 A 5 kilogram rock is dropped from a bridge that is 200 meters above the surface of a river that is 10 meters deep. As the rock falls through the air, the force of resistance on the rock (in newtons) is equal to 0.8 times the velocity in meters per second (the units on 0.8 are newtons per meter per second). As the rock falls through the water the resistance in newtons is equal to 5.0 times the rock's velocity (same units as the 0.8 coefficient). Acceleration due to gravity is 9.8 meters per second squared.

- How long after the rock is released will it hit the water?

- (b) How long after the rock hits the water will it hit the bottom of the river?
 (c) What is the rock's velocity when it hits the bottom?

Exercise 2.1.10

- (a) In the case that $V(t) = V_0$ is constant, use the integrating factor approach to show that the solution to (2.19) with $q(0) = 0$ is given by

$$q(t) = CV_0(1 - e^{-t/(RC)}).$$

- (b) What is the limiting value of the charge on the capacitor as $t \rightarrow \infty$?
 (c) Show that the product RC has the dimension of time. Hint: look back at Reading Exercise 2.1.10. The product RC is called the **RC time constant** for this circuit.
 (d) How long does it take the charge on the capacitor to attain 99 percent of the limiting value in part (b)? Compare this to the rule of thumb that a capacitor takes $5RC$ time units to effectively reach full charge.

Exercise 2.1.11 Capacitors can be used in circuits to act as filters that allow sinusoidal input voltages in a given frequency range to pass through while attenuating (reducing) other frequencies. As an example, consider the circuit of Figure 2.3 (similar to the circuit of Figure 2.2, but with additional leads to measure the voltage $V_C(t)$ across the capacitor). Suppose the input voltage is given by $V(t) = V_0 \sin(\omega t)$ for some radial frequency ω and amplitude V_0 . Take $R = 10^6$ ohms and $C = 10^{-6}$ farads.

- (a) Show that a general solution to (2.19) in this case is

$$q(t) = De^{-t} + A \cos(\omega t) + B \sin(\omega t)$$

for an arbitrary constant D , with $A = -\frac{V_0 \omega}{10^6(\omega^2+1)}$ and $B = \frac{V_0}{10^6(\omega^2+1)}$.

- (b) Use $q(t) = CV_C(t)$ with $V_C(t) = V_C^+(t) - V_C^-(t)$ (the voltage over the capacitor C) to find $V_C(t)$. Argue that for large time t (about $t = 5$ seconds) the quantity $V_C(t)$ is, for practical purposes, given by

$$V_C(t) = \frac{-V_0 \omega}{\omega^2 + 1} \cos(\omega t) + \frac{V_0}{\omega^2 + 1} \sin(\omega t).$$

- (c) A function of the form $f(t) = A_1 \cos(\omega t) + A_2 \sin(\omega t)$ is sinusoidal with frequency ω and amplitude $\sqrt{A_1^2 + A_2^2}$ (see Exercise 4.2.9). Use this to show that the amplitude of $V_C(t)$ is given by

$$\text{amplitude of } V_C(t) = \frac{V_0}{\sqrt{\omega^2 + 1}}.$$

- (d) We can consider $V(t)$ as the input to the circuit and $V_C(t)$ as the output, perhaps to some other portion of the circuit. Take $V_0 = 1$ and plot the amplitude $\frac{V_0}{\sqrt{\omega^2 + 1}}$ of V_C as a function of ω . How does this amplitude compare to the amplitude V_0 of $V(t)$ for low frequencies $\omega \approx 0$? How does the amplitude of $V_C(t)$ compare to V_0 at high frequencies, as $\omega \rightarrow \infty$? Can you see why this circuit is called a *low-pass filter*?

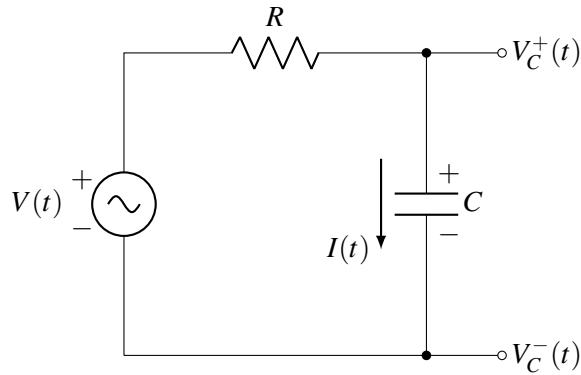


Figure 2.3: Single loop RC series circuit with voltage source $V(t)$, current $I(t)$ in the direction indicated.

Exercise 2.1.12 In trying to solve a linear ODE written in the form

$$u'(t) - h(t)u(t) = g(t), \quad (2.20)$$

it's not at all obvious how anyone might hit upon the idea of multiplying both sides of (2.20) by $e^{-H(t)}$ where $H'(t) = h(t)$. One way you might arrive at this inspiration is to note that the left side of (2.20) looks a little like the result of applying the product rule for derivatives to the product $w(t)u(t)$ for some function $w(t)$, namely

$$(w(t)u(t))' = w(t)u'(t) + w'(t)u(t), \quad (2.21)$$

although the left side of (2.20) and the right side of (2.21) aren't quite the same. But if we multiply the left side of (2.20) by an arbitrary function $w(t)$ we obtain $w(t)u'(t) - w(t)h(t)u(t)$, and comparison to the right side of (2.21) shows that the $w(t)u'(t)$ terms will match no matter what we choose for $w(t)$. If we can choose $w(t)$ so that $-w(t)h(t)u(t) = w'(t)u(t)$ then we're in business: the quantity $w(t)u'(t) - w(t)h(t)u(t)$ will be an exact derivative.

Show that the condition $-w(t)h(t)u(t) = w'(t)u(t)$ leads to the conclusion that $w(t) = e^{-H(t)}$, where $H'(t) = h(t)$. This is precisely the integrating factor in (2.10).

2.2 Separable Equations

The other common class of first-order ODEs that can often be solved analytically are those that are separable. We now give a precise definition.

Definition 2.2.1 — Separable First-Order ODEs. A first-order ODE that can be written in the form

$$u'(t) = g(t)h(u(t)) \quad (2.22)$$

for some functions g and h is said to be **separable**.

That is, an ODE $u' = f(t, u)$ is separable if $f(t, u) = g(t)h(u)$ for some functions g and h .

■ Example 2.6

- (a) The ODE $u'(t) = \sin(t)u^2(t)$ is separable since it can be written as $u' = f(t, u)$ with $f(t, u) = \sin(t)u^2$, which splits as $f(t, u) = g(t)h(u)$ with $g(t) = \sin(t)$ and $h(u) = u^2$.

- (b) The ODE $v'(t) = \sin(v(t))$ is separable since it can be written as $v' = f(t, v)$ with $f(t, v) = \sin(v)$, which splits as $f(t, v) = g(t)h(v)$ with $g(t) = 1$ and $h(v) = \sin(v)$.
- (c) The ODE $u'(t) = u(t) + t$ is not separable. To see this note that it can be written as $u' = f(t, u)$ with $f(t, u) = t + u$. It might seem obvious that writing $f(t, u) = g(t)h(u)$ is impossible for any choice of g and h , but to show this conclusively, suppose that $f(t, u) = g(t)h(u)$. Then $f(0, 0) = 0$ so that $g(0)h(0) = 0$ and either $g(0) = 0$ or $h(0) = 0$. From $f(1, 0) = g(1)h(0)$ and $f(0, 1) = g(0)h(1)$ conclude at least one of $f(1, 0)$ or $f(0, 1)$ equals zero, contradicting the fact that $f(1, 0) = f(0, 1) = 1$. ■

How will you know if a first-order ODE of the form $u' = f(t, u)$ is separable? It will almost always be obvious at a glance that $f(t, u)$ splits into a product $g(t)h(u)$, as in (a) and (b) of Example 2.6. If $f(t, u)$ doesn't obviously split, as in case (c) above, the ODE probably isn't separable.

2.2.1 Application: Falling Objects

Quadratic Air Resistance

Consider an object with mass m falling straight down under the influence of gravitational force near the earth's surface. Let us choose a coordinate system in which downward is positive and use v to denote the object's velocity (which will be a function of time t). Technically, v is a vector, but since we are interested only in vertical motion we treat v as a scalar. The gravitational force on the object is $F_g = mg$ where g denotes gravitational acceleration and is positive. However, the falling object will also experience a force due to air resistance. It is quite common to model the force of air resistance as proportional to the square of the object's speed, though this is by no means a fundamental law of nature. Reading Exercise 2.2.1 gives some justification for this choice, however.

Reading Exercise 2.2.1 A spherical object with cross sectional area A falls at speed $v > 0$ through a fluid (e.g., air or water) with density ρ , and so experiences a resistive force of magnitude F_r . Suppose F_r is of the form $F_r = KA^a\rho^b v^c$ for some constants a, b , and c , and dimensionless constant K . Find choices for a, b , and c that yield a dimensionally consistent formula. What might K depend upon?

We can use the result of Reading Exercise 2.2.1 to write $F_r = kv^2$ where $k = KA\rho$ is a constant with dimension $[k] = ML^{-1}$. But because F_r is a resistive force, F_r should point upward (the negative direction), that is, $F_r < 0$. We will thus write

$$F_r = -kv^2, \quad (2.23)$$

with k positive (recall Modeling Tip 1.1.1). Keep in mind that (2.23) only works when the object is moving downward ($v > 0$); if $v < 0$ (upward motion) the quantity v^2 remains positive, and F_r given by (2.23) would remain upward, now assisting the object's motion.

More generally, we can posit that (2.23) holds whether the object is spherical or not; the constant k will certainly depend on the object's properties, e.g., its size, shape, or cross-sectional area, and other factors.

An ODE for the Object's Motion

In view of the discussion above, the net force F on the object is $F = F_g + F_r$ or

$$F = mg - kv^2 \quad (2.24)$$

where g and k are both positive. We can combine Newton's second law $F = ma$ with $a = v'(t)$ and (2.24) to obtain $mv' = mg - kv^2$ or

$$v'(t) = g - \frac{k}{m}v^2(t). \quad (2.25)$$

If the object is dropped at time $t = 0$ with zero initial velocity then $v(0) = 0$ and then (2.25) is an initial value problem for $v(t)$. The ODE (2.25) is nonlinear, so a solution cannot be obtained using the integrating factor approach. We'll instead use a technique known as **separation of variables**, but rather than start with (2.25) (which has some unpleasant algebra that clouds the central issues) let's begin with a more straightforward example and return to (2.25) shortly. See also the project "A Shot in the Water" in Section 2.5.

Reading Exercise 2.2.2 Verify that (2.25) is separable.

2.2.2 Separation of Variables: A First Example

Consider the problem of finding a general solution to the ODE $u'(t) = tu^2(t)$ and then finding the solution with initial condition $u(1) = 4$. This ODE is separable and has the form $u' = g(t)h(u)$ with $g(t) = t$ and $h(u) = u^2$. As with the integrating factor technique for linear equations, the initial goal in separation of variables is to manipulate the ODE so that integration actually makes progress toward a solution.

Step 1 is to separate $u'(t) = tu^2(t)$ as

$$\frac{u'(t)}{u^2(t)} = t. \quad (2.26)$$

The point of Step 1 is to get $u'(t)$ and $u(t)$ on one side of the equation and all instances of the independent variable t on the other, to prepare both sides of the equation for an integration.

Step 2 is to integrate both sides of (2.26) with respect to t to obtain

$$\int \frac{u'(t)}{u^2(t)} dt + C_1 = \int t dt + C_2. \quad (2.27)$$

The right side of (2.27) is easy to integrate and is $t^2/2 + C_2$. The left side of (2.27) looks problematic—how can an antiderivative be found when we don't know what $u(t)$ is? But in fact the integral can be done with the substitution $w = u(t)$. Then $dw = u'(t) dt$ and the integral distills down to evaluating

$$\int \frac{dw}{w^2} + C_1 = -\frac{1}{w} + C_1,$$

since an antiderivative for $1/w^2$ is $-1/w$. Now (2.27) can be written as

$$-\frac{1}{u(t)} = t^2/2 + C \quad (2.28)$$

with $C = C_2 - C_1$. All derivatives are now gone and solving for $u(t)$ is an algebra problem.

Step 3 is to solve (2.28) for $u(t)$. Multiply (2.28) through by -1 and reciprocate to find

$$u(t) = -\frac{1}{t^2/2 + C}. \quad (2.29)$$

This is a general solution to the ODE $u'(t) = tu^2(t)$.

Step 4 is to solve for the value of C necessary to obtain the solution with $u(1) = 4$. Substituting $t = 1$ and $u(1) = 4$ into (2.29) yields $4 = -\frac{1}{1/2+C}$, which can be solved to find $C = -3/4$. The solution to $u'(t) = tu^2(t)$ with the required initial condition is therefore

$$u(t) = -\frac{1}{t^2/2 - 3/4}.$$

An Alternate Notation for Separation of Variables

There is a common approach to separation of variables that many people favor. It revolves around using the Leibniz notation du/dt instead of $u'(t)$. Let's again find a general solution to the ODE $u'(t) = tu^2(t)$ with this notational approach, and find the specific solution with the initial condition $u(1) = 4$.

Begin by switching to Leibniz notation for the derivative of u and write the ODE as

$$\frac{du}{dt} = tu^2,$$

where the argument t of the function u has been suppressed. This ODE is separable, of the form $du/dt = g(t)h(u)$ with $g(t) = t$ and $h(u) = u^2$.

Step 1 is to separate variables by treating du/dt as a fraction and writing the ODE as

$$\frac{du}{u^2} = t dt. \quad (2.30)$$

Compare (2.30) to (2.26). This manipulation is an abuse of the Leibniz notation, since du/dt is not really a fraction. You may have seen this before in a calculus course. The notation is designed so that this kind of abuse usually works.

Step 2 is to integrate both sides of (2.30), treating the left side as an integral with respect to u and the right side as an integral with respect to t . This yields

$$\int \frac{du}{u^2} + C_1 = \int t dt + C_2, \quad (2.31)$$

where constants of integration C_1 and C_2 have been explicitly added to both sides of (2.31). Compare (2.31) to (2.27). Working both integrals in (2.31) yields

$$-\frac{1}{u} = t^2/2 + C \quad (2.32)$$

with $C = C_2 - C_1$. Compare (2.32) to (2.28); the integral on the left in (2.31) has been evaluated using a substitution but without actually introducing a new variable w . Instead we substitute $u = u(t)$ and $du = u'(t) dt$.

Steps 3 and 4 are the same as before. Solve (2.32) for u to find a general solution

$$u = -\frac{1}{t^2/2 + C}.$$

Then substitute in $t = 1$, $u = 4$ into this general solution and solve for $C = -3/4$ as before.

2.2.3 The General Procedure for Separation of Variables

The steps for separation of variables on a general separable ODE $u'(t) = g(t)h(u(t))$ are similar to those for (2.26). We'll use the Leibniz notational approach.

1. **Separate:** Write the ODE (2.22) in the form

$$\frac{du}{h(u)} = g(t) dt. \quad (2.33)$$

2. **Integrate:** Integrate both sides of (2.33) with respect to t :

$$\int \frac{du}{h(u)} + C_1 = \int g(t) dt + C_2, \quad (2.34)$$

where the constants of integration have been explicitly added to both sides (and will later be lumped together). Let $G(t)$ denote an antiderivative for $g(t)$, so the right side of (2.34) will become $G(t) + C_2$. The left side of (2.34) involves finding an antiderivative for $1/h(u)$ with respect to u . Define

$$Q(u) = \int \frac{du}{h(u)}$$

to be the required antiderivative, so the left side of (2.34) is $Q(u) + C_1$. This means that (2.34) can be expressed as $Q(u) + C_1 = G(t) + C_2$, or

$$Q(u) = G(t) + C \quad (2.35)$$

with arbitrary constant $C = C_2 - C_1$. Again, notice that all derivatives have vanished: finding $u(t)$ is now an algebra problem.

3. **Solve:** The next step is to solve (2.35) for $u = u(t)$, by whatever algebraic manipulations are needed, carrying C along as an arbitrary constant. This will produce a general solution to the ODE (2.22), which can be written (conceptually, anyway) as

$$u(t) = Q^{-1}(G(t) + C). \quad (2.36)$$

4. **Obtain the initial condition:** If an initial condition is given, the last step is to determine C in the general solution (2.36) to satisfy the initial condition $u(t_0) = u_0$. This means solving $u_0 = Q^{-1}(G(t_0) + C)$ for C . If we back up to (2.35), this yields $C = Q(u_0) - G(t_0)$.

Reading Exercise 2.2.3 Solve $u'(t) = u^2(t) + 1$ with $u(0) = 0$ using separation of variables. Hint: write the right side of the ODE as $f(t, u) = g(t)h(u)$ where $g(t) = 1$ and $h(u) = u^2 + 1$.

2.2.4 Example: Solving the Falling Object ODE

Let's now find a general solution to the ODE (2.25) using separation of variables, and then find the solution with initial data $v(0) = 0$. The solution process here involves a couple of twists and turns that are typical in this process. There are many variations on this computation.

To begin, write the ODE using Leibniz notation as

$$\frac{dv}{dt} = -\frac{k}{m}v^2 + g. \quad (2.37)$$

Separate (2.37) by dividing both sides by $kv^2/m - g$ to obtain

$$\frac{dv}{kv^2/m - g} = -dt. \quad (2.38)$$

Integrate both sides of (2.38) with respect to the appropriate variable (v on the left, t on the right) to obtain

$$\int \frac{dv}{kv^2/m - g} = - \int 1 dt. \quad (2.39)$$

The integral on the right in (2.39) is $-t + C_1$ for any constant C_1 . The integral on the left isn't difficult, but we leave this to the reader in Exercise 2.2.6. An antiderivative is

$$\int \frac{dv}{kv^2/m - g} = \frac{1}{2} \sqrt{\frac{m}{gk}} \ln \left| \frac{v - \sqrt{mg/k}}{v + \sqrt{mg/k}} \right| + C_2. \quad (2.40)$$

The algebra is a bit simpler with the definition $\alpha = \sqrt{mg/k}$, in which case (2.40) becomes

$$\int \frac{dv}{kv^2/m - g} = \frac{\alpha}{2g} \ln \left| \frac{v - \alpha}{v + \alpha} \right| + C_2. \quad (2.41)$$

From (2.39) and (2.41) we obtain (setting $C = C_1 - C_2$, an arbitrary constant)

$$\frac{\alpha}{2g} \ln \left| \frac{v - \alpha}{v + \alpha} \right| = -t + C. \quad (2.42)$$

The next step is to solve for v , which is merely algebra. Multiply both sides of (2.42) by $2g/\alpha$; the right side becomes $-2gt/\alpha - 2gC/\alpha$, but since C is arbitrary so is $-2gC/\alpha$. It would make sense to call this new constant something like C' (or just leave it as $-2gC/\alpha$) but it's common practice to label it C again and press on. This process in which arbitrary constants are redefined on the fly without renaming is fairly common, but some care is needed (see below). We now have

$$\ln \left| \frac{v - \alpha}{v + \alpha} \right| = -\frac{2g}{\alpha} t + C \quad (2.43)$$

where C is an arbitrary constant. Exponentiate both sides of (2.43) to obtain

$$\left| \frac{v - \alpha}{v + \alpha} \right| = e^{C} e^{-2gt/\alpha}. \quad (2.44)$$

Since C is an arbitrary (real) constant, isn't e^C also arbitrary? No. For any real number C , e^C is a *positive* real number. If we want to rename it, it might be better to give it a name that reminds us of this fact, e.g., let $C^+ = e^C$. Then (2.44) becomes

$$\left| \frac{v - \alpha}{v + \alpha} \right| = C^+ e^{-2gt/\alpha}. \quad (2.45)$$

The absolute values in (2.45) can now be dropped if we're careful. To see this, note that an equation like $|z| = A$ means that $z = \pm A$. With this in mind, (2.45) can be written as

$$\frac{v - \alpha}{v + \alpha} = \pm C^+ e^{-2gt/\alpha}.$$

But if C^+ is an arbitrary positive constant, then $\pm C^+$ is an arbitrary nonzero constant, which we will call C . Then we have

$$\frac{v - \alpha}{v + \alpha} = C e^{-2gt/\alpha}.$$

Finally, routine algebra and the fact that $\alpha = \sqrt{mg/k}$ shows that a general solution to (2.25) is

$$v(t) = \sqrt{\frac{mg}{k}} \left(\frac{1 + C e^{-2t\sqrt{kg/m}}}{1 - C e^{-2t\sqrt{kg/m}}} \right). \quad (2.46)$$

The constant C appears in two places in (2.46), which is fine. This solution could be written with C appearing only once, but why bother? (Recall Remark 1.4.1.) Also, the restriction that $C \neq 0$ can be removed, since taking $C = 0$ in (2.46) yields $v(t) = mg/k$, which is also a solution to $v'(t) = g - kv^2(t)/m$, with initial data $v(0) = mg/k$.

To obtain the initial condition $v(0) = 0$ we need

$$\sqrt{\frac{mg}{k}} \left(\frac{1 + C}{1 - C} \right) = 0,$$

which gives $C = -1$. From (2.46) the solution with this initial data is thus

$$v(t) = \sqrt{\frac{mg}{k}} \left(\frac{1 - e^{-2t\sqrt{kg/m}}}{1 + e^{-2t\sqrt{kg/m}}} \right). \quad (2.47)$$

This solution may also be expressed as

$$v(t) = \sqrt{\frac{mg}{k}} \tanh(t\sqrt{kg/m}), \quad (2.48)$$

where $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ is the hyperbolic tangent function.

Reading Exercise 2.2.4 Use (2.47) with $m = 1$ kg, $g = 9.8$ meters per second squared, and $k = 0.004$ (units: newtons per (meter per second) squared) and plot $v(t)$ for $t = 0$ to $t = 20$ seconds. Does this seem reasonable? (Recall that down is the positive direction.)

Reading Exercise 2.2.5 Use (2.47) to show that $\lim_{t \rightarrow \infty} v(t) = \sqrt{mg/k}$. Then compute the dimension of $\sqrt{mg/k}$ and provide a physical interpretation of this quantity.

See also the project “A Shot in the Water” in Section 2.5 for some explorations of quadratic resistance to motion.

2.2.5 Example: Solving the Logistic Equation

Let us solve the logistic equation (1.10) for population growth from Section 1.3. First, write the equation as

$$\frac{du}{dt} = ru(1 - u/K).$$

Separate it as

$$\frac{du}{u(1 - u/K)} = r dt. \quad (2.49)$$

This isn’t the only way to separate—the constant r could be taken to the denominator on the left side of (2.49) instead of leaving it on the right, but it won’t matter.

The next step is to integrate. The antiderivative of the right side of (2.49) is easy: it is $rt + C_1$. Integrating the left side is more complicated and requires a partial fraction expansion in u . Write

$$\frac{1}{u(1 - u/K)} = \frac{A}{u} + \frac{B}{1 - u/K} = \frac{(B - A/K)u + A}{u(1 - u/K)}$$

for some constants A and B , to be determined. The numerator of the rightmost expression above must equal 1 for all u (the numerator of the leftmost expression), which requires $B - A/K = 0$ and $A = 1$, two equations in unknowns A and B . Solve these equations to find $A = 1$ and $B = 1/K$. Then a bit of algebra yields the partial fraction expansion

$$\frac{1}{u(1 - u/K)} = \frac{1}{u} + \frac{1/K}{1 - u/K} = \frac{1}{u} + \frac{1}{K - u}.$$

Integrating both sides above in u shows that

$$\int \frac{du}{u(1 - u/K)} = \ln|u| - \ln|K - u| + C_2.$$

Use this along with the antiderivative $rt + C$ to see that integrating both sides of (2.49) produces

$$\ln|u| - \ln|K-u| = rt + C, \quad (2.50)$$

where all constants are lumped on the right into C . All derivatives have now disappeared and what remains is an algebra problem.

The function u can be found by first exponentiating both sides of (2.50) (note $e^{\ln|u| - \ln|K-u|} = e^{\ln|u|}e^{-\ln|K-u|} = |u|/|K-u|$) to find

$$\frac{|u|}{|K-u|} = e^C e^{rt}$$

or equivalently,

$$\left| \frac{u}{K-u} \right| = Ce^{rt} \quad (2.51)$$

for some $C > 0$. Equation (2.51) is equivalent to $u/(K-u) = \pm Ce^{rt}$, or

$$\frac{u}{K-u} = Ce^{rt}$$

for a constant C that can be positive or negative. (However, taking $C = 0$ also yields valid solution $u = 0$ to (1.10).) Finally, solve this last equation for u and find

$$u(t) = \frac{K}{1 + e^{-rt}/C}.$$

Redefine C as $1/C$ and write

$$u(t) = \frac{K}{1 + Ce^{-rt}}. \quad (2.52)$$

This is a general solution to the logistic equation.

To obtain an initial condition $u(0) = u_0$ requires

$$\frac{K}{1+C} = u_0,$$

which leads to $C = K/u_0 - 1$. Then (2.52) can be written as

$$u(t) = \frac{Ku_0}{u_0 + e^{-rt}(K-u_0)}. \quad (2.53)$$

See Exercise 2.2.8 for an opportunity to compare (2.53) to some real data.

2.2.6 Exercises

Exercise 2.2.1 Find a general solution for each separable ODE, and then find the specific solution with the given initial condition.

- (a) $u'(t) = u(t) + 3$, $u(0) = 3$
- (b) $u'(t) = 2u(t) + 4$, $u(0) = 0$
- (c) $u'(t) = -3u(t) + 3$, $u(0) = 5$
- (d) $u'(t) = tu(t) + t$, $u(0) = 2$
- (e) $u'(t) = \sin(t)u(t) + \sin(t)$, $u(0) = 4$
- (f) $u'(t) = au(t) + b$, $u(0) = u_0$, where a, b , and u_0 are constants
- (g) $u'(t) = \sin(t)u(t)$, $u(0) = 1$
- (h) $u'(t) = t^2u(t)$, $u(1) = 2$
- (i) $u'(t) = e^t u(t)$, $u(0) = 3$

Exercise 2.2.2 Solve the Newton cooling ODE (2.14) with initial condition $u(0) = u_0$, by using separation of variables.

Exercise 2.2.3 Solve the Hill-Keller ODE (1.3) with $v(0) = 0$ using separation of variables.

Exercise 2.2.4 The *viscosity* μ of a fluid is a measure of the fluid's resistance to deformation or flow and has dimension $[\mu] = ML^{-1}T^{-1}$. (As an example of a highly viscous fluid, think of motor oil or honey.) Suppose a sphere of radius r falls moves through a fluid at speed v and so experiences a drag force F .

- Suppose the drag force F depends only on v , μ , and r . Show that the only dimensionally consistent formula for F in terms of these variables is of the form $F = kr\mu v$, where k is a dimensionless constant.
- Suppose the object has mass m and is falling straight down in a container filled with a fluid of viscosity μ , under the influence of gravity. If the object has velocity $v(t)$ (take $v > 0$ as the downward direction, and $g > 0$ as gravitational acceleration) use Newton's second law to show that $v(t)$ satisfies

$$v'(t) = g - \frac{kr\mu}{m} v(t). \quad (2.54)$$

- Find a general solution to (2.54). What is the terminal velocity of the object, in terms of k, r, μ , and m ?

Exercise 2.2.5 Solve the logistic equation with harvesting (1.12) with initial data $u(0) = u_0$ and so demonstrate that (1.13) is correct. Hint: you can either solve it directly with separation of variables, or you can note that (1.12) can be written as a standard logistic equation $u' = \tilde{r}u(1 - u/\tilde{K})$ with $\tilde{r} = r - h$ and $\tilde{K} = ((1 - h/r)K)$. Then use the logistic equation solution (1.11).

Exercise 2.2.6 Evaluate the integral on the left in (2.39) by writing it as

$$\int \frac{dv}{kv^2/m - g} = \frac{m}{k} \int \frac{dv}{v^2 - a^2},$$

with $a = \sqrt{mg/k}$ (note $mg/k > 0$ here). Evaluate the integral on the right above by using

$$\frac{1}{v^2 - a^2} = \frac{1}{2a} \frac{1}{v-a} - \frac{1}{2a} \frac{1}{v+a}$$

and use this to show that

$$\int \frac{dv}{kv^2/m - g} = \frac{1}{2} \sqrt{\frac{m}{gk}} \ln \left| \frac{v - \sqrt{mg/k}}{v + \sqrt{mg/k}} \right|.$$

Exercise 2.2.7 Use separation of variables to solve the compartmental salt tank DE (2.16) with initial condition $x(0) = 3$, and so demonstrate that (2.17) is correct.

Exercise 2.2.8 Table 2.1 contains population data concerning the growth of a species of yeast in a closed vessel (from a classic study [35]). This data may also be found at the book website [8].

Use this data and the solution (2.53) to the logistic DE to find good estimates of the constants r and K ; a graphical approach would be fine for now. Hint: start by plotting the data. You know u_0 . The solution to the logistic equation levels out at $p(t) = K$, so you can find a good guess at the value of K . How well does the solution fit the data? What do you predict as the maximum sustainable population based on this model?

Time (hours)	0	1	2	3	4	5	6	7	8
Population (millions)	9.6	18.3	29.0	47.2	71.1	119.1	174.6	257.3	350.7

Time (hours)	9	10	11	12	13	14	15	16	17
Population (millions)	441.0	513.3	559.7	594.8	629.4	640.8	651.1	655.9	659.6

Table 2.1: Yeast population (millions) as a function of time (hours).

Exercise 2.2.9 This problem is based on the SIMIODE modeling project [123]. Table 2.2 contains data for the distance that a shuttlecock (the projectile used in badminton) falls in a given time; the data is from [97]. The goal is to determine whether the ODE (2.25), namely $v'(t) = g - \frac{k}{m}v^2(t)$, provides a good model for an object falling in the presence of quadratic air resistance, if m , g , and k are suitably chosen. Or perhaps another model is better, something with linear resistance, like the Hill-Keller ODE.

- (a) Although we may reasonably take $g \approx 9.8$ meters per second squared in (2.25), we do not know m or k ; moreover, these variables appear only as a ratio k/m , so we cannot estimate them individually from the data. However, let $\tilde{k} = k/m$, so the ODE (2.25) becomes

$$v'(t) = g - \tilde{k}v^2(t). \quad (2.55)$$

It is the variable \tilde{k} that we will estimate. Verify that the solution to (2.55) with $v(0) = 0$ is

$$v(t) = \sqrt{\frac{g}{\tilde{k}}} \left(\frac{1 - e^{-2t\sqrt{g/\tilde{k}}}}{1 + e^{-2t\sqrt{g/\tilde{k}}}} \right) \quad (2.56)$$

or equivalently, that

$$v(t) = \sqrt{\frac{g}{\tilde{k}}} \tanh(t\sqrt{g/\tilde{k}}). \quad (2.57)$$

- (b) Compute the distance $d(t)$ fallen by the shuttlecock with $d(0) = 0$ as

$$d(t) = \int_0^t v(\tau) d\tau,$$

with v given by (2.56) or (2.57). Explain why this is the correct formula for $d(t)$. You may find it helpful to recall that $\int \tanh(x) dx = \ln(\cosh(x))$, where $\cosh(x) = (e^x + e^{-x})/2$.

- (c) Plot the data in Table 2.2, then plot $d(t)$ from part (b) with $g = 9.8$ and a guess at \tilde{k} ($\tilde{k} = 1$ is a good start). Adjust \tilde{k} until you obtain the best visual fit possible.
- (d) Does this model with quadratic air resistance seem reasonable?
- (e) A linear model for air resistance is easily obtained in the same manner as (2.25) and leads to $v'(t) = g - \frac{k}{m}v(t)$ or

$$v'(t) = g - \tilde{k}v(t) \quad (2.58)$$

where $\tilde{k} = k/m$ again. Justify this model with reasoning similar to that which led to (2.25).

- (f) Solve (2.58) with initial condition $v(0) = 0$. Then repeat parts (b)-(d) with this model, and find the optimal value for \tilde{k} . Compare the fit of this model to the fit obtained with the quadratic model. Is one convincingly superior?

See the project “Shuttlecocks and Model Selection” in Section 3.5.4 for more on this problem and the issue of which model is best.

Time (s)	0	0.347	0.47	0.519	0.582	0.65	0.674	0.717	0.766
Distance (m)	0	0.61	1.00	1.22	1.52	1.83	2.00	2.13	2.44

Time (s)	0.823	0.87	1.031	1.193	1.354	1.501	1.726	1.873
Distance (m)	2.74	3.00	4.00	5.00	6.00	7.00	8.50	9.50

Table 2.2: Distance (meters) fallen by a shuttlecock in time (seconds).

2.3 Qualitative and Graphical Insights

Based on the material in the last two sections, you might think that the subject of differential equations is purely computational, devoted to finding analytical solutions to ODEs. That is a part of the subject, but there's much more to it. Differential equations are highly geometric in nature, and even in cases where one can write down an analytical solution, the geometric analysis described in this section is of great value. It's always a good start, and sometimes it will be all you have.

2.3.1 Direction Fields

Let's start with a concrete example. We analyzed Newton's law of cooling as embodied by (2.14) with a constant ambient temperature A . But if you reexamine the derivation, the same reasoning allows for A to be time-dependent, in which case the temperature $u(t)$ of an object in an environment with time-varying ambient temperature $A(t)$ obeys

$$u'(t) = -k(u(t) - A(t)) \quad (2.59)$$

for some positive constant k . The ODE (2.59) is a linear constant-coefficient nonhomogeneous equation, that can be solved using the integrating factor approach. Let's consider (2.59) in the case that $k = 0.2$ and $A(t) = 10 + 5\sin(t/2)$; the precise units on the temperature scale don't matter. That is, the object sits in an environment with sinusoidally varying temperature, average value 10 degrees, but with excursions between 5 and 15 degrees. How will $u(t)$ behave?

Information about the solution can be gleaned from the ODE without actually solving. The approach is primarily graphical. Visualize a pair of tu axes (see the left panel of Figure 2.4) on which one might graph a solution $u(t)$. Suppose such a solution passes through the point $t = 5$, $u = 8$ (to make a random choice), that is, $u(5) = 8$. What can be said about the solution at this point, besides the fact that it passes through $t = 5, u = 8$? Substituting $t = 5$ and $u(5) = 8$ into the right side of the ODE (2.59) shows that

$$u'(5) = -(0.2)(u(5) - A(5)) = -(0.2)(8 - (10 + 5 \sin(5/2))) \approx 0.998. \quad (2.60)$$

The ODE itself tells us directly that as $u(t)$ passes through $t = 5, u = 8$ this solution has a slope of about 0.998. This can be indicated graphically, as in the left panel of Figure 2.4, by drawing a vector with its tail at the point $(t, u) = (5, 8)$ with a slope of 0.998; the length of the vector is not important, only the slope, so for visual appeal we use vector $\langle 2, 1.996 \rangle$, which has slope 0.998. This conclusion concerning the slope of the solution $u(t)$ that passes through the point $t = 5, u = 8$ requires only elementary arithmetic.

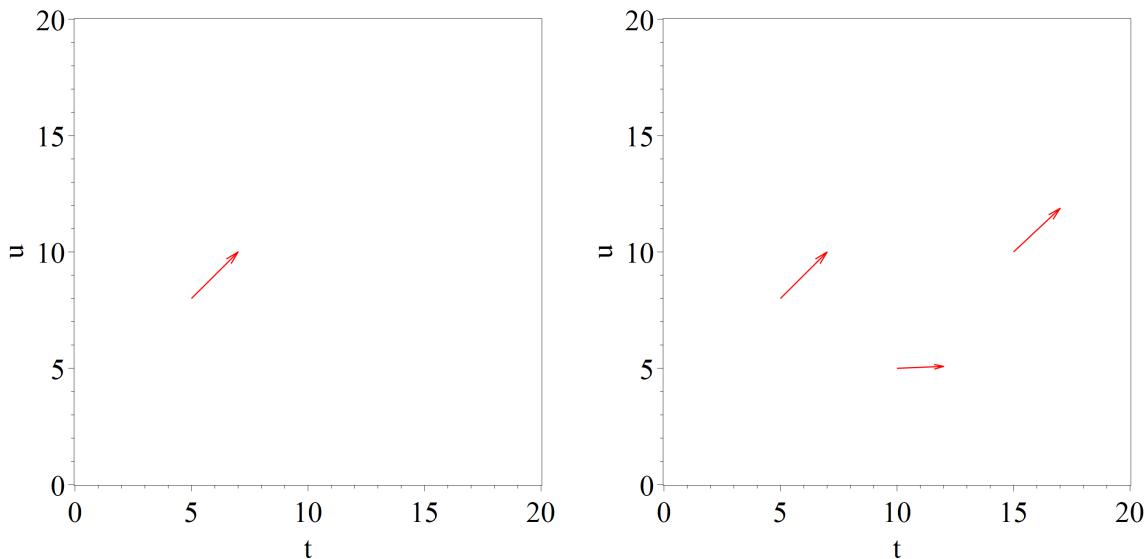


Figure 2.4: Left panel: vector to indicate the slope of the solution to the Newton-cooling ODE $u'(t) = -k(u(t) - A(t))$ with $k = 0.2$ and $A(t) = 10 + 5 \sin(t/2)$ at $t = 5, u = 8$. Right panel: vectors to indicate solution slope to this ODE at several points.

The essential takeaway here is that the solution to (2.59) that passes through $t = 5, u = 8$ must be tangent to this vector, whose slope we can determine from the right side of ODE (2.59).

Reading Exercise 2.3.1 Emulate the computation in (2.60) at $t = 10, u = 5$ to compute $u'(10)$ for the solution that satisfies $u(10) = 5$. Repeat for the point $t = 15, u = 10$ to compute $u'(15)$ for the solution that satisfies $u(15) = 10$.

In Reading Exercise 2.3.1 $u'(10) \approx 0.041$ for the solution passing through $t = 10, u = 5$, and $u'(15) \approx 0.938$ for the solution passing through $t = 15, u = 10$. We can plot a vector with tail at each of these points and corresponding slope, in the same fashion as the left panel of Figure 2.4. These two vectors are shown in Figure 2.4 in the right panel. If we pick enough points in the tu plane, perform this computation at each point, and draw the corresponding vectors, then the result is a picture that shows the slope of the solution passing through each point.

This process is highly repetitive and must be done quickly and accurately if it is to be of any value—a perfect job for a computer. The left panel of Figure 2.5 shows what is obtained if this is

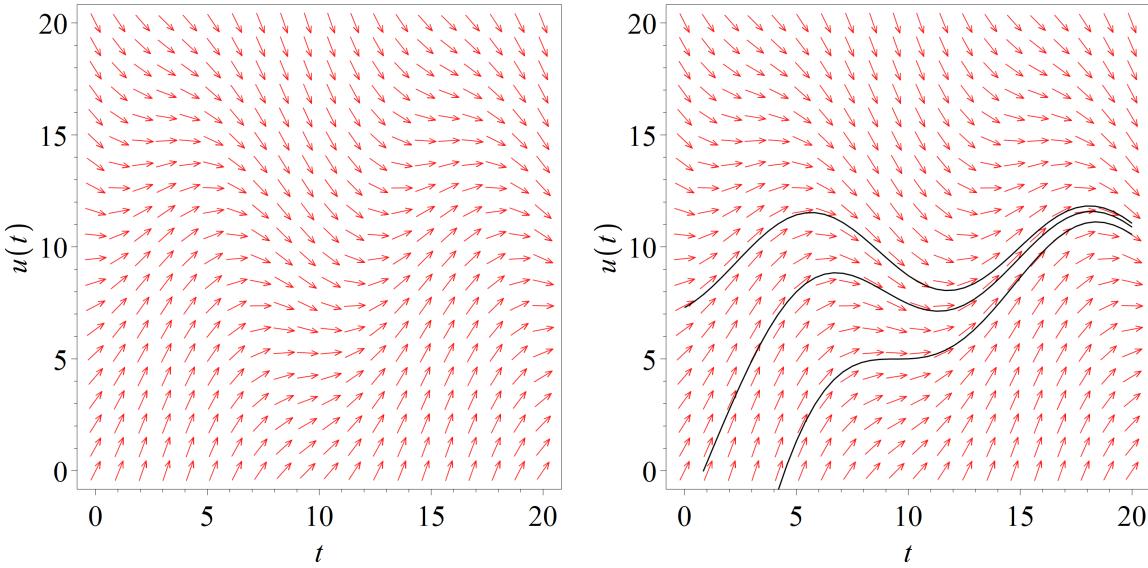


Figure 2.5: Left panel: vector field for the Newton-cooling ODE $u'(t) = -k(u(t) - A(t))$ with $k = 0.2$ and $A(t) = 10 + 5 \sin(t/2)$. Right panel: vector field for this ODE, with superimposed solution curves.

done at many more points. The resulting figure is called the **vector field** or the **slope field** or the **direction field** for the ODE. The vectors in the left panel in Figure 2.5 illustrate the slope that a solution should have at each point. The length of the vectors doesn't matter, only the slope; the length can be chosen from aesthetic considerations. Solution curves can be drawn by following the direction field arrows from left to right, in the direction of increasing t , as illustrated in the right panel of Figure 2.5.

Although this process doesn't provide a formula for the solution or prove anything quantitative in nature, it can be highly illuminating and help build intuition. For example, Figure 2.5 makes a compelling case that all solutions to (2.59) asymptotically approach the same solution curve, itself some kind of periodic function. This can be proved by noting that a general solution to (2.59) with $k = 0.2$ and $A(t) = 10 + 5 \sin(t/2)$ can be found using the integrating factor approach and is

$$u(t) = 10 + \frac{20}{29} \sin(t/2) - \frac{50}{29} \cos(t/2) + Ce^{-t/5}.$$

All solutions therefore decay to the curve $u = 10 + \frac{20}{29} \sin(t/2) - \frac{50}{29} \cos(t/2)$ as t increases.

Reading Exercise 2.3.2 Verify that $u(t) = 10 + \frac{20}{29} \sin(t/2) - \frac{50}{29} \cos(t/2)$ is itself a solution to $u'(t) = -(0.2)(u(t) - A(t))$ with $A(t) = 10 + 5 \sin(t/2)$.

For any ODE $u' = f(t, u)$, sketching the direction field comes down to choosing many points $t = t_0$, $u = u_0$ in the tu plane, computing the slope $u'(t_0) = f(t_0, u_0)$ of the solution that passes through each point, and then plotting a vector with tail at $t = t_0$, $u = u_0$ with appropriate slope. This can be done for any ODE of the form $u' = f(t, u)$, and the entire computation involves only arithmetic. We are thus empowered to graphically analyze any first-order scalar ODE. Of course, it's a lot of arithmetic, so software packages like Maple, Mathematica, Matlab, etc., are helpful.

2.3.2 Autonomous Equations

There is a special type of first-order equation that's even easier to analyze graphically, and we've already encountered a number of examples. General first-order ODEs are of the form $u' = f(t, u)$,

but in many cases the right side does not depend explicitly on t .

Definition 2.3.1 — Basic ODE Terminology. A first-order scalar ordinary differential equation is **autonomous** if it is of the form $u'(t) = f(u(t))$ (or $u' = f(u)$).

Physical systems modeled by autonomous ODEs are often referred to as **time-invariant**, especially in engineering.

■ **Example 2.7** Consider the following ODEs:

1. $v' = P - kv$, where k is a constant (the Hill-Keller ODE, (1.3)).
2. $u' = rc_1 - \frac{r}{V}u$, where r, c_1 , and V are constants (this is (1.5)).
3. $\frac{du}{dt} = ru(1 - u/K)$, where r and K are constants (the logistic equation (1.10)).
4. $u' = -k(u - \cos(t))$, k a constant.

The ODEs (1)-(3) in the list above are autonomous, and have been considered in earlier sections. The ODE in (4) is not autonomous, due to the explicit dependence of the right hand side on t . However, most of the differential equations arising from the applications we've considered so far (and those to come) are autonomous. ■

Reading Exercise 2.3.3 Is the falling object ODE $v'(t) = g - kv^2(t)/m$ autonomous? (This is (2.25).)

Autonomous ODEs have exceptionally simple direction fields, since the right side of the ODE $u' = f(u)$ does not depend on t . The slope of a solution curve that passes through $t = t_0$, $u = u_0$ is $f(u_0)$, and this greatly reduces the work necessary to compute the direction field, since we only have to compute $f(u)$ for some range of u ; t does not factor into the computation. Direction fields for autonomous ODEs also have a characteristic appearance, as you will see below.

■ **Example 2.8** Consider the ODE $v'(t) = 11 - v(t)$, which is the Hill-Keller ODE (1.3) with $k = 1$ and $P = 11$. The direction field is shown in the left panel of Figure 2.6, with $0 \leq t \leq 10$ and $0 \leq v \leq 20$. The slope of the vector at any point (t, v) does not depend on t , only v . As a consequence, the vectors on any horizontal line (lines of constant v coordinate) all have the same slope. Note that there is a constant solution to $v'(t) = 11 - v(t)$ at $v = 11$, where all vectors have slope $v' = 11 - 11 = 0$. This solution at $v = 11$ is indicated by the horizontal line.

We can economize our efforts in sketching a direction field for an autonomous ODE by eliminating the t axis, as shown in the right panel in Figure 2.6. Think of the figure on the right as a compressed version of the direction field, flattened to just the one-dimensional v axis. The constant solution at $v = 11$ becomes the black dot in the panel on the right. The solutions $v(t)$ to $v'(t) = 11 - v(t)$ that pass through points with $0 < v < 11$ are all increasing, since there $11 - v(t) > 0$, and so $v'(t) > 0$. Solutions passing through points with $v > 11$ are decreasing, since there $11 - v(t) < 0$, and so $v'(t) < 0$. The arrows above and below $v = 11$ in the right panel in Figure 2.6 indicate how solutions behave, either increase or decrease, in each region.

The figure in the right panel in Figure 2.6 is called a **phase portrait** or a **phase line portrait** for this ODE. From this graphical analysis, it appears that all solutions asymptotically approach the constant solution $v(t) = 11$. ■

2.3.3 Phase Portraits

There is no need to draw the direction field before sketching a phase portrait for an autonomous ODE. It can be done directly, and usually with very little effort. First, let's make a definition.

Definition 2.3.2 — Basic ODE Terminology. If $u(t)$ is a solution to an autonomous ODE $u' = f(u)$ and $u(t)$ is constant, then the function $u(t)$ is called an **equilibrium solution** for the ODE.

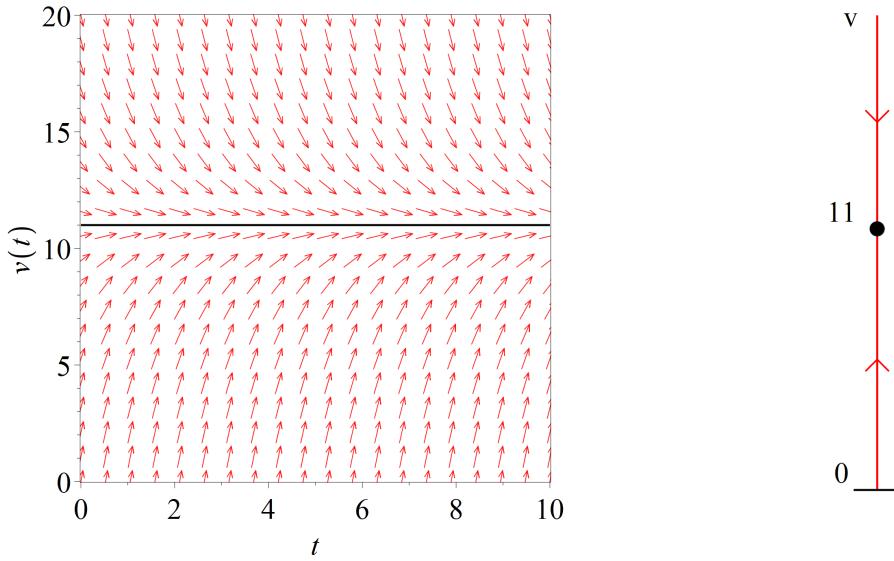


Figure 2.6: Left panel: vector field for the Hill-Keller ODE $v'(t) = 11 - v(t)$. Right panel: phase portrait for this ODE.

■ **Example 2.9** Consider the autonomous ODE $u' = -2u + 4$. The equilibrium solutions for this ODE are functions $u(t)$ that are constant, so $u(t) = u^*$ for some constant u^* . This means that $u'(t) = 0$. Substituting $u'(t) = 0$ and $u(t) = u^*$ into the ODE $u' = -2u + 4$ yields $0 = -2u^* + 4$. We can solve to find that $u^* = 2$, so $u(t) = 2$ is the only equilibrium solution for this ODE. ■

More generally, if $u(t) = u^*$ is an equilibrium solution for an autonomous ODE $u' = f(u)$, then $u'(t) = 0$. Substituting $u'(t) = 0$ and $u(t) = u^*$ into the ODE $u' = f(u)$ shows that $f(u^*) = 0$. That is, the equilibrium solutions $u(t) = u^*$ for $u' = f(u)$ can be found by determining all solutions to $f(u^*) = 0$. Any such number u^* is called a **fixed point** or **critical point** for the ODE. If $u(t)$ is not an equilibrium solution for the ODE then $f(u(t)) > 0$ or $f(u(t)) < 0$ at any time t , and so $u(t)$ satisfies $u'(t) > 0$ (the solution is increasing) or $u'(t) < 0$ (the solution is decreasing), respectively.

A Recipe for Drawing Phase Portraits

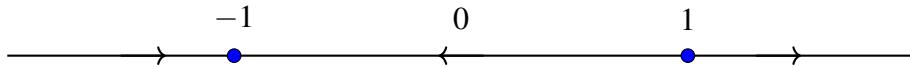
Here's how to draw a phase portrait for an autonomous ODE $u' = f(u)$. The phase portrait in the right panel of Figure 2.6 was vertically oriented, but most people draw them horizontally; it doesn't matter. It may help to refer to the right panel of Figure 2.6.

1. Find the fixed points for the ODE by finding all solutions to the equation $f(u) = 0$. Plot each fixed point as a dot on a line, the u axis.
2. The fixed points will divide the u axis up into one or more intervals, some of which may stretch to ∞ or $-\infty$. In each interval either $f(u) > 0$ or $f(u) < 0$. If $f(u) > 0$ in an interval, draw an arrow in the direction of increasing u . If $f(u) < 0$ in an interval, draw an arrow in the direction of decreasing u .

Let's look at an example.

■ **Example 2.10** Let us sketch a phase portrait for the autonomous ODE $u'(t) = u^2(t) - 1$. In this case the ODE is $u' = f(u)$ with $f(u) = u^2 - 1$ and the fixed points are the solutions to $f(u) = 0$, that is, $u^2 - 1 = 0$. These fixed points are $u = -1$ and $u = 1$, shown as the blue dots in the u axis in Figure 2.7.

The fixed points divide the u axis into intervals $(-\infty, -1)$, $(-1, 1)$, and $(1, \infty)$. If $u < -1$ then $f(u) > 0$, so solutions in this interval increase. If $-1 < u < 1$ then $f(u) < 0$, so solutions in this interval decrease. If $u > 1$ then $f(u) > 1$ and solutions in this interval increase. This is summarized

Figure 2.7: Phase portrait for the ODE $u' = u^2 - 1$.

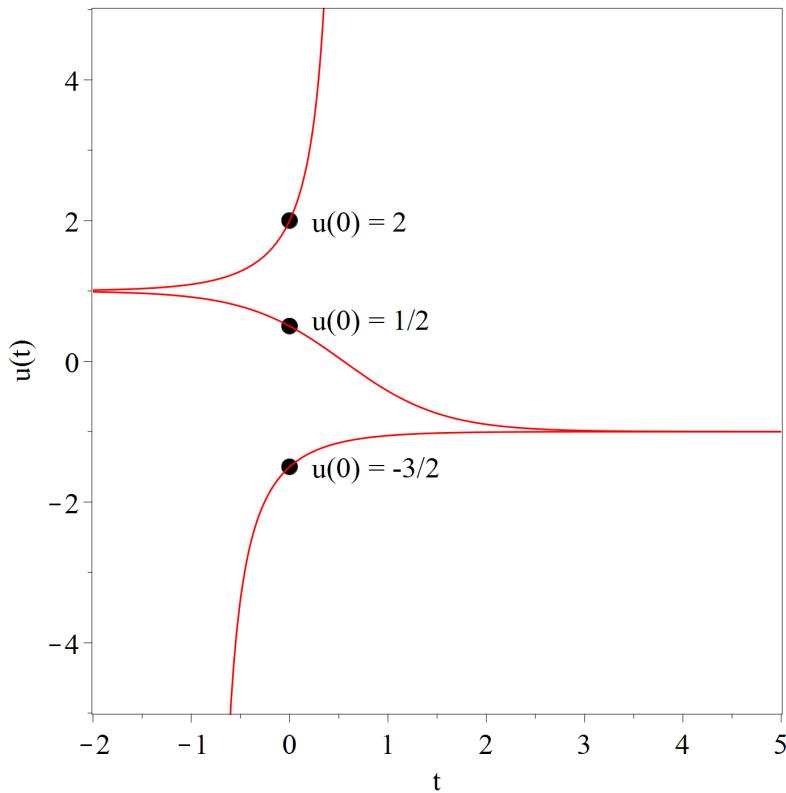
in the phase portrait of Figure 2.7; the arrows between the fixed points indicate whether solutions increase or decrease in that interval. ■

Reading Exercise 2.3.4 Sketch a phase portrait for the autonomous ODE $u'(t) = u^2(t) - 2u(t) - 3$ on the range $-5 \leq u \leq 5$.

The Purpose of the Phase Portrait

The phase portrait in Figure 2.7 is not an end unto itself, but rather a tool for understanding how solutions to the autonomous ODE $u'(t) = u^2(t) - 1$ behave. To illustrate, consider three solutions to $u'(t) = u^2(t) - 1$, one that satisfies $u(0) = -3/2$, another that satisfies $u(0) = 1/2$, and another that satisfies $u(0) = 2$. We can use the phase portrait to sketch the graph of each solution.

Let's first consider the solution with $u(0) = -3/2$. This solution passes through the point $t = 0$, $u = -3/2$, and this point is shown as a black dot in Figure 2.8. The phase portrait of Figure 2.7 indicates that for $u < -1$, which is the region in which this point lies, solutions are increasing. A solution curve through $t = 0, u = -3/2$ is shown that is consistent with this requirement. It appears that the solution asymptotically approaches -1 as t increases. The solution is also sketched for $t < 0$ and, in accordance with the phase portrait, is increasing with respect to t .

Figure 2.8: Three solutions to $u' = u^2 - 1$ plotted on $-2 \leq t \leq 5$, with initial conditions $u(0) = -3/2$, $u(0) = 1/2$, and $u(0) = 2$.

A solution curve that satisfies $u(0) = 1/2$ is also shown in Figure 2.8. In the region $-1 < u < 1$, the phase portrait of Figure 2.7 indicates that solutions are decreasing, and the solution curve

through $t = 0, u = 1/2$ is consistent with this observation. It appears that the solution asymptotically approaches -1 as t increases. The solution can also be sketched for $t < 0$ and appears to approach $u = 1$ as $t \rightarrow -\infty$.

Finally, the a solution curve with $u(0) = 2$ is shown in Figure 2.8 and is consistent with the phase portrait, which indicates the solution should be increasing, since $u > 1$.

The phase portrait allows us to make qualitative conclusions about the behavior of solutions, without actually solving the autonomous ODE. In this case the solutions that satisfy $u(0) = -3/2$ and $u(0) = 1/2$ asymptotically approach the equilibrium solution $u = -1$ as t increases. The solution with $u(0) = 2$ grows without limit, and as it turns out, has a vertical asymptote around $t \approx 0.55$. However, it should be noted that the phase portrait gives us qualitative information about the long-term behavior of solutions, not precise values. As such, if we had drawn Figure 2.8 by hand we would not have quantitative information about the scaling on the t axis.

Reading Exercise 2.3.5 Use the phase portrait you drew in Reading Exercise 2.3.4 to sketch what solutions to the ODE $u'(t) = u^2(t) - 2u(t) - 3$ with initial data $u(0) = -4$, $u(0) = 0$, and $u(0) = 4$ would look like, on a pair of tu axes. Make the u axis range at least $-5 \leq u \leq 5$, and consider $t < 0$ as well as $t \geq 0$.

Remark 2.3.1 Example 2.10 and the associated Figure 2.8 illustrate that a solution $u(t)$ to an ODE $u' = f(t, u)$ with $u(t_0) = u_0$ may blow up to ∞ (or $-\infty$) in finite time. In Figure 2.8 the solution to $u' = u^2 - 1$ with initial condition $u(0) = 2$ has a vertical asymptote at a time $t = t_f$ with $t_f \approx 0.55$; for $t \geq t_f$ the solution $u(t)$ is not defined. The same thing can happen when moving backward in time: a solution $u(t)$ to an ODE may have an asymptote at $t = t_i$ with $t_i < t_0$, and so $u(t)$ is not defined for $t \leq t_i$. This is also illustrated in Figure 2.8, in which the solution with $u(0) = -3/2$ has a vertical asymptote somewhere around $t \approx -0.6$. Thus a solution $u(t)$ to an ODE $u' = f(t, u)$ with $u(t_0) = u_0$ is only defined on some interval (t_i, t_f) , though this interval may be of the form $(-\infty, t_f)$, (t_i, ∞) , or $(-\infty, \infty)$.

2.3.4 Fixed Points and Stability

From the phase portrait in Figure 2.7 it is apparent that solutions that start sufficiently close to $u = -1$ approach this fixed point as t increases. This is not the case for the fixed point at $u = 1$; if a solution $u(t)$ starts with $u(t_0) = u_0$ it will not approach $u = 1$ (unless $u_0 = 1$ to begin with). The fixed point at $u = -1$ is asymptotically stable and the fixed point at $u = 1$ is unstable, terms we will now define more precisely.

Suppose that $u(t) = u^*$ is an equilibrium solution for an autonomous ODE. Informally,

- The fixed point $u = u^*$ is **stable** if all solutions that start sufficiently close to u^* stay close to u^* , although these solutions need not satisfy $\lim_{t \rightarrow \infty} u(t) = u^*$.
- The fixed point $u = u^*$ is **asymptotically stable** if all solutions $u(t)$ that start sufficiently close to u^* approach u^* , that is, $\lim_{t \rightarrow \infty} u(t) = u^*$.
- If the fixed point $u = u^*$ is not stable it is **unstable**.

The above is the slightly informal, intuitive definition of these terms. The precise definition is more technical.

Definition 2.3.3 — Basic ODE Terminology. Suppose $u(t) = u^*$ is an equilibrium solution to an autonomous ODE. Let $u(t)$ be a solution to the ODE with initial condition $u(t_0) = u_0$ for some u_0 . Then

- The fixed point u^* is **stable** if for each $\varepsilon > 0$ there is some real number $\delta > 0$ so that if $|u_0 - u^*| < \delta$ then $|u(t) - u^*| < \varepsilon$ for all $t > t_0$.
- The fixed point u^* is **asymptotically stable** if it is stable and for some $\delta > 0$, if $|u_0 - u^*| <$

δ then $\lim_{t \rightarrow \infty} u(t) = u^*$.

- If the fixed point is not stable it is **unstable**.

Asymptotically stable fixed points are also called **sinks**. Unstable fixed points like $u = 1$ in Example 2.10, in which the arrows point away from the fixed point on both sides, are also called **sources**. There is also another possibility.

■ **Example 2.11** Consider the ODE $u' = u^2$. The only fixed point for this ODE is the solution to $u^2 = 0$, that is, $u = 0$. If $u < 0$ or if $u > 0$ then $u'(t) > 0$ and the solution is growing. The phase portrait is shown in Figure 2.9. In this case solutions with initial condition $u(t_0) < 0$ increase toward the fixed point $u = 0$, while solutions with initial condition $u(t_0) > 0$ grow, apparently without bound. Fixed points such as this are called **semi-stable**. However, according to Definition 2.3.3 these fixed points are unstable—not all solutions that start sufficiently close to $u = 0$ stay close to $u = 0$. ■

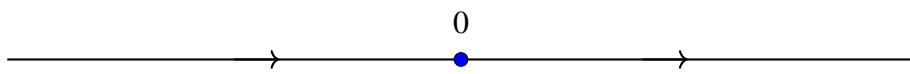


Figure 2.9: Phase portrait for the ODE $u' = u^2$.

Reading Exercise 2.3.6 Use your phase portrait from Reading Exercise 2.3.4 to characterize each fixed point as asymptotically stable or unstable.

A Small Refinement in Phase Portrait Sketching

Some authors graphically distinguish stable fixed points from unstable fixed points in a phase portrait by drawing a filled dot for a stable fixed point and a circle for an unstable fixed point. In this case Figure 2.7 for the ODE $u' = u^2 - 1$ would be drawn as shown in Figure 2.10.

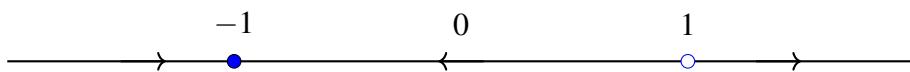


Figure 2.10: Version 2 of a phase portrait for the ODE $u' = u^2 - 1$.

A further refinement is to draw semi-stable fixed points as half-filled circles, so the phase portrait for $u' = u^2$, shown Figure 2.9, would now be drawn as shown in Figure 2.11.

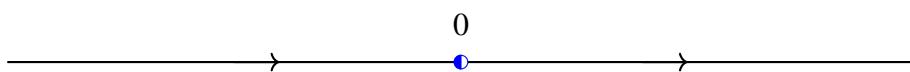


Figure 2.11: Version 2 of a phase portrait for the ODE $u' = u^2$.

Reading Exercise 2.3.7 Improve your phase portrait from Reading Exercise 2.3.4 to graphically indicate the stability of each fixed point by using a solid dot or empty circle as in Figure 2.10.

2.3.5 Determining the Stability of Fixed Points

Sign Changes in f at Fixed Points

The fixed points for autonomous ODEs $u' = f(u)$ are either stable or unstable. Suppose $u(t) = u^*$ is a fixed point for $u' = f(u)$, so $f(u^*) = 0$. Suppose also that for some $\delta > 0$ there is an open interval of the form $(u^* - \delta, u^* + \delta)$ around u^* that contains no other fixed points. Based on our work sketching phase portraits, the stability of u^* can be determined by looking at the sign of f on either side of u^* . In particular, a fixed point can be characterized as follows:

- **Stable:** If $f(u) > 0$ for $u^* - \delta < u < u^*$ (to the left of u^*) and $f(u) < 0$ for $u^* < u < u^* + \delta$ (to the right of u^*) then u^* is stable.
- **Semi-stable:** If $f(u) > 0$ for all $u \in (u^* - \delta, u^* + \delta)$ (with $u \neq u^*$) or $f(u) < 0$ for all u in the interval $(u^* - \delta, u^* + \delta)$ (with $u \neq u^*$) then u^* is semi-stable.
- **Unstable:** If $f(u) < 0$ for $u^* - \delta < u < u^*$ (to the left of u^*) and $f(u) > 0$ for $u^* < u < u^* + \delta$ (to the right of u^*) then u^* is unstable.

And as remarked above, semi-stable fixed points are a particular type of unstable fixed point. Examining the sign of f on each side of a fixed point is usually the most straightforward method for determining stability.

The Sign of f' at a Fixed Point

There is another way to determine the stability of a fixed point that can be useful. In the vast majority of cases we consider, the function f in $u' = f(u)$ will be continuously differentiable everywhere that f is defined. In such cases it even easier to test the stability of fixed points. In particular, suppose u^* is a fixed point:

- **Stable Fixed Point:** Suppose that $f'(u^*) < 0$. An argument from basic calculus then shows that $f(u)$ is strictly decreasing near $u = u^*$. Thus $f(u) > f(u^*) = 0$ on some interval $u^* - \delta < u < u^*$ and $f(u) < f(u^*) = 0$ on some interval $u^* < u < u^* + \delta$. In this case the fixed point is stable.
- **Unstable Fixed Point:** Suppose that $f'(u^*) > 0$. An argument from basic calculus then shows that $f(u)$ is strictly increasing near $u = u^*$. Thus $f(u) < f(u^*) = 0$ for $u^* - \delta < u < u^*$ and $f(u) > f(u^*) = 0$ for $u^* < u < u^* + \delta$. In this case the fixed point is unstable.

However, if $f'(u^*) = 0$ then the stability of the fixed point cannot be determined using this method; this may remind you of the second derivative test for maxima and minima in calculus. The direct approach of examining f on each side of u^* would be applicable, though.

Reading Exercise 2.3.8 Determine the stability of each fixed point for $u' = f(u)$ with $f(u) = u^2 - 2u - 3$ by examining the sign of f' at each fixed point.

■ **Example 2.12** Let us sketch a phase portrait for the logistic equation (1.10), $u' = ru(1 - u/K)$, and classify the stability of the fixed points. Doing so will illustrate the power of this phase portrait technique—it won't be necessary to specify the constants r or K , as it would be in order to have the computer draw a direction field. We merely need to know that both r and K are positive. Also, since the equation models a population, only the region $u \geq 0$ is relevant.

To begin, note that the logistic equation is autonomous, of the form $u' = f(u)$ with $f(u) = ru(1 - u/K)$. The fixed points are the solutions to $f(u) = 0$, that is, $ru(1 - u/K) = 0$, which are easily found to be $u = 0$ and $u = K$. Draw a u axis horizontally and put dots at the fixed points, $u = 0$ and $u = K$, labeled appropriately, as in Figure 2.12. As it turns out, the fixed point at $u = 0$ is unstable and the one at $u = K$ is stable, so we'll just draw appropriate dots right now.

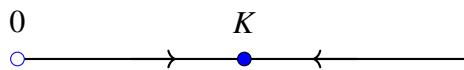


Figure 2.12: Phase portrait for the logistic equation $u' = ru(1 - u/K)$.

The next step is to determine how the solutions behave between the fixed points, by looking at whether u' (or $f(u)$) is positive or negative in each interval $0 < u < K$ and $K < u$, ignoring the physically irrelevant region $u < 0$. It's easy to see that $f(u) = ru(1 - u/K) > 0$ holds if $0 < u < K$ (substitute in $u = K/2$ and find $f(K/2) = r(K/2)(1 - (K/2)/K) = rK/4 > 0$, using $r, K > 0$). As a result, we draw an arrow pointing to the right somewhere between $u = 0$ and $u = K$, to indicate that solutions in this region increase. The same analysis for $u > K$ shows that $f(u) < 0$ here (e.g., $f(2K) = -2rK < 0$) so solutions decrease and the appropriate arrow is to the left. As remarked, we

ignore the region $u < 0$ since this is not physically relevant (although $u' < 0$ there, if $u < 0$ made sense).

The phase portrait also makes it clear that the fixed point at $u = K$ is asymptotically stable. The fixed point at $u = 0$ is unstable, since any solutions that start close to $u = 0$ in the range $0 < u < K$ move away from $u = 0$. An alternative approach to determine the stability of each fixed point is to look at the sign of f' at each fixed point. Given that $f(u) = ru(1 - u/K)$ we have $f'(u) = r - 2ru/K$. Then $f'(0) = r > 0$, so the fixed point $u = 0$ is unstable. But $f'(K) = r - 2r < 0$, so the fixed point $u = K$ is stable. ■

Take note: we just figured out how solutions to the logistic equation behave and all we did was some straightforward algebra and elementary derivatives, while completely avoiding the work that was necessary to solve the logistic ODE in Section 2.2.5.

Reading Exercise 2.3.9 Use the phase portrait in Figure 2.12 to sketch the graph of solutions $u(t)$ to the logistic equation with $u(0) = K/2$, and with $u(0) = 2K$.

Reading Exercise 2.3.10 What happens when you try to find a fixed point (constant solution) to a general first-order ODE $u' = f(t, u)$? As a specific example, consider $u'(t) = u(t) + \sin(t)$. What happens if you try to obtain $u(t) = u^*$ for some constant u^* and all t ?

Reading Exercise 2.3.11 Consider the autonomous ODE $u'(t) = 0$. Explain why $u(t) = c$ is a fixed point for any real number c . Argue that all of these fixed points are stable, but not asymptotically stable.

Reading Exercise 2.3.12 Sketch a phase portrait to show that all solutions to the salt tank ODE (2.16) approach the equilibrium solution $x(t) = x^* = 20$ kg (no need to solve the ODE.) Of course you can confine your attention to $x \geq 0$, although it won't change the conclusion.

■ **Example 2.13** Consider the problem of constructing an autonomous ODE $u' = f(u)$ (by specifying the function f) that has a stable fixed point at $u = -1$, an unstable fixed point at $u = 1$, and a stable fixed point at $u = 2$. Since f must be zero at any fixed point, the obvious choice is to take f as a polynomial with roots at $u = -1$, $u = 1$, and $u = 2$. A first try might be $f(u) = (u+1)(u-1)(u-2)$. Plotting f , or just testing select choices of u , shows that $f(u) < 0$ for $u < -1$, $f(u) > 0$ for $-1 < u < 1$, $f(u) < 0$ for $1 < u < 2$, and $f(u) > 0$ for $u > 2$. In each case the sign of f is exactly the opposite of what we want for the indicated stability. But multiplying by -1 fixes the issue, so take $f(u) = -(u+1)(u-1)(u-2)$. An autonomous ODE with the given phase portrait is

$$u' = -(u+1)(u-1)(u-2).$$

The right hand side can be multiplied by any positive constant, or indeed, by any positive function of u , and the ODE still has the desired behavior. ■

2.3.6 Bifurcations

In this section we'll take a brief look at the notion of **bifurcations**.

Bifurcations in the Harvested Logistic Equation

The differential equations in which we are interested may often contain unspecified parameters. For example, the Hill-Keller equation (1.3) contains parameters k and P , while the harvested logistic equation (1.12), reproduced here for convenience,

$$u'(t) = ru(t)(1 - u(t)/K) - hu(t), \quad (2.61)$$

depends on parameters r, K , and h . Recall that $u(t)$ is the population of a species (e.g., fish) with growth rate r in an environment with carrying capacity K , being harvested at a rate h . In many

cases the behavior of solutions depends critically on the relative values of the parameters, and this dependence may be exactly what is of interest, rather than the solution for any particular choice of parameters. For example, how will solutions to the harvested logistic equation (2.61) behave if h is very large? Could a sufficiently large value for h drive the species to extinction?

Let's explore this issue by sketching some phase portraits for (2.61) under a variety of assumptions about r and h . The fixed points are solutions to

$$ru(1 - u/K) - hu = 0,$$

and are easily found to be $u = 0$ and

$$u^* = K(1 - h/r). \quad (2.62)$$

Only the physically relevant region $u \geq 0$ is of interest. One thing is clear from (2.62). If $h < r$ then $u^* > 0$ and there is a physically meaningful solution in which the fish population has a constant positive value. That is, $u(t) = u^* > 0$ is a solution to (2.61). But if $h \geq r$ then there is no equilibrium corresponding to a positive population. When parameters like the harvesting rate h change, the phase portrait may change dramatically, and solutions to the ODE may drastically change behavior. Changes in the number or stability of fixed points when parameters in the ODE change are called **bifurcations**.

Let's consider the phase portrait for the harvested logistic equation for each case, $h < r$, $h > r$, and $h = r$. Define $f(u) = (r - h)u - ru^2/K$, which is the right side of (2.61) after simplifying.

- $h < r$: This is the situation in which the harvesting rate h does not exceed the intrinsic growth rate r . In this case there are two fixed points, $u = 0$ and from (2.62), $u = u^*$. We find that $u^* > 0$, since $1 - h/r > 0$ if $h < r$. In the interval $0 < u < u^*$, solution directions can be determined by computing $f(u^*/2) = \frac{K(h-r)^2}{4r}$ (after a bit of algebra) and since this quantity is positive, solutions in this region increase. We can also see that if $u > u^*$ then $f(u) < 0$, say, by computing $f(2u^*) = -\frac{2K(h-r)^2}{r}$, which is negative. The resulting phase portrait is shown in Figure 2.13.

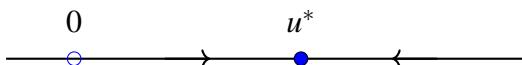


Figure 2.13: Phase portrait for $u' = ru(1 - u/K) - hu$ with $h < r$.

The arrows make it easy to see that $u = 0$ is unstable and $u = u^*$ is stable. But as an additional check we can also compute $f'(u) = r - h - 2ru/K$ so that $f'(0) = r - h > 0$ (unstable) and $f'(u^*) = -r(1 - h/r) < 0$ (stable), which is in accordance with the phase portrait.

- $r < h$: This is the situation in which the harvesting rate h exceeds the intrinsic growth rate r . In this case the fixed point given by (2.62) satisfies $u^* < 0$ and ceases to be physically relevant. Here $u = 0$ is the only nonnegative fixed point. For $u > 0$ compute $f(u) = (r - h)u - ru^2/K < 0$ (since $r - h < 0$). The corresponding phase portrait is shown in Figure 2.14.

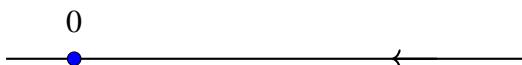


Figure 2.14: Phase portrait for $u' = ru(1 - u/K) - hu$ with $h > r$.

It's clear that if $h > r$ then 0 is now a stable fixed point, if the only concern is solutions for which $u \geq 0$. You can also check that since $r < h$, $f'(0) = r - h < 0$, confirming that

the origin is stable. The analysis of the cases $h < r$ and $h > r$ illustrates the notion of a bifurcation: as h increases from $h < r$ to $h > r$ the equilibrium population u^* is lost, and the fixed point $u = 0$ changes from unstable to stable.

- $r = h$: See Reading Exercise 2.3.13.

Reading Exercise 2.3.13 Consider the harvested logistic equation (2.61) in the razor's edge case that $h = r$. Show that $u = 0$ is the only fixed point, draw a phase portrait, and use it to show that $u = 0$ is semi-stable, at least if we are willing to consider $u < 0$ in addition to $u \geq 0$.

The analysis above lets us make a strong and physically relevant conclusion without solving the harvested logistic equation (2.61): If $h < r$ (the harvesting rate h is less than the growth rate r) then there is a stable equilibrium solution $u^* = K(1 - h/r)$ for the population which all solutions asymptotically approach. If $h > r$ then there is no positive equilibrium solution and the species is doomed to extinction.

Bifurcation Diagrams

The above observations and phase portraits can be amalgamated into a single figure called a **bifurcation diagram**, shown in Figure 2.15. Each vertical line is a phase portrait for (2.61) for a different value of the parameter h (r and K are fixed). In these phase portraits the region $u < 0$ has been included; even though it is non-physical, it is mathematically interesting. The leftmost vertical line is a typical phase portrait in the case that $h < r$, where the fixed point $u = u^*$ is stable and $u = 0$ is unstable. The rightmost line is a phase portrait in the case that $h > r$, and the middle shows the case $h = r$ examined in Reading Exercise 2.3.13. Each fixed point is filled, unfilled, or half-filled according to its stability. The solid lines between the vertical phase portraits indicate the trajectories of the relevant fixed points that are stable as h increases (and we move to the right). The dashed lines indicate the trajectories of the fixed points that are unstable. As h exceeds r the fixed points briefly coalesce and then separate again, with $u = u^*$ going from stable to unstable as u^* moves from a positive to negative value, while $u = 0$ goes from being unstable to stable. This type of bifurcation at $h = r$ is called a **transcritical bifurcation** and the fixed points $u = u^*$ and $u = 0$ undergo an **exchange of stability**. Of course from a purely physical perspective in which only the region $u \geq 0$ is of interest, the fixed point $u = u^*$ disappears when $h > r$.

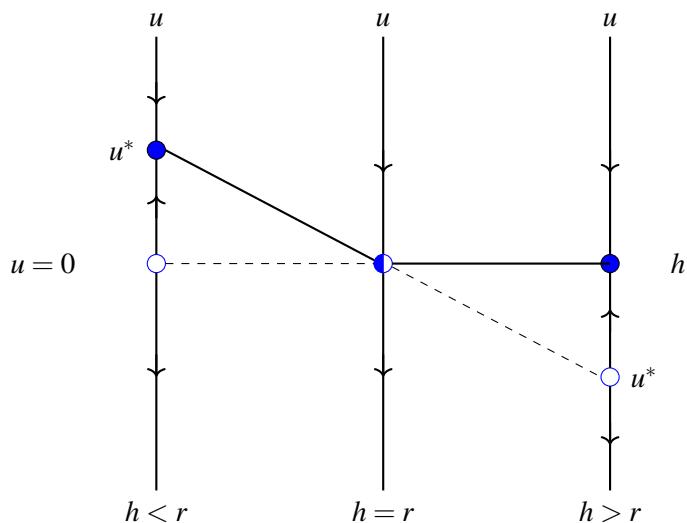


Figure 2.15: Bifurcation diagram for the harvested logistic equation (2.61).

2.3.7 Exercises

Exercise 2.3.1 For each ODE below and points $t = t_0$, $u = u_0$, compute the slope $u'(t_0)$ of the solution that passes through each point. Then plot appropriate vectors on a pair of tu axes to form a (crude) direction field; make the vectors fairly short, perhaps length $1/4$ or so.

- (a) $u'(t) = u(t) - 2t$, for (t_0, u_0) pairs $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$.
- (b) $u'(t) = u^2(t) + t + 1$, for (t_0, u_0) pairs $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$.
- (c) $u'(t) = -u(t)$, for (t_0, u_0) pairs $(0, 1)$, $(0, 2)$, $(1, 1)$, and $(1, 3)$.
- (d) $u'(t) = -1/u(t)$, for (t_0, u_0) pairs $(0, 1)$, $(0, 2)$, $(1, 1)$, and $(1, 3)$.

Exercise 2.3.2 Use whatever technology you have available to sketch a direction field for the given ODE on the specified range. If the ODE is autonomous, visually identify the equilibrium solutions, if any.

- (a) $u'(t) = u(t) - 2t$ for $0 \leq t \leq 2$ and $0 \leq u \leq 2$.
- (b) $u'(t) = u^2(t) + t + 1$ for $-2 \leq t \leq 2$ and $-2 \leq u \leq 2$.
- (c) $u'(t) = -u(t)$ for $-2 \leq t \leq 2$ and $-2 \leq u \leq 2$.
- (d) $u'(t) = -1/u(t)$ for $-2 \leq t \leq 2$ and $-2 \leq u \leq 2$.
- (e) $u'(t) = u(t)(u(t) - 3)$ for $-2 \leq t \leq 5$ and $-2 \leq u \leq 5$.
- (f) $u'(t) = (u(t) - 1)(u(t) + 1)$ for $-2 \leq t \leq 5$ and $-2 \leq u \leq 5$.
- (g) $u'(t) = t \sin(u) - t^2/4$, $-2 \leq t \leq 5$ and $-2 \leq u \leq 5$.
- (h) $u'(t) = \cos(u+t)$ for $-2 \leq t \leq 5$ and $-2 \leq u \leq 5$.

Exercise 2.3.3 For each ODE sketch a phase portrait by hand, following the procedure of Examples 2.10, 2.11, and 2.12. Classify each fixed point as asymptotically stable or unstable. Use the result to sketch solutions for the given initial conditions on pair of tu axes with a reasonable range for u .

- (a) $u'(t) = -u(t)$, sketch solutions with $u(0) = 2$ and $u(0) = -2$.
- (b) $v'(t) = 11 - 2v(t)$, sketch solutions with $v(0) = 0$ and $v(0) = 15$.
- (c) $v'(t) = 11 - kv(t)$ (k a positive constant), sketch solutions with $v(0) = 0$ and $v(0) = 15/k$.
- (d) $u'(t) = -(u(t) - 1)(u(t) - 3)$, sketch solutions with $u(0) = 1/2$, $u(0) = 2$, and $u(0) = 4$.
- (e) $u'(t) = u(t)(1 - u(t)) - u(t)/10$ (the harvested logistic equation (1.12) with $r = 1$, $K = 1$, and $h = 1/10$), sketch solutions with $u(0) = 1/2$ and $u(0) = 3/2$. Note only $u \geq 0$ makes physical sense here. What is the long-term fate of the species?
- (f) $u'(t) = u(t)(1 - u(t)) - 2u(t)$ (the harvested logistic equation (1.12) with $r = 1$, $K = 1$, and $h = 2$), sketch solutions with $u(0) = 1/2$ and $u(0) = 3/2$. Note only $u \geq 0$ makes physical sense here. What is the long-term fate of the species?
- (g) $u'(t) = rc_1 - ru(t)/V$ (the conservation law ODE (1.5) with $r, c_1, V > 0$). Recall that this model is only appropriate for $u \geq 0$. Label the fixed point(s) in terms of r, c_1 , and V , and sketch solutions for which $u(0) = 0$ and $u(0) = 2c_1V$.
- (h) $v'(t) = g - kv^2(t)/m$ (the falling body ODE (2.25) with $m, g, k > 0$). Recall that this model is only appropriate for $v \geq 0$. Label the fixed point(s) in terms of m, g , and k , and sketch solutions for which $v(0) = 0$ and $v(0) = 2\sqrt{mg/k}$.

Exercise 2.3.4 In each part below make up an autonomous ODE $u' = f(u)$ (by finding a suitable function f) that has the indicated fixed points with the indicated stability. Hint: review Example 2.13.

- (a) Fixed points $u = 1$ (stable) and $u = 3$ (unstable).
- (b) Fixed points $u = -3$ (stable), $u = 0$ (unstable), $u = 4$ (stable), and $u = 5$ (unstable).
- (c) Fixed points $u = 1$ (semistable) and $u = 3$ (stable). Hint: include a term $(u - 1)^2$ in $f(u)$.
- (d) Fixed points $u = k\pi$ for all integers k , unstable when k is even, stable when k is odd. Hint: tricky problems like this come up *periodically*.

Exercise 2.3.5 Each ODE below depends on a parameter h . Assume the parameter h can be a real number of any sign. A bifurcation occurs in the ODE for a single value of the parameter, call it $h = h^*$. Determine this value, sketch representative phase portraits for each of $h < h^*$, $h > h^*$, and $h = h^*$. Use this to make a bifurcation diagram in the spirit of Figure 2.15.

- (a) $u' = hu - u^2$.
- (b) $u' = hu - u^3$. The resulting bifurcation diagram illustrates a *pitchfork bifurcation*.
- (c) $u' = ru(1 - u/K) - h$, with $r = 1$ and $K = 1$. This is similar to the harvested logistic equation (2.61), but here the harvest rate is a constant h , instead of hu . Assume $h > 0$, and confine your attention to the region $u \geq 0$.

2.4 The Existence and Uniqueness of Solutions

2.4.1 Some Inspiration from Calculus 1

Let's forget about differential equations for a moment and instead go back to Calculus 1, by considering the equation

$$2x - \cos(x) = 0. \quad (2.63)$$

The goal is to find a real number x that satisfies (2.63). Consider trying to solve (2.63) by using the elementary algebraic operations $+$, $-$, \times , \div , as well as square roots, inverse cosines, etc. You will not succeed, yet a plot of the function $f(x) = 2x - \cos(x)$ clearly reveals that (2.63) has a real root between $x = 0$ and $x = 1$, as illustrated in Figure 2.16. It's pretty clear this root is unique (there can be only one) since f appears to be strictly increasing; once f exceeds zero it can never decrease back to zero.

These assertions can be made watertight as follows: the function $f(x)$ is continuous, with $f(0) = -\cos(0) = -1 < 0$ and $f(1) = 2 - \cos(1) > 0$ (since $\cos(1) < 2$ is definitely true). Since $f(x)$ is continuous and changes from negative to positive values between $x = 0$ and $x = 1$, the intermediate value theorem says that $f(x)$ must be zero somewhere between $x = 0$ and $x = 1$. This is true even though we can't write down the solution. As asserted above, this solution is unique. To see this, note that $f'(x) = 2 + \sin(x)$, and so $f'(x) > 0$ for all x (since $\sin(x) \geq -1$ for all x). As such, f is strictly increasing and cannot have two roots, for if $f(a) = 0$ and $f(b) = 0$ with $a \neq b$, then by the mean value theorem it must be that $f'(c) = (f(b) - f(a))/(b - a) = 0$ for some c between a and b . But $f'(c) = 2 + \sin(c) \geq 1$ for all c , so f cannot have two distinct roots.

The moral of the above discussion is that we can use tools from elementary calculus to establish that certain algebraic equations must have a solution, and that the solution is unique, even if we cannot write down the solution in any simple form. This is of value, because knowing that a solution exists and is unique gives us the license and the confidence to go hunting for it using approximate or numerical methods. In the above example, once it has been established that $f(x) = 0$ has a unique

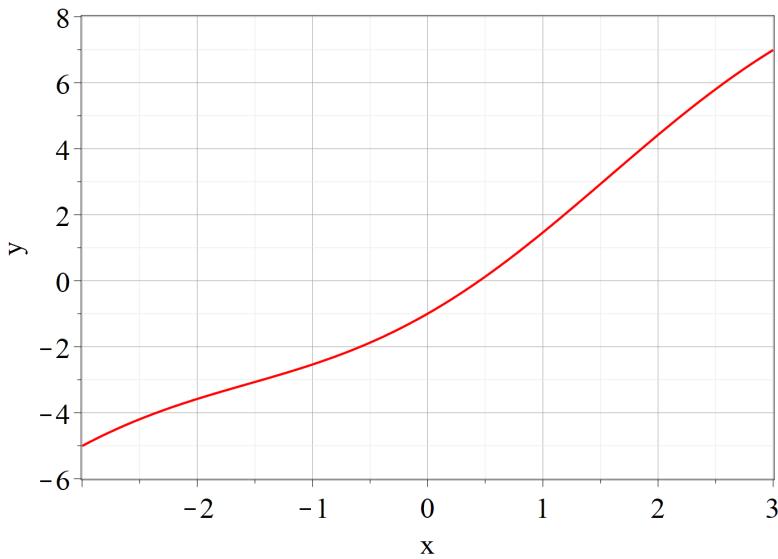


Figure 2.16: Graph of $f(x) = 2x - \cos(x)$.

solution somewhere between $x = 0$ and $x = 1$, a numerical method like Newton's method can be used to approximate the root. Without knowledge of the existence or uniqueness of the solution, we don't know whether numerical methods will succeed, or if they apparently do succeed, whether they found anything real, relevant, or unique.

The same observations hold for differential equations. It is of value to establish conditions under which we can be assured that an ODE has a solution, and that the solution is unique, even when the solution can't be written out in any simple form.

Reading Exercise 2.4.1 Show that the polynomial $p(x) = x^5 + x^3 + x + 5$ has a unique real root (solution to $p(x) = 0$), and this root lies in the interval $-2 < x < 2$.

2.4.2 What Are Solutions to ODEs?

We've been talking about solutions to ODEs, yet have never officially defined what constitutes a solution. This might seem like kind of a silly issue, since in every case so far we can verify the proposed solution works by substituting it into the ODE of interest; all solutions thus far have been quite explicit and elementary, so we've been able to do this. But later in the text we'll see differential equations with solutions that cannot be written out explicitly. Let's take a moment and carefully define what is meant by a solution to an ODE, and several additional useful notions.

Definition 2.4.1 — Solution to an ODE. Let $f(t, u)$ be a function of two variables defined on a rectangle R in the tu plane, where R is defined by the inequalities $a < t < b$ and $c < u < d$. A function $u(t)$ is a **solution** to $u'(t) = f(t, u(t))$ in R if

- $u(t)$ is defined for $a < t < b$ and satisfies $c < u(t) < d$ for each such t (so the graph of u stays in R).
- $u(t)$ is differentiable (hence continuous) for $a < t < b$ and $u'(t) = f(t, u(t))$ at each t .

In many cases the function $f(t, u)$ will be defined for all points (t, u) in the plane, that is, R can be taken as the entire tu plane.

As noted above, in every example so far we could easily verify that solutions fit the above definition. But let's take a look at an example with a slight subtlety.

■ **Example 2.14** Consider the ODE

$$u'(t) = u^2(t) \quad (2.64)$$

with the initial condition $u(0) = 1$. The ODE here is $u' = f(u)$ with $f(u) = u^2$. Solving by separation of variables shows that

$$u(t) = \frac{1}{1-t}. \quad (2.65)$$

The function $u(t)$ is plotted in Figure 2.17, on the range $-3 < t < 3$. The graph of the solution

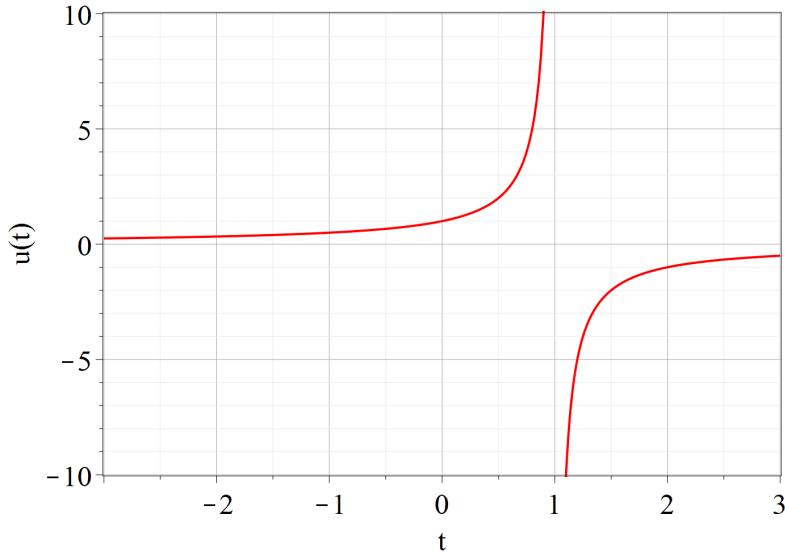


Figure 2.17: Solution to $u'(t) = u^2(t)$ with $u(0) = 1$.

has a vertical asymptote at $t = 1$, as is clear from (2.65). It might be tempting to conclude that (2.64) with $u(0) = 1$ has a solution for all $t \neq 1$, but this cannot be made to fit Definition 2.4.1. The solution must be defined and differentiable at all points in some interval of the form $a < t < b$. In the present case the solution curve that passes through $u(0) = 1$ can be extended in the direction of increasing t up to, but not including, $t = 1$. There is no way to push the solution past this point, as a differentiable function. However, the solution is defined for all $t < 1$. For this initial value problem the solution is defined on the interval $-\infty < t < 1$, but no larger interval. We could take the rectangle R in Definition 2.4.1 as the rectangle defined by $-\infty < t < 1$ and $-\infty < u < \infty$, since $f(u) = u^2$ is defined at all points in this rectangle. ■

The issue raised in Example 2.14 also appeared in Example 2.10 and associated Figure 2.8, as noted in Remark 2.3.1.

Reading Exercise 2.4.2 Go back and look at Example 2.10, and in particular at Figure 2.8. For each of the initial conditions $u(0) = -3/2$, $u(0) = 1$, and $u(0) = 2$, estimate the largest interval $a < t < b$ on which each solution exists. It is possible that $a = -\infty$ or $b = \infty$.

Interval of Existence of a Solution

Given a solution $u(t)$ to an ODE $u' = f(t, u)$ defined on some interval $a < t < b$, it may be possible to extend the solution to a larger interval, possibly even $-\infty < t < \infty$. The largest interval to which the solution can be extended is called the **maximum domain** of the solution. If a solution cannot be extended, it is frequently because the solution has a vertical asymptote, but other things can go

wrong, too. The maximum domain depends on the ODE, of course, but also on the initial condition. In Reading Exercise 2.4.2 you found that the solution with $u(0) = -3/2$ has maximum domain of about $-0.7 < t < \infty$, while the solution with $u(0) = 1/2$ has maximum domain $-\infty < t < \infty$, and $u(0) = 2$ has maximum domain $-\infty < t < 0.5$, roughly. When the solution has a vertical asymptote at a finite time $t = T$ mathematicians say, informally, that the solution **blows up** at time T .

In Chapter 5 we'll encounter solutions to ODEs that are not differentiable or even continuous, and so don't fit Definition 2.4.1. As a result we will have to reinterpret our notion of solution. But Definition 2.4.1 will do for now.

2.4.3 The Existence-Uniqueness Theorem for ODEs

Under what circumstances can we be assured that an ODE with given initial data has a solution in the sense of Definition 2.4.1? When will the solution be unique? As an example, consider the ODE

$$u'(t) = t \cos(u(t)) - \sin(t) \quad (2.66)$$

with initial condition $u(2) = 3$. This ODE is not linear, nor is it separable. A solution almost certainly cannot be written down in any simple form. Yet a glance at the direction field for (2.66), shown in the left panel of Figure 2.18, ought to convince you that a solution should exist, because a solution curve that follows the arrows can be drawn through the point $t = 2, u = 3$, both forward and backwards in t , as shown in the right panel of Figure 2.18. The curve can be extended in each direction until exiting the picture, via the sides or bottom or top.

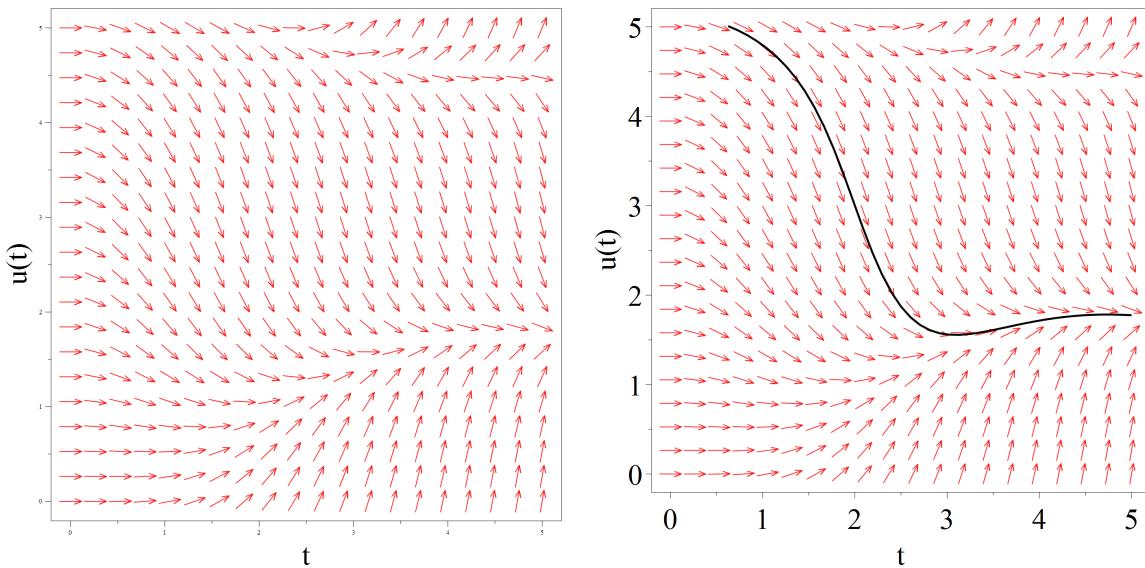


Figure 2.18: Direction field for (2.66) (left) and direction field for (2.66) with solution curve through $t = 2, u = 3$ (right).

You can imagine that if we drew the direction field on an arbitrarily fine grid we could sketch a graph $u = u(t)$ of a function that exactly obeys $u' = t \cos(u) - \sin(t)$ at every point. Moreover, since the direction field points in a single well-defined direction at every point, our hand would be forced: following the direction field arrows gives no choice, and so the solution curve is unique.

These observations are essentially correct, even if they do not constitute a mathematical proof that a unique solution exists. However, there are several restrictions on the types of the ODEs for which we can assert that unique solutions exist. Specifically, the function $f(t, u)$ that defines the right side of $u'(t) = f(t, u(t))$ has to be continuous and $\frac{\partial f}{\partial u}$ should also be continuous on some

rectangle in the tu plane (in more advanced texts weaker conditions on f are permitted). We summarize these facts in the following theorem.

Theorem 2.4.1 — Existence and Uniqueness of Solutions to ODEs. Let R denote a rectangle $a < t < b$ and $c < u < d$ in the tu plane. Suppose that the function $f(t, u)$ is continuous at each point in R and that $\frac{\partial f}{\partial u}$ is continuous at each point in R . Then for any point $t = t_0, u = u_0$ in R , the ODE $u'(t) = f(t, u(t))$ has a unique solution with $u(t_0) = u_0$ on some interval $t_0 - \delta_1 < t < t_0 + \delta_2$ with $\delta_1 > 0$ and $\delta_2 > 0$.

A fairly accessible proof of Theorem 2.4.1 can be found in [24].

■ **Example 2.15** To illustrate Theorem 2.4.1, consider the right panel of Figure 2.18, where a solution to (2.66) with $u(2) = 3$ is shown. Graphically, this solution exists up to at least $t = 5$, where it exits the right side of the graph, and backward in t to roughly $t = 0.7$, where it exits the top. That is, a solution exists for at least $0.7 < t < 5$. Theorem 2.4.1 can be used to prove these assertions. Take R as the rectangle defined by $0 < t < 5$ and $0 < u < 5$, and note that $f(t, u) = t \cos(u) - \sin(t)$ is a continuous function on this rectangle; in fact f is continuous on the whole tu plane. Also, $\frac{\partial f}{\partial u} = -t \sin(u)$ is also continuous on R , and the whole tu plane. By Theorem 2.4.1 there is solution to $u' = f(t, u)$ with $u(2) = 3$ that exists on some time interval $2 - \delta_1 < t < 2 + \delta_2$, and this solution is unique.

By using more advanced techniques in this case it can be shown that the solution exists for $-\infty < t < \infty$, but the point here is that it exists on some interval containing $t = 2$. ■

With a few notable exceptions, every ODE in this book will satisfy the conditions of the existence-uniqueness theorem, and so it will be possible to assert that the solutions with specified initial conditions exist and are unique, even when they cannot be exhibited explicitly. Even in the exceptional cases, other considerations will allow us to conclude a unique solution exists.

Reading Exercise 2.4.3 The ODE

$$v'(t) = g - kv^r(t)/m \quad (2.67)$$

can be used to describe an object falling in the presence of air resistance, for $v > 0$. Here $g > 0$ is gravitational acceleration, m is the mass of the object, and $r \geq 1$ is a real number. Equation (2.25) was the special case $r = 2$, while the case $r = 1$ was explored in Exercise 2.2.9. The ODE (2.67) can't be solved in any simple form for a general choice of r . Use Theorem 2.4.1 to show that if $r \geq 1$, there is a unique solution to (2.67) for any initial data $v(0) = v_0$ when v_0 is positive.

Implications for Sketching Solutions

The existence-uniqueness theorem 2.4.1 drives an important geometric conclusion concerning solution curves to an ODE $u' = f(t, u)$: two distinct solution curves cannot cross or even touch each other. That is, the situations in Figure 2.19 are impossible, if both curves in each panel represent solutions to an ODE that satisfies the hypotheses of Theorem 2.4.1.

The situation in the left panel, in which both curves pass through a common point $t = t_0, u = u_0$, is clearly impossible with rather casual reasoning: If both curves are solutions to $u' = f(t, u)$ then at the point of intersection we must have $u'(t_0) = f(t_0, u_0)$, so both curves must have the same slope, $f(t_0, u_0)$. By inspection, this is clearly false. The situation in the right panel is more delicate, for here the curves touch and have the same slope. The existence-uniqueness theorem precludes this possibility as well, but the analysis is more subtle.

■ **Example 2.16** If the hypotheses of Theorem 2.4.1 are not met, then it may be the case that no solution to the initial value problem exists, or there may be multiple solutions. As an example, consider the initial value problem

$$u' = f(u)$$

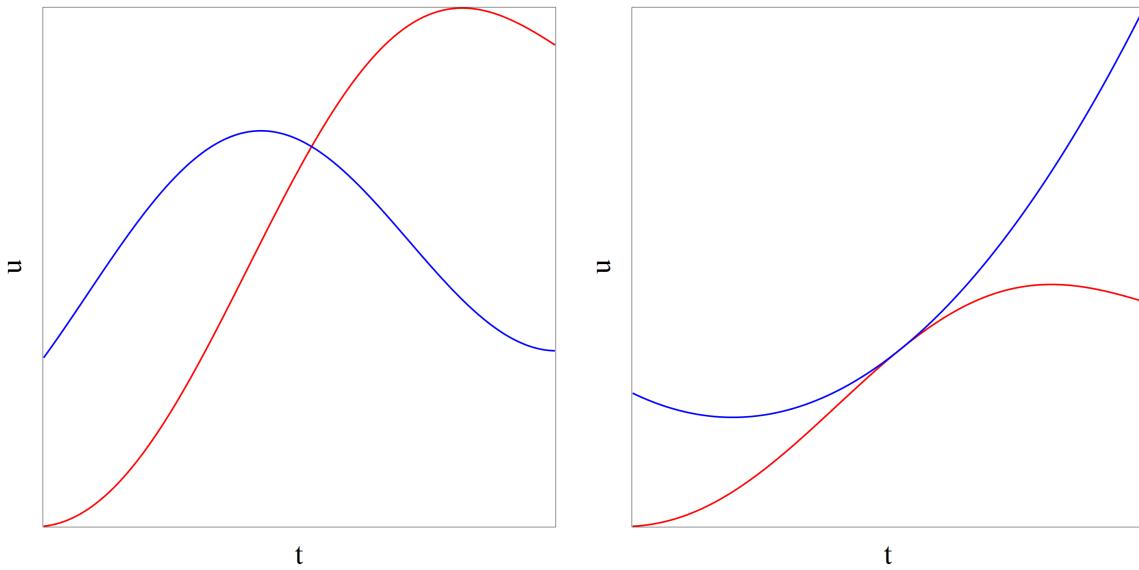


Figure 2.19: Left panel: crossing solution curves for an ODE. Right panel: solutions touching with the same slope. Both scenarios are impossible if Theorem 2.4.1 applies.

with $f(u) = 3u^{2/3}$ and $u(0) = 0$. We do have to be a bit careful with the definition of $f(u)$. To compute $f(u)$ for any real number u , first compute $u^{1/3}$ (every real number has a unique real cube root) and then square and multiply by 3, so $f(u) = 3(u^{1/3})^2$. You can check that $u(t) = 0$ (the zero function) satisfies $u' = f(u)$ with $u(0) = 0$. But so does the function $u(t) = t^3$. The function $f(u)$ is continuous (plot it) but $\frac{\partial f}{\partial u} = 2|u|^{2/3}/(3u)$ is not continuous or even defined at $u = 0$, which violates the hypotheses of Theorem 2.4.1.

In general, the continuity of f is enough to guarantee the existence of at least one solution to the initial value problem, but as this example demonstrates, it is not enough to guarantee that any solution is unique. ■

2.4.4 Exercises

Exercise 2.4.1 Use Theorem 2.4.1 to show that each initial value problem $u' = f(t, u)$, $u(t_0) = u_0$ has a unique solution for some t interval around t_0 . Make sure to specify what f is, and why it has the required properties.

- (a) $u'(t) = u(t) + 3$, $u(0) = 3$
- (b) $u'(t) = -u^2(t) + \sin(t)$, $u(1) = 4$
- (c) $u'(t) = 1/u(t)$, $u(2) = 2$
- (d) $u'(t) = ru(t)(1 - u(t)/K)$, $u(t_0) = u_0$ (the logistic equation).

Exercise 2.4.2 Consider a very general form of Newton's law of cooling given by

$$u'(t) = h(u(t) - A) \tag{2.68}$$

for the temperature $u(t)$ of an object, where $h(x)$ is a continuously-differentiable function for all x and A is the ambient temperature. The usual Newton's law of cooling, equation (2.14), is the special case $h(x) = -kx$.

- (a) Use Theorem 2.4.1 to show that (2.68) has a unique solution for any initial condition $u(0) = u_0$. Hint: the ODE is $u' = f(t, u)$ with $f(t, u) = h(u - A)$.
- (b) Why would the condition $h(0) = 0$ make sense to impose on h ? Hint: if $u(0) = A$ what should the solution $u(t)$ equal for all t ? What is $u'(t)$?
- (c) Why would it be reasonable to impose the condition that h must be a decreasing function? Hint: if $u_1(t)$ and $u_2(t)$ are solutions to $u' = h(u - A)$ with $u_1(0) > u_2(0) \geq A$ (object 1 is hotter than object 2), how should $u'_1(0)$ and $u'_2(0)$ be related? What if $A \leq u_1(0) < u_2(0)$?

Exercise 2.4.3 Find the solution to each initial value problem and then use it to find the maximum domain of the solution.

- (a) $u'(t) = 2 - u(t)$, $u(0) = 2$
- (b) $u'(t) = 1 + u^2(t)$, $u(0) = 0$
- (c) $u'(t) = e^{u(t)}$, $u(0) = 0$
- (d) $u'(t) = 1/u(t)$, $u(0) = 3$

2.5 Modeling Projects

In this section we offer three modeling opportunities, based on projects from the SIMIODE website [9]. The projects concern the law of mass action for modeling chemical reactions, the behavior of a bullet as it moves through water, and mathematical models for loans.

2.5.1 Project: Money Matters 2

This modeling project is based on the SIMIODE Modeling Scenario “Finance—Savings and Loans,” [119].

Borrowing money to pay for things is one of the inescapable privileges and curses of adulthood. Buying a home usually involves borrowing a large sum of money. The options available when one shops around for financing a home purchase are dizzying: loan payback periods ranging from 10 to 30 years, fixed and variable interest rates, and flexible down payments, all make it hard to figure out the best deal. Many people who obtain a home loan are surprised at how slowly the amount they owe decreases, despite making substantial monthly payments. Differential equations can be used to model these situations and come to some conclusions on which to base financial decisions.

Monthly Payments

To begin, let’s suppose you take out a loan of \$250,000 to buy a house. Typically this means you have to offer about \$50,000 as a down payment, but we are not concerned with this here. The interest rate on the loan is 3 percent annually and the loan term is 15 years. What exactly does this mean? In what follows we do not model any additional payments you might make, e.g., points, escrow, or mortgage insurance. We’ll just focus on the basics: you borrow money, the bank charges interest, you have to pay the interest and pay back the amount you borrowed.

Let’s measure time in months, 12 per year, with time 0 as the moment when you acquire the loan and interest begins accruing. At the end of the first month the amount of interest charged is

$$\text{interest in first month} = \frac{1}{12}(0.03)(250,000) = \$625.00.$$

That’s how much interest you owe at the end of the first month. At that time you will also make a payment. The payments must cover the interest and some portion of **principal** (the money you initially borrowed), and are calculated so that at time 180 months (15 years) the amount you owe is

zero. In this case that amount, as you can check below, is \$1726.45 per month. Thus at the end of the first month the amount you owe is

$$\begin{aligned}\text{balance at end of month 1} &= 250,000 + \frac{1}{12}(0.03)(250,000) - 1726.45 \\ &= \$248,898.55.\end{aligned}$$

Of the \$1726.45 you paid, \$625.00 went to pay interest, the other \$1101.45 went to pay off the actual principal.

This process repeats in the next month, with a balance of \$248,898.55 in place of \$250,000. Your payment of \$1726.45 remains the same each month, at least in this type of loan. Thus at the end of the second month you will be charged $\frac{1}{12}(0.03)(248898.55) = \622.25 in interest, where we are rounding to the nearest penny. At the end of the second month you make your payment and your balance is then

$$\begin{aligned}\text{balance at end of month 2} &= 248898.55 + \frac{1}{12}(0.03)(248898.55) - 1726.45 \\ &= \$247,794.35.\end{aligned}$$

The process repeats for another 178 months.

If $p_0 = 250000$ denotes the loan amount, $r = 0.03$ the annual interest rate, $b = 1726.45$ the monthly payment, and p_k the amount owed at the end of month k , then the above computations can be summarized as

$$p_k = \left(1 + \frac{r}{12}\right)p_{k-1} - b. \quad (2.69)$$

Modeling Exercise 5.1.1 Equation (2.69) is called a **difference equation**. Use (2.69) to compute p_2, p_3, \dots, p_{180} . We recommend you use technology; a spreadsheet would do nicely. Alternatively, pseudocode for this computation is shown in Figure 2.20. No adjustments are made here for rounding up to the nearest cent, but you can easily do that; it makes little difference. And this payment, \$1726.45, leaves a balance of 93 cents after 180 months. Verify this.

```
begin
    r:=0.03;
    p[0]:=250000;
    b:=1726.45;
    for k from 1 to 180 do
        p[k]:=(1+r/12)*p[k-1]-b;
    end do;
return
```

Figure 2.20: Pseudocode for loan computation.

Compounding Continuously

Suppose that instead of computing the interest monthly (12 times per year) interest is computed daily, or more conveniently, 360 times per year, or 30 times per month. The daily payment is $b/30$ and (2.69) can be replaced by

$$p_k = \left(1 + \frac{r}{360}\right)p_{k-1} - \frac{b}{30},$$

where now p_k is the balance at the end of day k . In this case the balance at the end of one year is \$236,582.89, versus \$236,599.34 when computed on a monthly basis. After ten years the balance computed daily is \$95,866.53, versus \$96,081.82 for monthly compounding. The daily compounding yields a final balance of -\$374.18 instead of zero. It seems that compounding more frequently makes little difference in the end.

What if compounding was performed n times per year? In this case

$$p_k = \left(1 + \frac{r}{n}\right) p_{k-1} - \frac{12b}{n}. \quad (2.70)$$

Here p_k denotes the balance at the end of k time periods, each of length $1/n$ years, or $12/n$ months. Note that $12b$ is the amount paid annually.

If we increase n toward infinity, (2.70) models the situation in which interest is compounded continuously and payments are made at a constant continual rate per unit time. What does (2.70) become in this case?

Modeling Exercise 5.1.2 Show that (2.70) can be written as

$$\frac{p_k - p_{k-1}}{1/n} = rp_{k-1} - 12b. \quad (2.71)$$

Modeling Exercise 5.1.3 The quantity p_k is the balance of the loan at time $t = k/n$ years, since each iteration of (2.70) steps time forward $1/n$ years. If we consider the loan balance as a function of time t this means that $p_k = p(k/n)$. Show that (2.71) can be expressed as

$$\frac{p(k/n) - p(k/n - 1/n)}{1/n} = rp(k/n - 1/n) - 12b$$

or better yet, as

$$\frac{p(t) - p(t - \Delta t)}{\Delta t} = rp(t - \Delta t) - 12b \quad (2.72)$$

where $t = k/n$ and $\Delta t = 1/n$.

Modeling Exercise 5.1.4 Argue that in the limit $n \rightarrow \infty$ (2.72) becomes

$$p'(t) = rp(t) - 12b. \quad (2.73)$$

Assume that p is a continuous function, so $p(t - \Delta t) \rightarrow p(t)$ as $\Delta t \rightarrow 0$.

Equation (2.73) is the differential equation that models the loan on the assumption of continuous compounding. Time t is in years, and $12b$ is the rate at which the loan is repaid on an annual basis. If p has the dimension of value, say $[p] = V$ then the interest rate r has dimension $[r] = T^{-1}$ and $[12b] = VT^{-1}$.

Modeling Exercise 5.1.5 Sketch a phase portrait for (2.73) under the assumption that r and b are unspecified constants and $p \geq 0$. What practical interpretation can you give to the fixed point?

Modeling Exercise 5.1.6 Solve (2.73) with initial condition $p(0) = 250000$ dollars, $r = 0.03$ per year, and $b = 1726.45$ per month. Use this to compute the loan balance at times $t = 5$ and $t = 10$ years, and compare to the balances when compounding is done monthly, namely \$178,794.88 (at 5 years) and \$96081.81447 (at 10 years). They should be fairly close, within a fifth of a percent or less.

Modeling Exercise 5.1.7 Solve $p(t) = 0$ for t to find that time $t = T$ when you pay off the loan.

Modeling Exercise 5.1.8 How much interest do you pay over the life of the loan? To compute this, note that interest accrues at a rate of $rp(t)$ dollars per year, continually; this is the first term on the right in (2.73). The total interest paid should therefore be the accumulation or sum

$$\int_0^T rp(t) dt, \quad (2.74)$$

where T , the time the loan is paid off, is from the last exercise. Show that the expression in (2.74) has the dimension V (value), or units of dollars. Compute the integral in this case, with $p(0) = 250000$ dollars, $r = 0.03$ per year, and $b = 1726.45$ per month..

The exercises above should illustrate that the continuously compounded model for the loan agrees very closely with the more standard discrete model of monthly compounding, but the continuous ODE model has the advantage of being easier to manipulate. To convince you, here are a few scenarios to consider. You'll want to keep these principles and techniques in mind when you take out a large loan for a house or car in the future.

Modeling Exercise 5.1.9 Show that the solution to the ODE (2.73) with initial condition $p(0) = p_0$ and with r, b , and p_0 undefined is

$$p(t) = p_0 e^{rt} + \frac{12b}{r} (1 - e^{rt}). \quad (2.75)$$

Modeling Exercise 5.1.10 Suppose you take out a 30-year mortgage instead of a 15-year. The rate on 30-year mortgages is usually a few tenths of a percent higher, so let us use $r = 0.033$. Assume as before that $p_0 = 250000$. You can compute the monthly payment by substituting $p_0 = 250000, r = 0.033$ into (2.75), then set $p(30) = 0$ and solve the resulting equation for b . This gives the necessary monthly payment. How does it compare to the 15-year payment?

Modeling Exercise 5.1.11 Use the procedure of Exercise 5.1.8 to compute how much interest you will pay over the 30-year life of the loan.

2.5.2 Project: Chemical Kinetics

This project is based on the SIMIODE Modeling Scenario “Kinetics—Rate of Chemical Reactions” [23].

Chemists often use differential equations to model chemical reactions. The rate at which a reaction proceeds is frequently determined primarily by the concentrations of the reactants, a process usually referred to as the **law of mass action**. For example, concerning hydrogen peroxide it is known that, “The rate of decomposition is dependent on the temperature and concentration of the peroxide, as well as the pH and the presence of impurities and stabilizers” [11]. In this project we consider reactions involving a reactant A in which the reaction rate is dependent upon $[A]$, the concentration of A (perhaps in units of moles per liter) at time t . Note that in this project, as is common in chemistry, the notation $[A]$ denotes the concentration of a chemical species, not a physical dimension. The dimension of a chemical concentration is typically moles per volume.

The rates of reaction studied in the elementary texts are often of the form

$$\frac{d[A]}{dt} = -k[A]^m, \quad \text{with} \quad [A](0) = [A_0],$$

where k is a positive **reaction rate constant** and m is the **order** of the reaction, usually an integer. Here $[A](0) = [A_0]$ is the initial concentration of reactant. Study at this level is frequently restricted to $m = 0, 1$, or 2 , and these are called *zeroth-, first-, and second-order* reactions, respectively.

In most classes and textbooks, you have probably just been given information like rate constants and the order of the reaction. Here, we're also going consider experimental data, and then see how

one would actually go about determining the rate constant and order of a reaction from such data. In Section 3.4 of the next chapter, we'll consider even more sophisticated methods for estimating these parameters.

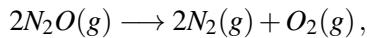
The Order of a Reaction

In practice chemists often have some idea of the order of a reaction, rooted in an understanding of basic chemistry, but sometimes estimating the order and rate constant helps them to understand the nature of the reaction. We proceed with the most common and elementary chemical reaction kinetics models.

Zeroth-Order Reactions

We begin with an excerpt from [133, p. 657].

Zeroth-order reactions are most often encountered when a substance such as a metal surface or an enzyme is required for the reaction to occur. For example, the decomposition reaction of nitrous oxide,



occurs on a hot platinum surface. When the platinum surface is completely covered with N_2O molecules, an increase in the concentration of N_2O has no effect on the rate, since only those N_2O molecules on the surface can react. Under these conditions the rate is a constant because it is controlled by what happens on the platinum surface rather than by the total concentration of N_2O

For a generic reaction we shall use $y(t) = [A] = [A(t)]$, where $[A]$ is the concentration of reactant A (in moles or moles/liter) with time, t , in seconds. Here the rate equation for the decomposition of nitrous oxide, a differential equation, is

$$\frac{dy}{dt} = -ky^0 = -k. \quad (2.76)$$

We will also have an initial condition such as $y(0) = y_0$.

Modeling Exercise 5.2.1 Show that the solution to (2.76) is $y(t) = y_0 - kt$.

Zeroth-order reactions are fairly self-evident from data, for a plot of $y(t)$ against t reveals a linear function starting at $y = y_0$ with a negative slope of $-k$. Such data can be analyzed with any suitable software and the slope of the best-fit line through this data computed. This slope indicates the value of $-k$. Since zeroth-order reactions are quite rare, we move on to first-order reactions.

First-Order Reactions

Let us look at general first-order reactions, namely

$$\frac{dy}{dt} = -ky^1 = -ky \quad (2.77)$$

with initial condition $y(0) = y_0$.

Modeling Exercise 5.2.2 Use the separation of variables technique to show that

$$\ln(y) = -kt + \ln(y(0)). \quad (2.78)$$

(You can skip the last step in separation of variables, solving for y explicitly.)

The next step in separation of variables would be to solve (2.78) explicitly for $y = y(t)$. But chemists are primarily interested in demonstrating that a reaction is first-order, as well as finding the reaction rate constant k , and this is sometimes easier by using (2.78) as it stands. Chemists refer to (2.78) as the **integrated form** of the rate law. In mathematics, however, it is traditional to push on to an explicit solution for $y(t)$. Exponentiating both sides of (2.78) shows that

$$y = y(t) = y(0)e^{-kt}.$$

A First-Order Example

Let's look at a specific example of a first-order reaction, and how one might deduce that the reaction is first-order from experimental data, as well as estimate the rate constant k . In the study of chemical reactions one of the simplest reactions is that of the decomposition of a substance, say hydrogen peroxide (H_2O_2). For example, one might go to the medicine chest to find hydrogen peroxide (or iodine) to flush and clean a cut, only to discover that what is in the bottle does not produce a white froth when applied to the cut, as the medicine is supposed to do. If this is the case, the hydrogen peroxide is old and has lost its powers. This is an example of the decomposition of H_2O_2 into water and oxygen ($2H_2O_2 \rightarrow 2H_2O + O_2$) and we can use the basic law of mass action to conjecture a rate (differential) equation for H_2O_2 . Our conjecture for how the concentration $[H_2O_2]$ changes over time is that

$$\frac{d[H_2O_2]}{dt} = -k \cdot [H_2O_2]^m \quad (2.79)$$

for some number m and rate constant k . We seek two things: (1) to determine if this reaction is first-order, i.e., if $m = 1$, and (2) to determine the value of the rate constant k .

Table 2.3 shows time-concentration data pairs for an experiment concerning the decomposition of H_2O_2 .

Time (seconds)	$[H_2O_2]$ (mol/L)
0	1.00
120	0.91
300	0.78
600	0.59
1200	0.37
1800	0.22
2400	0.13
3000	0.08
3600	0.05

Table 2.3: Data collected on the reaction $2H_2O_2(g) \rightarrow 2H_2O + O_2(g)$, from [133, p. 682].

Modeling Exercise 5.2.3 On the book website [8] you will find Matlab, Maple, Mathematica, and Sage files that contain this data displayed in Table 2.3. In this exercise we will use the various softwares' capabilities for fitting lines and curves to data to estimate parameters like k and m in (2.79).

- Create a vector called `log_of_data` that contains the natural log of the concentration of H_2O_2 data from Table 2.3 in the vector `data`.
- From (2.78) we can see that a first-order reaction will produce a linear relationship between $\ln(y)$ and t . For the data in Table 2.3, plot $\ln[H_2O_2]$ against t and verify that this reaction is first-order.

- (c) Each software environment contains commands for finding the best fit line to a data set; these are demonstrated in the provided files. Use this information to determine the reaction constant k , noting that $\ln(y(0)) = \ln([H_2O_2](0)) = 0$. Plot the resulting line $y = -kt$ on the same axes as the data.
- (d) The fit in (c) should be excellent, but slightly better results may be obtained by including the y -intercept $\ln(y(0))$ in the line-fitting process (rather than forcing it to be zero as in (c)). This puts all the data points on a more equal footing, rather than forcing the line to go through the initial data point. Fit a line of the form $\ln(y) = -kt + b$ to the data, using whatever software you've chosen. Does it improve the fit substantially?

Second-Order Reactions

Now let us look at second-order reactions, modeled by

$$\frac{dy}{dt} = -ky^2$$

with $y(0) = y_0$.

Modeling Exercise 5.2.4 Use the separation of variables technique to show that

$$\frac{1}{y} = kt + \frac{1}{y(0)}. \quad (2.80)$$

You do not need to actually solve for $y(t)$ explicitly.

Equation (2.80) can be solved explicitly for $y = y(t)$, as we will do below, but again the chemist really is interested in determining the nature (order) of the reaction and the parameter k , and will often stop at this point with (2.80). But if desired, an explicit form for $y(t)$ can be found by inverting both sides in (2.80) to obtain

$$y = y(t) = \frac{1}{kt + \frac{1}{y(0)}} = \frac{y(0)}{y(0)kt + 1}.$$

Modeling Exercise 5.2.5 Revisit the decomposition of hydrogen peroxide in Table 2.3 and show that it is not a second-order reaction. Offer as complete a defense as you can—data fitting, plots, verbal argument, etc.

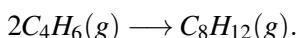
We summarize our conclusions up to this point in Table 2.4.

Order of reaction	Differential Equation	Integrated Form	Solution Form
0	$y'(t) = -k$	$y(t) = y(0) - kt$	$y(t) = y(0) - kt$
1	$y'(t) = -ky(t)$	$\ln(y(t)) = -kt + c$	$y(t) = y(0)e^{-kt}$
2	$y'(t) = -ky^2(t)$	$\frac{1}{y(t)} = kt + \frac{1}{y(0)}$	$y(t) = \frac{y(0)}{y(0)kt + 1}$

Table 2.4: Summary of zeroth-, first-, and second-order kinetics in a differential equation model, the integrated form of the solution through which chemists can possibly obtain a linear plot to confirm the order of the reaction, and a complete solution for a fully developed model.

Decomposition of C₄H₆

Consider the reaction of C₄H₆, butadiene, to form its dimer, a chemical structure formed from two similar sub-units. The reaction is



Some time-concentration data concerning this reaction is shown in Table 2.5.

Time t in s	$[C_4H_6]$ mol/L
0	0.01000
1000	0.00625
1800	0.00476
2800	0.00370
3600	0.00313
4400	0.00270
5200	0.00241
6200	0.00208

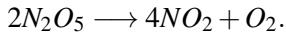
Table 2.5: Data for the reaction of C_4H_6 , butadiene, to form its dimer, from [133, p. 654].

Modeling Exercise 5.2.6 On the book website [8] you will find Matlab, Maple, Mathematica, and Sage files that contain the data in Table 2.5 and support the analysis you are asked to do below.

- (a) Plot the data in Table 2.5.
- (b) From the plot make a conjecture as to the order ($m = 0, 1, 2$) of the reaction.
- (c) Conduct a complete analysis, determining the order and the parameters. Plot the data and the model, and be sure to defend your model. Explain what the order is and what the order is not in consideration of the possibilities $m = 0, 1$, and $m = 2$.

Decomposition of N_2O_5

Consider the reaction describing the decomposition of N_2O_5 , dinitrogen pentoxide:



Some time-concentration data for this reaction is shown in Table 2.6

Time t in s	$[N_2O_5](t)/M$
0	0.310
600	0.254
1200	0.208
1800	0.172
2400	0.141
3000	0.116
3600	0.0964
4200	0.0812
4800	0.0669
6000	0.0464

Table 2.6: Data for the decomposition of N_2O_5 , from [60].

Modeling Exercise 5.2.7 On the book website [8] you will find Matlab, Maple, Mathematica, and Sage files that contain the data in Table 2.6 and support the analysis you are asked to do below.

- (a) Plot the data in Table 2.6.
- (b) From the plot make a conjecture as to the order ($m = 0, 1, 2$) of the reaction.
- (c) Conduct a complete analysis to determine the reaction order and parameters. Plot the data and the model you obtain and defend the model.

2.5.3 Project: A Shot in the Water

This project is based on the SIMIODE Modeling Scenario “A Shot in the Water” [31].

It's a classic scene from an action movie: our heroine is fleeing the bad guys and dives over the side of a boat to escape, then begins to swim away underwater. The villains draw guns and proceed to fire at the heroine, but the view beneath the surface shows the bullets slowing dramatically as they enter. The resistance of the water quickly renders the bullets harmless. Is this realistic? See [4] for some actual experiments and footage.

The Model

The situation can be modeled exactly as we modeled in Section 2.2.1. Specifically, assume the bullet is fired directly downward into the water. As we did previously, we take downward as the positive coordinate direction, and let $v(t)$ denote the bullet's velocity. With quadratic resistance to motion due to the water, Newton's second law of motion leads to the ODE (2.25), reproduced here:

$$v'(t) = g - \frac{k}{m}v^2(t). \quad (2.81)$$

Assume that the bullet enters the water at time $t = 0$. The main difference between Section 2.2.1 and here is that we are now taking $v(0) = v_0 > 0$, rather than $v(0) = 0$ as in Section 2.2.1.

Analysis

Modeling Exercise 5.3.1 Sketch a phase portrait for (2.81); you may restrict your attention to the region $v \geq 0$ (bullet moving downward). What is the fixed point for this ODE in terms of m , g , and k , and what is the fixed point's physical meaning?

Let's assume that bullet has a mass of 55 grains (a typical unit and mass for a rifle bullet), which corresponds to $m = 3.563 \times 10^{-3}$ kg. We'll take $g = 9.81$ meters per second squared. A typical modern high velocity rifle has a muzzle velocity in excess of 1000 meters per second, so let's go with $v_0 = 1000$. The only unknown parameter at the moment is k . One way we can get an estimate of k is outlined in Modeling Exercise 5.3.2.

Modeling Exercise 5.3.2 Suppose, for argument's sake, that the bullet would fall at a terminal velocity of 1 meter per second if dropped in the water. Use this in conjunction with the answer to Modeling Exercise 5.3.1 to estimate k . What is the physical dimension of k ? It would be ideal to do an experiment to estimate k ; go for it, if you have the means, and let the author know the answer. I assume no responsibility for accidents.

Modeling Exercise 5.3.3 Show that under the assumptions above, the general solution

$$v(t) = \sqrt{\frac{mg}{k}} \left(\frac{1 + Ce^{-2t\sqrt{kg/m}}}{1 - Ce^{-2t\sqrt{kg/m}}} \right)$$

to (2.25) or (2.81) that was derived in Section 2.2.4 can be expressed as

$$v(t) = \frac{1 + Ce^{-19.62t}}{1 - Ce^{-19.62t}}. \quad (2.82)$$

Modeling Exercise 5.3.4 Adjust the constant C in the general solution (2.82) to obtain $v(0) = 1000$ meters per second.

Modeling Exercise 5.3.5 Let's suppose the bullet is harmless once $v(t) \leq 10$ meters per second. Use the solution $v(t)$ with $v(0) = 1000$ from Reading Exercise 5.3.4 to determine that time t^* when $v(t^*) = 10$ meters per second. You may have to solve numerically; it might be helpful to plot $v(t)$. Then compute the distance traveled by the bullet from $t = 0$ (when the bullet enters the water) until $t = t^*$. How deep must our heroine be to escape serious injury?

Modeling Exercise 5.3.6 Suppose our estimate of the terminal velocity of the bullet is wrong, say it descends at a terminal velocity of 2 meters per second. How much difference does that make?

3. Numerical Methods for ODEs

3.1 The Need for Numerics

Many ODEs of interest cannot be solved in any simple analytical form, so if quantitative information is needed, we have to turn to numerical methods for approximating solutions. Understanding how these methods work is an important part of the effective application of ODEs to real-world problems. In this chapter we present some classical methods for numerically approximating a solution to an ODE. These methods and extensions form the basis for more modern algorithms that are implemented by number of software packages. The numerical ODE solvers in these packages accept a variety of input arguments that allow the user to specify how accurately the solver tracks the solution. It's helpful to have some understanding of what these input arguments do, so that accurate solutions can be obtained efficiently. We thus spend some time in Section 3.3 discussing error control and adaptive stepsizing, an essential part of any good ODE solver. The goal is not to make you an expert in numerical ODE methods or to teach you to write code, but rather to help you become an informed user of existing codes.

It is also true that one often develops an ODE model for a given physical situation but in which certain parameters like growth rates, spring constants, resistances, etc., are unknown and must be estimated from data. Section 3.4 is devoted to some basic ideas and examples concerning this task, known as parameter estimation.

3.1.1 Logistic Example: Time-Varying Parameters

In [58] the authors consider the logistic equation (1.10) as a model for a population under the general circumstance that the intrinsic growth rate r and carrying capacity K are known functions of time. In this case (1.10) becomes

$$u'(t) = r(t)u(t) \left(1 - \frac{u(t)}{K(t)}\right), \quad (3.1)$$

where $r(t)$ and $K(t)$ are specified functions of time, with $K(t) > 0$. Let's consider this equation under the assumption that $r = 1$ and $K(t) = 1 + 0.25 \sin(2\pi t)$, where t is time in years. This choice for $K(t)$ might represent seasonal variation in the carrying capacity of the environment. In (3.1)

this yields the ODE

$$u'(t) = u(t) \left(1 - \frac{u(t)}{1 + 0.25 \sin(2\pi t)} \right). \quad (3.2)$$

Unfortunately, (3.2) is not separable, nor is it linear. No solution technique we've seen so far allows us to solve this ODE, except in the trivial case that $u(0) = 0$. Because the equation is not separable, it cannot be autonomous, so even the method of phase portraits from Section 2.3 is not applicable.

However, we can sketch a direction field for (3.2), and this is shown in the left panel of Figure 3.1. The right panel shows how to visualize the solution with, for example, initial condition $u(0) = 0.3$. The hypotheses of Theorem 2.4.1 are also straightforward to verify for the ODE (3.2), so as the right panel in Figure 3.1 suggests, a unique solution with $u(0) = 0.3$ exists. What if the value of $u(4)$ for this solution is desired? How can this kind of quantitative information be obtained without drawing pictures? That is the subject of this chapter.

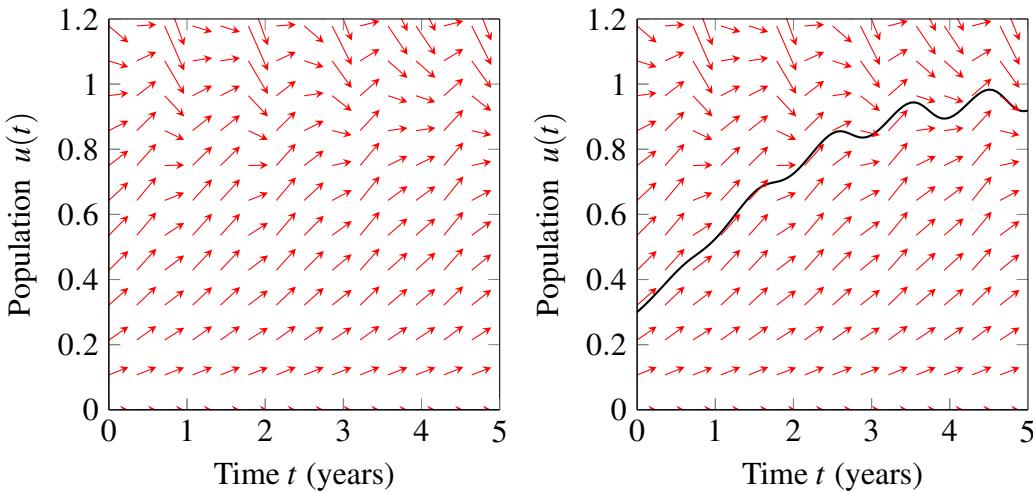


Figure 3.1: Left panel: the direction field for (3.2). Right panel: same, but with solution curve for $u(0) = 0.3$ superimposed.

Reading Exercise 3.1.1 Verify that the existence-uniqueness theorem (Theorem 2.4.1) applies to (3.2) with $u(0) = 0.3$.

3.1.2 Euler's Method

The Tangent Line Approximation

Recall an elementary technique from Calculus 1, the tangent line approximation, or more generally **linearization**. Consider a function $u(t)$ defined on some interval (a, b) , and let t^* be a base point in this interval, as illustrated in Figure 3.2, in which the graph of $u(t)$ is the solid curve.

Assume u is continuously differentiable on (a, b) . The tangent line to the graph of u at the point $(t^*, u(t^*))$ is $y = L(t)$ where $L(t)$ is the function

$$L(t) = u(t^*) + u'(t^*)(t - t^*) \quad (3.3)$$

and is graphed as the dashed blue line in Figure 3.2. The function $L(t)$ is linear with respect to t and designed so that $L(t^*) = u(t^*)$ and $L'(t^*) = u'(t^*)$. As is strongly suggested by Figure 3.2, $L(t)$ is a good approximation to $u(t)$ as long as t is sufficiently close to the base point t^* where the tangent line and graph of u coincide.

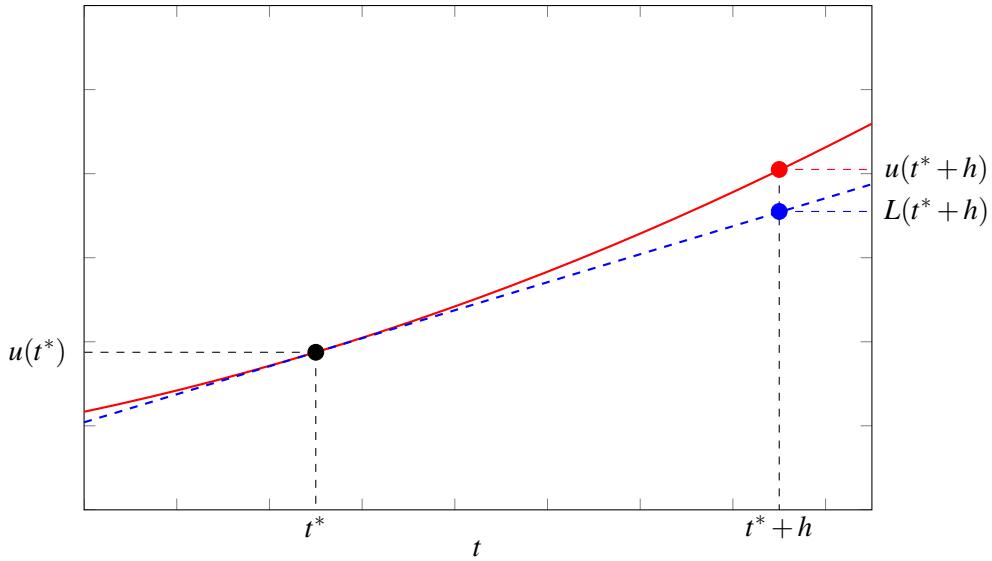


Figure 3.2: Graph of $y = u(t)$ (solid red curve) and the linearization $y = L(t)$ (dashed blue line) of u at $t = t^*$, with points $(t^*, u(t^*))$, $(t^* + h, u(t^* + h))$ and $(t^* + h, L(t^* + h))$ as labeled.

Consider the choice $t = t^* + h$ as illustrated in Figure 3.2. From (3.3) we obtain

$$L(t^* + h) = u(t^*) + u'(t^*)h. \quad (3.4)$$

If h is sufficiently close to zero, then Figure 3.2 suggests that $L(t^* + h)$ is a good approximation to $u(t^* + h)$, and so (3.4) justifies the approximation

$$u(t^* + h) \approx u(t^*) + u'(t^*)h. \quad (3.5)$$

We can use (3.5) with knowledge of $u(t^*)$, h , and $u'(t^*)$ to approximate $u(t^* + h)$, and this forms the basis for a reasonable method to quantitatively track a solution to an ODE $u' = f(t, u)$ starting with a given initial condition.

Reading Exercise 3.1.2 Suppose $u(t) = 2t^2$ and $t^* = 1$. Compute the linearization $L(t)$ as given by (3.3). Then compute $u(t^* + h)$ and $L(t^* + h)$ for each of $h = 1, 0.1, 0.01$, and $h = 0.001$ and observe how the accuracy of the linearization improves as h gets smaller. What is the relation between h and the difference $|u(t^* + h) - L(t^* + h)|$?

Reading Exercise 3.1.3 With $u(t)$ and t^* as in Reading Exercise 3.1.2, compute and simplify the expression $u(t^* + h) - L(t^* + h)$ (leave h undefined). How does this quantity depend on h ?

3.1.3 Evaluate, Extrapolate, Repeat as Necessary

Let's look at how tangent line extrapolation in the form (3.5) can be used to approximate the solution to an ODE. As a first example, consider instead the ODE

$$u'(t) = u(t) - 3t^2 \quad (3.6)$$

with initial condition $u(0) = 1$. The ODE (3.6) can be solved analytically using the integrating factor technique of Section 2.1. The solution is

$$u(t) = 3t^2 + 6t + 6 - 5e^t. \quad (3.7)$$

Knowing this will allow us to examine how well the numerical procedure approximates the exact solution.

To approach the problem numerically, define $f(t, u) = u - 3t^2$, so (3.6) can be written as $u' = f(t, u)$. Define $t_0 = 0$, the initial time, and $u_0 = u(0) = 1$. For this example we will construct a sequence of estimates u_1, u_2, \dots for $u(t)$ at times $t_1 = 0.25, t_2 = 0.5, t_3 = 0.75$, and so on, stepping forward in time increments of size 0.25. This increment is called the **step size** for the numerical solution. The process makes repeated use of (3.5).

To produce an estimate u_1 for $u(t_1)$ use (3.5) with base point $t^* = t_0$ and step size $h = 0.25$ to estimate

$$u(t_1) \approx u(t_0) + hu'(t_0). \quad (3.8)$$

To evaluate the right side above, note that $u(t_0) = u_0 = 1$ is given, as is $h = 0.25$, but what is $u'(t_0)$? This is where the ODE (3.6) comes into play: according to the ODE $u'(t_0) = f(t_0, u_0) = u_0 - 3t_0^2 = 1$, since $t_0 = 0$ and $u_0 = 1$. In (3.8) this yields an estimate u_1 for $u(t_1)$:

$$u_1 = u_0 + h\underbrace{f(t_0, u_0)}_{u'(t_0)} = 1 + (0.25)(1) = 1.25. \quad (3.9)$$

Equation (3.9) comprises one step of what is known as **Euler's method** for numerically approximating the solution to this ODE. Note that u_1 is only an estimate of $u(t_1)$, since the linearization $L(t)$ probably doesn't equal $u(t)$ away from $t^* = t_0$. The true value in this case is $u(0.25) \approx 1.2674$.

The next step is to extrapolate the solution to time $t_2 = 2h = 0.5$. This is done by setting $t^* = t_1 = 0.25$, and $u_1 = 1.25 \approx u(t_1)$ replaces u_0 . In this case (3.5) yields an approximation

$$u_2 = u_1 + h\underbrace{f(t_1, u_1)}_{\approx u'(t_1)}. \quad (3.10)$$

Compare (3.10) to (3.9). With $u_1 = 1.25, h = 0.25$, and $f(0.25, 1.25) \approx 1.0625$, (3.10) yields $u_2 = 1.5156$ as an approximation to $u(0.5)$. The true value is $u(0.5) \approx 1.5064$.

Reading Exercise 3.1.4 Compute u_3 , an estimate of $u(0.75)$, using the above procedure. Your estimate will be given by (3.5) with $t^* = t_2 = 0.5, h = 0.25, u_2 = 1.5156$, and $u'(t_2) \approx f(t_2, u_2)$. Carry all computations to four digits past the decimal. Repeat this process to compute an estimate $u_4 \approx u(1.0)$. In each case compare the estimate to the true value of the solution.

Figure 3.3 shows a plot of the true solution (3.7) (the solid black curve) and the Euler's method iterates based on step size $h = 0.25$, both superimposed on the direction field for the ODE (3.6); the Euler iterates are shown as the solid blue dots, connected by the dashed lines. At each iteration the algorithm extrapolates the solution from the current estimated point (t_k, u_k) to the next time $t = t_{k+1}$, to produce an estimate $u_{k+1} \approx u(t_{k+1})$. If the direction field deviates significantly from this extrapolating line then the next iterate is erroneous. We see this is the case as here in the step from $t_2 = 0.5$ to $t_3 = 0.75$, and from t_3 to $t_4 = 1$. One might think of the true solution as the result of taking infinitesimal steps of size h forward in time, so the true solution's graph tracks the direction field perfectly at all points.

Euler's Method In General

In general Euler's method works as follows. An ODE in the form $u'(t) = f(t, u(t))$ with initial condition $u(t_0) = u_0$ is given. To approximate the solution numerically choose step size h and set $t_k = t_0 + kh$ where $k = 0, 1, 2, \dots$. Approximations $u_k \approx u(t_k)$ for $k = 1, 2, \dots, N$ are constructed according to the linearized approximation

$$u_{k+1} = u_k + hf(t_k, u_k) \quad (3.11)$$

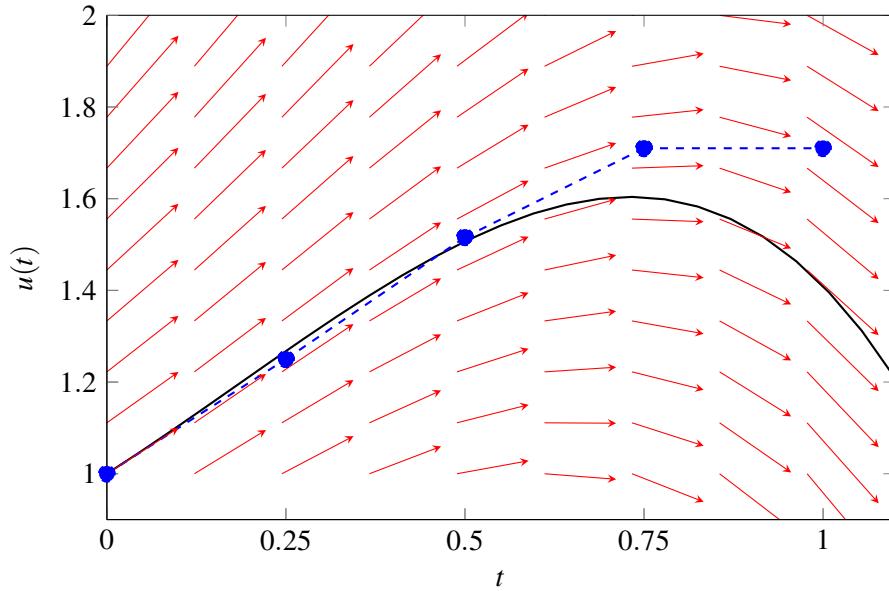


Figure 3.3: True solution $u(t)$ to (3.6) with $u(0) = 1$ (solid black curve), Euler iterates with step size $h = 0.25$ (solid blue dots connected by dashed lines) and direction field for (3.6) (red vectors).

for some chosen value of N . This iterative process is often called **marching** the numerical solution forward in time. Repeated application of (3.11) marches the solution from time $t = t_0$ out to time $t_N = t_0 + Nh$.

■ **Example 3.1** For the ODE (3.1) the computations for Euler's method with step size $h = 0.25$ are shown in Table 3.1, up to time $t = 1.5$ ($N = 6$). Euler's method produces an estimate of the

Iteration	t_k	u_k	$f(t_k, u_k)$
0	0.00	0.3000	0.2100
1	0.25	0.3525	0.2531
2	0.50	0.4158	0.2429
3	0.75	0.4765	0.1738
4	1.00	0.5199	0.2496
5	1.25	0.5823	0.3110
6	1.50	0.6601	0.2244

Table 3.1: Euler's method approximation for (3.1), step size $h = 0.25$.

solution at times $t = t_0, t_0 + h, t_0 + 2h, \dots, t_0 + Nh$. If these points are connected, we obtain a polygonal approximation to the true solution's graph. This is shown in Figure 3.4 out to $t = 5$, superimposed over an extremely accurate approximation to the true solution that is computed using more sophisticated methods that we discuss in the following sections. It looks like Euler's method does a reasonable job. But what if a more accurate solution is desired? ■

3.1.4 The Accuracy of Euler's Method

The most straightforward way to obtain greater accuracy is to run Euler's method with a smaller step size, and so track the direction field and true solution more closely. In general, taking smaller values for the step size h improves the estimate for $u(T)$ for any fixed choice of T , but at what computational cost?

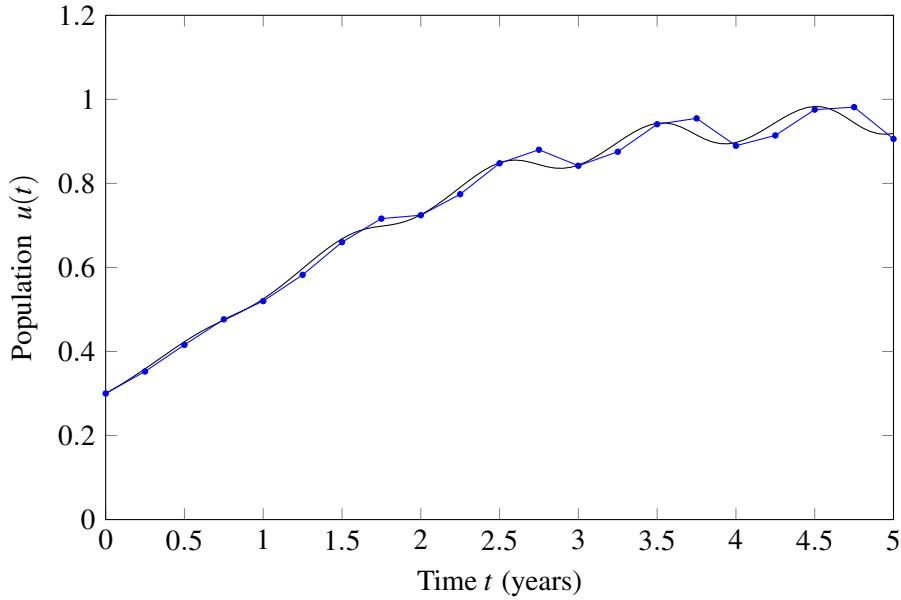


Figure 3.4: Euler's method approximation (blue dots/lines) for (3.1), step size $h = 0.25$, with true solution shown as solid black curve.

■ **Example 3.2** Let's estimate $u(1.0)$ for the solution to $u'(t) = u(t)$ with initial condition $u(0) = 1$, using Euler's method with step sizes $h = 1, h = 0.1, h = 0.01, h = 0.001$, and $h = 0.0001$. The true solution is $u(t) = e^t$ and so in this case $u(1) = e \approx 2.71828128$. This makes it easy to observe how the error in Euler's method behaves in this case. The given values for h require $N = 1, 10, 100, 1000$, and $N = 10000$ steps, respectively. The results are shown in Table 3.2, where $u_h(1.0)$ is the estimate of the true solution value using Euler's method with step size h , and the error is computed as $|e - u_h(1.0)|$. ■

Step size h	Euler estimate $u_h(1.0)$	Error
1.0	2.00000	0.718282
0.1	2.59374	0.124542
0.01	2.70481	0.013468
0.001	2.71692	0.001358
0.0001	2.71815	0.000136

Table 3.2: Euler's method estimate of $u(1) = e$ when applied to $u'(t) = u(t)$ with initial condition $u(0) = 1$ for various step sizes h , with error $|e - u_h(1.0)|$.

An examination of the error in Table 3.2, especially for $h \leq 0.1$, shows that each 10-fold decrease in h results in an approximate 10-fold decrease in the error. This is typical of Euler's method. For most ODEs the error in Euler's method approximation $u_h(T)$ to $u(T)$ for any fixed time $T > t_0$ is proportional to h , at least once h is sufficiently small. That is,

$$|u(T) - u_h(T)| \approx Ch, \quad (3.12)$$

for some constant C . We say that Euler's method is **first-order accurate**: the error is proportional to the first power of the step size.

Informal Analysis of the Error in Euler's Method

It's not hard to see why the error in Euler's method might be expected to follow the pattern of (3.12). The following argument is informal and gives the spirit of the result, but it can be made rigorous.

Suppose u is a continuously differentiable function defined on some interval (a, b) and t^* lies in the interval (a, b) , as illustrated in Figure 3.2. From Taylor's theorem it follows that

$$u(t^* + h) = \underbrace{u(t^*) + u'(t^*)h}_{L(t^*+h)} + \frac{1}{2}u''(s)h^2 \quad (3.13)$$

if $t^* + h$ is in (a, b) . Here s is some number between t^* and $t^* + h$ that is not known. Compare (3.13) to (3.4). The tangent line extrapolation for $u(t^* + h)$ will be in error by an amount $\frac{1}{2}u''(s)h^2$. Since Euler's method is just repeated tangent line extrapolation, (3.13) gives an estimate of the error made in each step. In particular, let $t^* = t_k$ in the k th Euler iteration and suppose $u(t^*) = u(t_k) = u_k$ (so the estimate $u_k \approx u(t_k)$ in Euler's method is perfect at the k th iteration). From (3.13) we see that the error made in the next Euler step is $\frac{1}{2}u''(s)h^2$. This makes geometric sense: recall from Calculus 1 that u'' is related to the curvature of u , and since we're using the tangent line to extrapolate, curvature in the graph of u is the source of this error.

So at each step in Euler's method the error made is $\frac{1}{2}u''(s)h^2$ for some s . If K is a bound on the absolute value of u'' on the interval of interest, that is, if $|u''(t)| \leq K$ on (t_0, T) , then the error made at the k th step is no larger than $\frac{1}{2}Kh^2$. Marching Euler's method from $t = t_0$ to $t = T$ in steps of size h requires $N = (T - t_0)/h$ steps. Taking N such steps, each with error no larger than $\frac{K}{2}h^2$ means that the maximum error made at $t = T$ is comparable to

$$\frac{NK}{2}h^2 = \frac{K(T - t_0)/h}{2}h^2 = \frac{K(T - t_0)}{2}h,$$

where again K is the maximum value of $|u''|$ on $[t_0, T]$.

■ **Example 3.3** To illustrate, consider the ODE $u' = u$ with $u(0) = 1$ on the interval $(0, 1)$ from Example 3.2. We have $t_0 = 0, T = 1$, and K as the maximum of the second derivative of e^t on $0 < t < 1$. It's easy to check that $K = e$ (since $(e^t)'' = e^t$), which yields an error estimate of $|e - u_h(1)| \approx eh/2$. For $h = 0.001$ this is $|e - u_h(1)| \approx 0.001359$, a very reasonable bound on the actual error of about 0.00135 from Table 3.2. ■

As we mentioned, the argument above isn't rigorous. For example, we assumed that $u_k = u(t_k)$ at the k th Euler step; that is, we assumed no error had been made up to the k th iteration. In reality the iterates u_k will drift farther and farther off of the true solution values as k increases. Nonetheless, the conclusion is essentially correct. A more rigorous version of the above analysis can be used to prove the following theorem.

Theorem 3.1.1 — Euler's Method Error. If u_h denotes the approximate solution to $u' = f(t, u)$ with initial condition $u(t_0) = u_0$ produced by Euler's method with step size h and if u'' is bounded on the interval $t_0 \leq t \leq T$ with $T = t_0 + Nh$ for some integer N , then

$$|u_h(T) - u(T)| \leq Ch$$

for some constant C .

For a proof of Theorem 3.1.1 see [18]. The constant C in Theorem 3.1.1 is not typically known, but the main point is that the error in Euler's method is proportional to the step size h . If you want an answer that is ten times more accurate, you will likely need to decrease h by a factor of ten, which means ten times more work to march the solution out to time $t = T$. This may seem reasonable, but it is possible to do much better.

3.1.5 Exercises

Exercise 3.1.1 For each initial value problem apply Euler's method, by hand (but use a calculator), using the indicated step size h and number of steps N . Carry computations to four significant figures. Then compute the value of the true (analytical) solution at time $T = t_0 + Nh$ using the methods of Chapter 2 and compare.

- (a) $u'(t) = u(t) + 3$, $u(0) = 1$, step size $h = 0.5$, $N = 2$ steps.
- (b) $u'(t) = -u(t) + 3t$, $u(0) = 2$, step size $h = 0.25$, $N = 4$ steps.
- (c) $u'(t) = 1/u(t)$, $u(0) = 2$, step size $h = 0.25$, $N = 4$ steps.
- (d) $u'(t) = tu(t)$, $u(1) = 3$, step size $h = 0.1$, $N = 5$ steps.

Exercise 3.1.2 For each initial value problem apply Euler's method, using whatever technology you have available, with the indicated step sizes h , to estimate $u(T)$ for the given value of T . For each step size h , compute the difference between the Euler estimate and the value of the analytic solution at $t = T$, obtained using the methods of Chapter 2. Does Theorem 3.1.1 seem to be accurate? With what value of C ?

- (a) $u'(t) = 1 - u(t)/3$, $u(0) = 2$. Estimate $u(5)$ using step sizes $h = 1, 0.1, 0.01$.
- (b) $u'(t) = te^{-u(t)}$, $u(0) = 1$. Estimate $u(3)$ using step sizes $h = 1, 0.1, 0.01$.
- (c) $u'(t) = u^2(t)$, $u(0) = 2$. Estimate $u(0.5)$ using step sizes $h = 0.5, 0.1, 0.01, 0.001$.
- (d) $u'(t) = 1/u(t)$, $u(0) = 2$. Estimate $u(4)$ using step sizes $h = 1.0, 0.1, 0.01, 0.001$.

Exercise 3.1.3 Apply Euler's method to the ODE $u'(t) = u(t) - t + 1$ with $u(0) = 0$, to estimate $u(1)$ using step sizes $h = 1, 0.1, 0.01$. Then find the analytical solution and compute the error for each step size. Why does this make perfect sense?

Exercise 3.1.4 A variation on the Hill-Keller ODE (1.3) is to take the resistive force as $F_r = -kmv^r(t)$ for some constants $k > 0$ and $r \geq 1$. This leads to the ODE

$$v'(t) = P - kv^r(t), \quad (3.14)$$

under the assumption that $v \geq 0$, where P still is the same propulsive effort as in Section 1.1.2. (Compare (3.14) to (2.67) in Example 2.4.3.) The ODE (3.14) has no simple analytical solution for a general r (except when $r = 1$ and $r = 2$). The equation must be solved numerically.

- (a) Verify that the choice $F = -kmv^r$ leads to the ODE (3.14).
- (b) Sketch a phase portrait for (3.14), limiting your attention to $v \geq 0$. Label the fixed point in terms of the constants P, k , and r .
- (c) Use Theorem 2.4.1 to show that (3.14) has a unique nonnegative solution for $v(t_0) = v_0$ when $v_0 > 0$.
- (d) Take $P = 11$, $r = 3/2$, and $k = 0.258$. Use the phase portrait from (b) to argue that $\lim_{t \rightarrow \infty} v(t) \approx 12.2$ (Bolt's top speed from the data in Table 1.1).
- (e) Use Euler's method with step sizes $h = 1$ and $h = 0.1$ to compute $v(t)$ with initial data $v(0) = 0$ out to time $t = 10$. In each case plot the numerical solution and compare it to the nature of the solution as expected based on the phase portrait.
- (f) Do two steps of $h = 2$ and $h = 5$, each by hand. Can you see why these step sizes are disastrous?

Exercise 3.1.5 Apply Euler's method to the ODE $u'(t) = u^2(t)$ with $u(0) = 1$ to estimate $u(1)$ using step sizes $h = 1, 0.1, 0.01, 0.001$. Then estimate $u(2)$ using step sizes $h = 0.1, 0.01$, and 0.001 . Explain what's going on. Hint: compute the analytical solution using separation of variables. Then recall Definition 2.4.1 and the notion of the maximum domain of a solution from Section 2.4.2.

Exercise 3.1.6 Consider the linear ODE

$$u'(t) = u(t) - \sin(t) + \cos(t).$$

- (a) Find a general solution.
- (b) Find the solution with initial condition $u(0) = 0$.
- (c) Sketch a direction field on the range $0 \leq t \leq 10, -5 \leq u \leq 5$, and superimpose the solution with $u(0) = 0$ on this direction field.
- (d) Apply Euler's method with step sizes $h = 1, 0.1, 0.01, 0.001$ with initial condition $u(0) = 0$ to estimate $u(10)$. Explain the poor estimates for $u(10)$ in light of the direction field from (c) and general solution from (a). Hint: what happens if Euler's method ever strays from the analytical solution curve? It might be helpful to plot the Euler iterates u_0, u_1, u_2, \dots for the case $h = 0.001$.

Exercise 3.1.7 This problem illustrates that if the step size is too large, Euler's method isn't just inaccurate—it may actually blow up, even if the true solution to the ODE decays. This should also be apparent in part (f) of Exercise 3.1.4.

Consider the differential equation $u'(t) = -10u(t)$ with $u(0) = 1$.

- (a) Find the analytical solution to this initial value problem, and show that it decays to zero as $t \rightarrow \infty$.
- (b) Apply Euler's method with step size $h = 0.1$ to estimate $u(5)$. Hint: after the first step the computations should be trivial.
- (c) Apply Euler's method with step size $h = 0.2$ to estimate $u(5)$. Hint: a simple pattern should emerge.
- (d) Apply Euler's method with step size $h = 1$ to estimate $u(5)$. Hint: again, there is a pattern.
- (e) Suppose we apply Euler's method with step size h . Show that the k th iterate u_k (an approximation to $u(kh)$) is given by

$$u_k = (1 - 10h)u_{k-1},$$

so that with initial iterate $u_0 = 1$ we have

$$u_k = (1 - 10h)^k. \tag{3.15}$$

Equation (3.15) should yield results in accordance with parts (b)-(d).

- (f) The analytical solution from part (a) decays to zero. How large can we take $h > 0$ in (3.15) to (at least) obtain decay to zero? Interpret the results of parts (b)-(d) in light of this analysis.

3.2 Improvements to Euler's Method

Shortcomings of Euler's Method

Let's take a look at the nature of the error in Euler's method, as a prelude to strategies for mitigating this error. We'll use the ODE $u'(t) = t + u(t)/2$ as a example. A typical step in Euler's method for this ODE is illustrated in Figure 3.5. For the k th iteration (k doesn't matter) we use $t_k = 0.2$, $u_k = 0.2$, shown as a black dot, and assume the true solution (shown as the solid black curve) passes precisely through this point. The dot at $(0.6, 0.418)$ on this curve highlights the true solution value. Euler's method with step $h = 0.4$ is used to extrapolate to $t = t_{k+1} = 0.6$ by using (3.11). This yields $u_{k+1} = u_k + hf(t_k, u_k)$ with $f(t, u) = t + u/2$, so here $u_{k+1} = 0.2 + (0.4)(0.3) = 0.32$. This linear extrapolation starting from $t_k = 0.2$, $u_k = 0.2$ is shown in Figure 3.5 as the dashed blue line segment, with the other end at $t_k = 0.6, u_k = 0.32$. The direction field for $u' = f(t, u)$ is also shown.

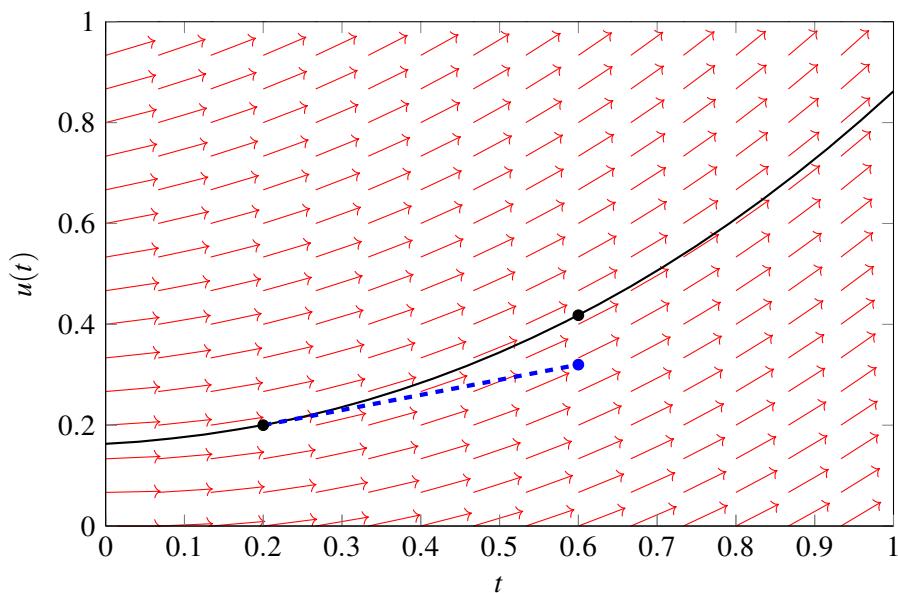


Figure 3.5: A typical Euler's method step (dashed blue segment) from $t_k = 0.2$ to $t_{k+1} = 0.6$, compared to the true solution (solid black curve) for the ODE $u'(t) = t + u(t)/2$ satisfying $u(0.2) = 0.2$.

The true solution is perfectly tangent to the direction field at all points. However, the Euler step is not tangent to the vector field, except at the initial point $t = t_k, u = u_k$. For $t > t_k$ the Euler extrapolation extends the solution linearly and takes no account of the fact that the direction field is changing. It should not be surprising that the estimate u_{k+1} is off a bit from the correct value for $u(t_{k+1})$.

3.2.1 Improving Euler's Method

The Euler extrapolation can be improved by recognizing that the direction field changes from $t = t_k$ to $t = t_k + h$ and then incorporating this observation into the extrapolation process. We again illustrate by using the ODE $u'(t) = t + u(t)/2$ with a typical step beginning at $t_k = 0.2$, $u_k = 0.2$, with the goal of extrapolating the solution to $t_{k+1} = t_k + h$ using step size $h = 0.4$. In Figure 3.6 the true solution through the point $t_k = 0.2, u_k = 0.2$ is shown as a solid black curve. The direction field for this ODE is also plotted.

The process involves three distinct stages. The first two are standard Euler steps starting from (t_k, u_k) , and the third stage involves combining the Euler steps to form an improved estimate for

u_{k+1} . The details:

- From (t_k, u_k) (the solid black dot in Figure 3.6 at coordinates $(0.2, 0.2)$) take the standard Euler step from t_k to $t_{k+1} = t_k + h$, as dictated by (3.11). However, this step is only a provisional estimate for $u(t_{k+1})$. Let w denote this estimate of $u(t_{k+1})$, so

$$w = u_k + hf(t_k, u_k). \quad (3.16)$$

In Figure 3.6 the step dictated by (3.16) is illustrated by the dashed blue line segment that starts at $(0.2, 0.2)$; w is the vertical coordinate of the right tip of this dashed blue segment, at coordinates $(0.6, 0.32)$, shown as a blue circle. This estimate is already significantly off of the true solution curve.

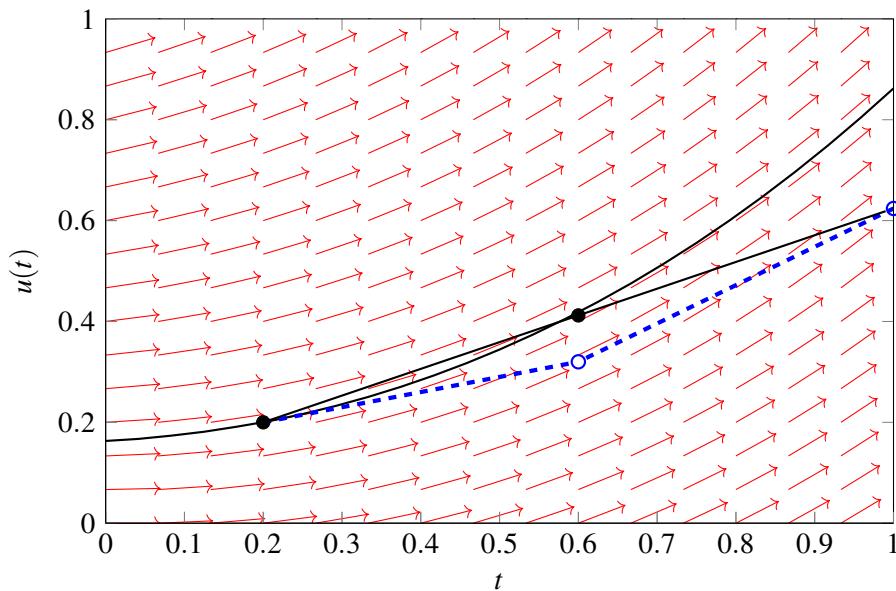


Figure 3.6: Illustration of the improved Euler's method step. The true solution to the ODE through $(0.2, 0.2)$ is shown as the solid black curve. Two Euler iterations are shown as dashed blue line segments. The improved Euler's method estimate is the midpoint of the solid black line segment, highlighted with a black dot near $(0.6, 0.412)$.

- Take a second Euler step of step size h starting at the point $(t_k + h, w)$ and extrapolate along the line with slope $f(t_k + h, w)$. This is illustrated by the rightmost dashed blue line segment in Figure 3.6. The point at the right end of this segment is highlighted by a blue circle, with t coordinate $t_k + 2h$ (here $t_k + 2h = 1$) and vertical coordinate

$$\tilde{w} = w + hf(t_k + h, w). \quad (3.17)$$

In Figure 3.6, $\tilde{w} \approx 0.624$.

- Construct an improved estimate u_{k+1} for $u(t_k + h)$ by taking the average value of u_k and \tilde{w} . From a geometric perspective this can be viewed as taking u_{k+1} to be the vertical coordinate of the midpoint of the line segment joining (t_k, u_k) to $(t_k + 2h, \tilde{w})$. This segment is shown as a solid black segment in Figure 3.6, and the midpoint as the solid black dot. The vertical

coordinate of this midpoint is the final estimate for u_{k+1} , and is

$$\begin{aligned} u_{k+1} &= \frac{u_k + \tilde{w}}{2} \\ &= \frac{u_k + w + hf(t_k + h, w)}{2} \\ &= u_k + h \left(\frac{f(t_k, u_k) + f(t_k + h, w)}{2} \right), \end{aligned} \quad (3.18)$$

where we have made use of (3.16) and (3.17). Equation (3.18) can also be viewed as linear extrapolation from $t = t_k$ to $t = t_k + h$, namely

$$u_{k+1} = u_k + hm \quad (3.19)$$

with slope

$$m = \frac{f(t_k, u_k) + f(t_k + h, w)}{2} \quad (3.20)$$

instead of the Euler slope $f(t_k, u_k)$.

The intuitive idea is that by evaluating $f(t, u)$ at both $t = t_k$ and $t = t_{k+1}$ with suitable choices for u and then averaging, the resulting slope m in (3.20) is more representative of the behavior of the solution over the interval $t_k \leq t \leq t_{k+1}$ than $f(t_k, u_k)$ alone. In Figure 3.6 it is clear that the improved Euler's method estimate, which appears to be right on the solution curve, is much superior to the standard Euler's method estimate.

■ **Example 3.4** Let's consider some specific numbers relevant to Figure 3.6. That figure is based on the specific choices $t_k = 0.2$, $u_k = 0.2$, and $h = 0.4$, with $f(t, u) = t + u/2$; here $u_k = u(t_k)$ exactly, that is, we assume no error at the k th iteration, for simplicity. The solution curve (solid black) is $u(t) = 4.6e^{(5t-1)/10} - 2t - 4$ and the value of this analytical solution at $t = t_{k+1} = 0.6$ is $u(0.6) \approx 0.418$. Euler's method produces the estimate

$$u(t_{k+1}) \approx u_k + hf(t_k, u_k) = 0.2 + 0.4f(0.2, 0.2) = 0.32.$$

The improved Euler's method gives (with intermediate computations shown)

$$\begin{aligned} w &= u_k + hf(t_k, u_k) = 0.2 + 0.4f(0.2, 0.2) = 0.32, \\ m &= \frac{f(t_k, u_k) + f(t_{k+1}, w)}{2} = \frac{f(0.2, 0.2) + f(0.6, 0.32)}{2} = 0.53, \\ u_{k+1} &= u_k + hm = 0.412, \end{aligned}$$

using (3.19) and (3.20). This is considerably more accurate than the Euler estimate. ■

3.2.2 The Improved Euler Method

Steps 1 to 3 above can be iterated to march the solution forward in time. The resulting algorithm is an improvement to Euler's method and is called—wait for it—the **improved Euler's method**. Let's consider an example.

■ **Example 3.5** In this example we perform two steps of the improved Euler's method with $h = 0.5$ to estimate $u(1.0)$, where u satisfies $u'(t) = u(t) + t + 1$ with initial condition $u(0) = 2$. For reference, the true solution here is $u(t) = 4e^t - t - 2$ and $u(1) = 4e - 3 \approx 7.873$. We use (3.19)

and (3.20) to march the solution out in time (equivalent to the single step (3.18), but we prefer to illustrate the intermediate details).

To begin note that $t_0 = 1, t_1 = 0.5$, and $t_2 = 1.0$. Set $u_0 = u(0) = 2$. The first iteration of the improved Euler's method is

$$\begin{aligned} w &= u_0 + hf(t_0, u_0) = 2 + 0.5f(0, 2) = 3.5, \\ m &= \frac{f(t_0, u_0) + f(t_1, w)}{2} = \frac{f(0, 2) + f(0.5, 3.5)}{2} = 4.0, \\ u_1 &= u_0 + hm = 2 + (0.5)(4.0) = 4.0. \end{aligned}$$

The second iteration produces

$$\begin{aligned} w &= u_1 + hf(t_1, u_1) = 4 + 0.5f(0.5, 4) = 6.75, \\ m &= \frac{f(t_1, u_1) + f(t_2, w)}{2} = \frac{f(0.5, 4) + f(1.0, 6.75)}{2} = 7.125, \\ u_2 &= u_1 + hm = 4 + (0.5)(7.125) = 7.5625. \end{aligned}$$

Standard Euler's method gives the estimate $u(1.0) \approx 6.0$, so the improved Euler's method is substantially better. ■

Reading Exercise 3.2.1 Verify that two steps of the standard Euler's method with step size $h = 0.5$ yields the estimate $u(1.0) \approx 6.0$.

Reading Exercise 3.2.2 Continue the computations of Example 3.5 to estimate $u(2.0)$. Compare the estimate to the true solution value.

Accuracy Of the Improved Euler Method

The improved Euler's method requires more work for each time step, but the payoff is much better accuracy. The improved Euler's method is also sometimes known as the **modified Euler's method** or **Heun's method**, although the latter term is sometimes used to refer to another closely related method. The improved Euler's method is also a simple example of a class of methods for numerically solving ODEs known as **Runge-Kutta methods**. Runge-Kutta methods will be further considered in the next section, as they are workhorses of modern numerical ODE solvers.

As with Euler's method, more accurate solution estimates can typically be obtained by using smaller step sizes. Let's consider an example that shows just how superior the improved Euler's method is compared to the standard Euler's method.

■ **Example 3.6** Consider the ODE $u'(t) = u(t)$ and initial condition $u(0) = 1$. The analytical solution is $u(t) = e^t$. We will use both Euler's method and the improved Euler's method to estimate $u(1) = e$, using step sizes $h = 1, 0.1, 0.01, 0.001$, and $h = 0.0001$ for each method. This requires $N = 1, 10, 100, 1000$, and $N = 10000$ steps, respectively. The results are tabulated in Table 3.3.

Step size h	Euler estimate	Euler Error	Improved Euler	Improved Euler Error
1.00	2.000000	7.183×10^{-1}	2.500000	2.183×10^{-1}
10^{-1}	2.593742	1.245×10^{-1}	2.714081	4.201×10^{-3}
10^{-2}	2.704813	1.347×10^{-2}	2.718237	4.497×10^{-5}
10^{-3}	2.716924	1.358×10^{-3}	2.718281	4.522×10^{-7}
10^{-4}	2.718146	1.359×10^{-4}	2.718281	4.530×10^{-9}

Table 3.3: Euler's method and improved Euler's method estimates of $u(1) = e$ for the ODE $u'(t) = u(t)$ with various step sizes h , with error $|e - u_h(1.0)|$.

An examination of the error for the improved Euler's method, especially for $h \leq 0.1$, shows that each 10-fold decrease in h results in approximately a 100-fold decrease in the error, so the error is roughly proportional to h^2 . This is typical of the improved Euler's method; for most ODEs the error made by the improved Euler's method is proportional to h^2 , i.e.,

$$|u(T) - u_h(T)| \approx Ch^2$$

for some constant C , which is a rather dramatic improvement over standard Euler's method. Since the error is approximately proportional to the second power of h we say that the improved Euler's method is **second-order accurate**. ■

3.2.3 Exercises

Exercise 3.2.1 Apply the improved Euler's method to each initial value problem, by hand (but use a calculator), using the indicated step size h and number of steps N . Carry computations to four significant figures. Then compute the value of the true (analytical) solution at time $T = t_0 + Nh$ and compare.

- (a) $u'(t) = u(t) + 3$, $u(0) = 1$, step size $h = 0.5$, $N = 2$ steps.
- (b) $u'(t) = -u(t) + 3t$, $u(0) = 2$, step size $h = 0.25$, $N = 4$ steps.
- (c) $u'(t) = 1/u(t)$, $u(0) = 2$, step size $h = 0.25$, $N = 4$ steps.
- (d) $u'(t) = tu(t)$, $u(1) = 3$, step size $h = 0.1$, $N = 5$ steps.

Exercise 3.2.2 For each initial value problem, apply the improved Euler's method with the indicated step sizes h , using whatever technology you have available, to estimate $u(T)$ for the given value of T . Compare these estimates to the true value of $u(T)$ obtained from an analytical solution.

- (a) $u'(t) = 1 - u(t)/3$, $u(0) = 2$. Estimate $u(5)$ using step sizes $h = 1, 0.1, 0.01$.
- (b) $u'(t) = te^{-u(t)}$, $u(0) = 1$. Estimate $u(3)$ using step sizes $h = 1, 0.1, 0.01$.
- (c) $u'(t) = u^2(t)$, $u(0) = 2$. Estimate $u(0.5)$ using step sizes $h = 0.5, 0.1, 0.01, 0.001$.
- (d) $u'(t) = 1/u(t)$, $u(0) = 2$. Estimate $u(4)$ using step sizes $h = 1.0, 0.1, 0.01$.

Exercise 3.2.3 Apply the improved Euler's method to the ODE $u'(t) = u(t) - t + 1$ with $u(0) = 0$, to estimate $u(1)$ using step sizes $h = 1, 0.1$, and 0.01 . Then find the analytical solution and compute the error for each step size. Why does this make perfect sense?

Exercise 3.2.4 (Compare the results here to Exercise 3.1.5.) Apply the improved Euler's method to the ODE $u'(t) = u^2(t)$ with $u(0) = 1$ to estimate $u(2)$ using step sizes $h = 1, 0.1, 0.01$, and 0.001 . Explain what's going on. Hint: compute the analytical solution using separation of variables. Then recall Definition 2.4.1 and the notion of the maximum domain of a solution from Section 2.4.2.

Exercise 3.2.5 (Compare the results here to Exercise 3.1.6.) Consider the linear ODE

$$u'(t) = u(t) - \sin(t) + \cos(t).$$

- (a) Find a general solution.
- (b) Find the solution with initial condition $u(0) = 0$.
- (c) Sketch a direction field on the range $0 \leq t \leq 10$, $-5 \leq u \leq 5$, and superimpose the solution with $u(0) = 0$ on this direction field.
- (d) Apply the improved Euler's method with step sizes $h = 1, 0.1, 0.01$, and 0.001 with initial condition $u(0) = 0$ to estimate $u(10)$. Explain the poor estimates for $u(10)$ in light of the direction field from (c) and general solution from (a). Hint: what happens if the improved Euler's method ever strays from the analytical solution curve? It might be helpful to plot the improved Euler's method iterates for $h = 0.001$.

Exercise 3.2.6 (Compare the results here to Exercise 3.1.7). This problem illustrates that if the step size is too large, the improved Euler's method (like Euler's method) isn't just inaccurate—it may actually blow up, even if the true solution to the ODE decays.

Consider the differential equation $u'(t) = -10u(t)$ with $u(0) = 1$.

- (a) Find the analytic solution to this initial value problem, and show that it decays to zero as $t \rightarrow \infty$.
- (b) Apply the improved Euler's method with step size $h = 0.1$ to estimate $u(5)$.
- (c) Apply the improved Euler's method with step size $h = 0.2$ to estimate $u(5)$. Hint: after the first step the computations should be trivial.
- (d) Apply the improved Euler's method with step size $h = 1$ to estimate $u(5)$.
- (e) Suppose we apply the improved Euler's method with step size h as detailed in (3.18). Show that u_{k+1} (an approximation to $u((k+1)h)$) is given by

$$u_{k+1} = (50h^2 - 10h + 1)u_k,$$

so that with initial iterate $u_0 = 1$ we have

$$u_k = (50h^2 - 10h + 1)^k. \quad (3.21)$$

The use of (3.21) should yield results in accordance with parts (b)-(d).

- (f) From part (a) we know that the analytical solution decays to zero. How large can we take step size h in (3.21) to (at least) obtain decay to zero? Interpret the results of parts (b)-(d) in light of this analysis.

3.3 Modern Numerical Methods

Numerical methods for solving ODEs form a vast field of research that is central to much modeling and analysis in engineering, science, and mathematics. There are many algorithms that are more sophisticated than what we've seen so far. Some algorithms are special-purpose, for certain types of ODEs, but many general purpose algorithms also exist. These algorithms are designed to work for non-experts on most problems. Modern software typically allows the user to select from a variety of algorithms, and also set a number of parameters that influence the algorithm's behavior, such as how accurately and efficiently solutions are approximated. It's thus helpful to know a little bit about why one might select one algorithm over another, and about the various parameters a user can set to control the algorithm's behavior. The goal in the section is not to make you an expert in ODE solvers, nor to teach you to program your own, but rather to help you become a knowledgeable user of modern ODE software.

3.3.1 The RK4 Algorithm

Runge-Kutta algorithms are a class of numerical ODE solvers that are commonly used as part of a general-purpose ODE solver. In particular, the classic **fourth-order Runge-Kutta algorithm** (often abbreviated RK4) forms the basis of many computer codes for solving ODEs numerically. Like the improved Euler's method, RK4 steps from an estimate of the solution value u_k at time t_k to an estimate u_{k+1} at time t_{k+1} using intermediate computations.

Suppose $u(t)$ is the solution to $u' = f(t, u)$ with initial condition $u(t_0) = u_0$. As with previous methods a step size h is chosen and we define $t_k = t_0 + kh$. The RK4 method steps from u_k to u_{k+1} using the following formulas:

$$\begin{aligned} m_1 &= f(t_k, u_k), \\ m_2 &= f(t_k + h/2, u_k + hm_1/2), \\ m_3 &= f(t_k + h/2, u_k + hm_2/2), \\ m_4 &= f(t_k + h, u_k + hm_3), \\ m &= \frac{1}{6}(m_1 + 2m_2 + 2m_3 + m_4), \\ u_{k+1} &= u_k + hm. \end{aligned} \tag{3.22}$$

These formulas might be seen as a generalization of the philosophy behind the improved Euler's method (itself a member of the Runge-Kutta family), that is, they may be viewed as a method to track the direction field more accurately from $t = t_k$ to $t = t_{k+1}$. We will not motivate or derive the formulas in (3.22), but will illustrate their effectiveness in solving ODEs below. For a derivation of (3.22) see [76].

■ **Example 3.7** Let's apply the RK4 formulas (3.22) to the initial value problem $u'(t) = u(t) + t$ with initial condition $u(0) = 1$ (analytical solution $u(t) = 2e^t - t - 1$) to estimate $u(1)$. We use step size $h = 1$, the largest step size possible here, since it takes us to the final time $t = 1$ in a single step. With $f(t, u) = u + t$, $t_0 = 0$, $t_1 = 1$, and $u_0 = 1$ the RK4, method yields

$$\begin{aligned} m_1 &= f(t_0, u_0) = f(0, 1) = 1, \\ m_2 &= f(t_0 + h/2, u_0 + hm_1/2) = f(0.5, 1.5) = 2, \\ m_3 &= f(t_0 + h/2, u_0 + hm_2/2) = f(0.5, 2) = 2.5, \\ m_4 &= f(t_1, u_0 + hm_3) = f(1, 3.5) = 4.5, \\ m &= \frac{1}{6}(m_1 + 2m_2 + 2m_3 + m_4) = 29/12 \approx 2.41667, \\ u_{k+1} &= u_0 + hm = 1 + (1)(2.41667) \approx 3.41667. \end{aligned}$$

The true value is $u(1) = 2e - 2 \approx 3.43656$, so the error here is only about 0.02, quite close, and that's with a step size of 1, the largest that can be taken. For comparison, Euler's method yields estimate $u(1) \approx 2$, error about 1.437, while the improved Euler's method yields $u(1) \approx 3$, error about 0.437. ■

■ **Example 3.8** Let's look at how RK4 behaves as the step size is decreased. As in Examples 3.2 and 3.6, consider the ODE $u'(t) = u(t)$ and initial condition $u(0) = 1$, with true solution $u(t) = e^t$. Let us estimate $u(1) = e$ using step sizes $h = 1, h = 0.1, h = 0.01$, and $h = 0.001$. This requires $N = 1, 10, 100$, and 1000 steps, respectively. The results are tabulated in Table 3.4, where $u_h(1.0)$ is the estimate of $u(1) = e$ using the RK4 method with step size h , and the error is computed as $|e - u_h(1.0)|$. It appears, at least asymptotically, that each 10-fold decrease in h results in approximately a 10^4 -fold decrease in the error, so the error is roughly proportional to h^4 . This is typical; for most ODEs the error in the RK4 method is proportional to h^4 , and obeys the inequality

$$|u(T) - u_h(T)| \approx Ch^4$$

Step size h	RK4 estimate $u_h(1.0)$	Error
1.0	2.708333	9.950×10^{-3}
0.1	2.718279	2.084×10^{-6}
0.01	2.718282	2.246×10^{-10}
0.001	2.718282	2.263×10^{-14}

Table 3.4: RK4 estimate of $u(1) = e$ for various step sizes h , with error $|e - u_h(1.0)|$.

for some constant C . We say that the RK4 method is **fourth-order accurate**. This is a huge improvement even over the improved Euler's method. ■

Numerical ODE solvers can be designed that satisfy an error bound of the form Ch^n for any n . However, larger n require more complicated solution formulas and many more computations for each iteration of the algorithm. Also, this asymptotic behavior only holds if the true solution $u(t)$ is sufficiently differentiable, for the constant C usually depends on $u^{(n+1)}$, the order $n + 1$ derivative of u . The RK4 method is considered a reasonable balance between accuracy and complexity.

Reading Exercise 3.3.1 Consider the ODE $u'(t) = -2u(t) + t^2$ with initial condition $u(0) = 0$. Use the RK4 method to estimate $u(1)$, by performing a single step of size $h = 1$. Compare the result to the true value for $u(1)$.

3.3.2 Adaptive Step Sizing and Error Control

Motivation

Let's return to the logistic equation in the form (3.1), in the case that $r = 20$ (a constant) and carrying capacity $K(t) = 10 + 0.1 \sin(2\pi t)$. The ODE becomes

$$u'(t) = 20u(t) \left(1 - \frac{u(t)}{10 + 0.1 \sin(2\pi t)} \right). \quad (3.23)$$

We'll use initial condition $u(0) = 0.5$. The ODE (3.23) has no analytical solution and so must be solved numerically. We do this with an RK4 method and a time step of 10^{-4} . The result is shown in Figure 3.7 and may be considered the true solution, at least to visual accuracy.

The step size used in numerically solving (3.23) determines how closely the method follows the direction field, and therefore the accuracy of the computed solution. The solution as shown in Figure 3.7 rises very rapidly from $t = 0$ to about $t = 0.4$, and then quickly levels off and varies rather slowly for $t \geq 0.5$. In the region $0 \leq t \leq 0.5$, where the solution changes rapidly, a small step size is required to maintain accuracy, especially in the vicinity of $0.4 < t < 0.5$ where the solution has a large second derivative. For $t \geq 0.5$, where the solution varies slowly, a much larger step size can be used without unduly sacrificing accuracy. This is illustrated in Figure 3.8, in which an RK4 method with a much larger step size of $h = 0.2$ is used (solid black curve), along with the more accurately computed solution (shown in red, dashed). The fixed step size procedure is obviously inaccurate in the range $0.4 < t < 1.5$.

This example highlights that we often have conflicting needs when solving ODEs numerically. In regions where the solution changes behavior rapidly, small steps are needed for accurate tracking. However, in regions where the solution changes slowly, much larger steps can be taken while maintaining good accuracy. The sledgehammer approach of taking small steps everywhere is wasteful and slow. What is needed is a method to monitor how accurately the method is tracking the true solution and then adapting the step size accordingly. Such an algorithm is called an **adaptive step size method**.

But how are we supposed to estimate the accuracy with which the true solution is being tracked when we don't know what the true solution is?

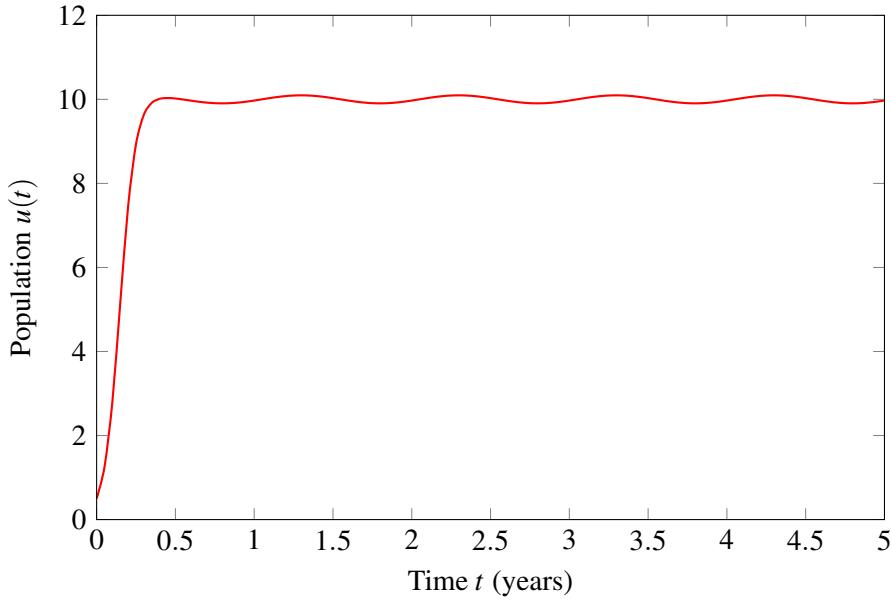


Figure 3.7: Solution to (3.23) with $r = 20$, $K(t) = 10 + 0.1 \sin(2\pi t)$, and initial condition $u(0) = 0.5$.

The Idea Behind Adaptive Step Sizing

Let's look at how one might implement an adaptive step size for the simplest numerical solver, Euler's method. Bear in mind that Euler would not be the method of choice. Moreover, the strategy below is not as efficient as it could be, but it will serve to get across the main point: one can estimate how much error is being made at each iteration and then increase or decrease the step size accordingly.

The situation is illustrated in Figure 3.9. Suppose we're marching out Euler's method in time and are currently at time $t = t_k$. Suppose also that the solution estimate $u_k = u(t_k)$ is exact. An Euler step is to be taken to estimate $u(t_k + h)$ by using (3.11), $u_{k+1} = u_k + hf(t_k, u_k)$. The estimate u_{k+1} for $u(t_k + h)$ will likely be a bit off of the true value. The goal is to estimate $u(t_k + h) - u_{k+1}$, the error made with an Euler step of size h , and then decrease h if this error is too large, or perhaps increase h if the error is very small, in order to gain efficiency by taking larger steps.

Error Analysis and Estimation

To estimate the error $u(t_k + h) - u_{k+1}$, perform a second-order Taylor expansion of $u(t)$ at the point $t = t_k$ in the form

$$u(t_k + h) = u(t_k) + hu'(t_k) + \frac{1}{2}u''(s)h^2. \quad (3.24)$$

Here s lies between t_k and $t_k + h$; since $h > 0$ it follows that $t_k < s < t_k + h$. Recall the assumption that $u_k = u(t_k)$ and note that $u' = f(t, u)$, so $u'(t_k) = f(t_k, u_k)$ and (3.24) becomes

$$u(t_k + h) = \underbrace{u_k + hf(t_k, u_k)}_{\text{Euler step } u_{k+1}} + \underbrace{\frac{1}{2}u''(s)h^2}_{\text{local truncation error}}. \quad (3.25)$$

As indicated, the quantity $u_k + hf(t_k, u_k)$ is precisely u_{k+1} , the estimate of $u(t_k + h)$ produced by Euler's method. The term $\frac{1}{2}u''(s)h^2$ is called the **local truncation error** (LTE) and is the error introduced by Euler's method in the step from $t = t_k$ to $t = t_k + h$. This error stems from the fact that the solution is extrapolated forward in time using the tangent line approximation, but $u(t)$ itself is (probably) not linear on this time interval.

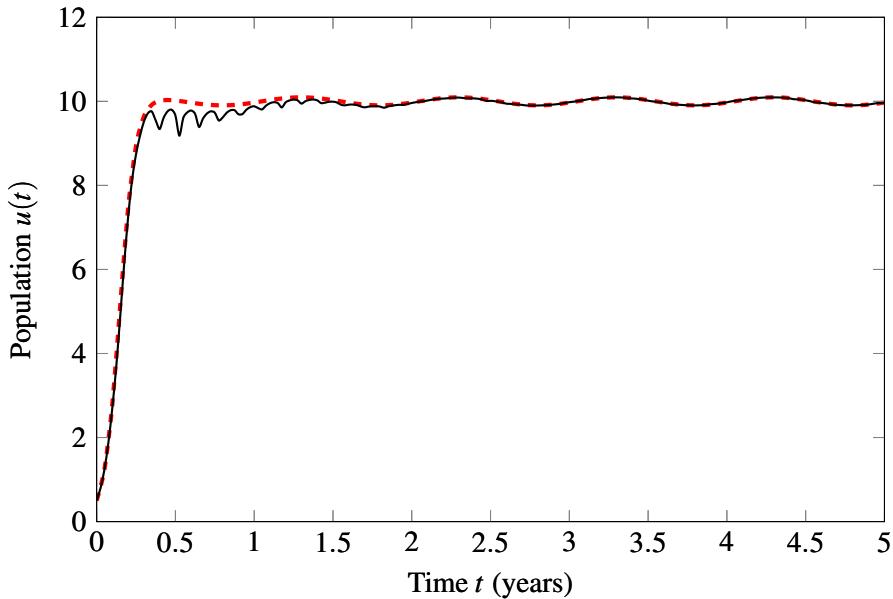


Figure 3.8: Solution to (3.23) with $r = 20$, $K(t) = 10 + 0.1 \sin(2\pi t)$, and initial condition $u(0) = 0.5$. The accurate solution is the red dashed curve, the RK4 method solution with fixed step size $h = 0.2$ is shown as the solid black curve.

If h is close to zero then $s \approx t_k$, because $t_k < s < t_k + h$. As a result, if u'' is continuous, the local truncation error will be close to $\frac{1}{2}u''(t_k)h^2$. For notational simplicity define $C = \frac{1}{2}u''(t_k)$, so the local truncation error is approximately Ch^2 . Equation (3.25) shows that if we start with $u_k = u(t_k)$ (no error at step k) and take a step of size h to estimate $u(t_{k+1})$ with Euler's method, then the local truncation error $u(t_k + h) - u_{k+1}$ is approximately

$$u(t_k + h) - u_{k+1} \approx Ch^2, \quad (3.26)$$

where $C = \frac{1}{2}u''(t_k)$.

Reading Exercise 3.3.2 Let $u(t) = 2e^t - t - 1$, which is a solution to $u'(t) = t + u(t)$ (so $f(t, u) = t + u$, the initial condition is irrelevant). Suppose we are marching Euler's method out in time and currently have $t_k = 1.5$ with $u_k = u(t_k) \approx 6.463378140$. For each step size $h = 1, 0.1, 0.01$, and $h = 0.001$ compute the Euler estimate $u_{k+1} = u_k + hf(t_k, u_k)$ as well as the actual value for $u(t_k + h)$, then make a table showing h versus $u(t_k + h) - u_{k+1}$. Does (3.26) seem to hold? Based on your table, what is the appropriate value for C here? Is it close to $\frac{1}{2}u''(t_k)$?

We can use (3.26) to estimate the error made when taking an Euler step of size h , which allows this step size to be adapted to control the error. If the value of C were known this would be easy, since the right side of (3.26) gives the approximate truncation error explicitly, but C is not known. However, the following approach allows both C and the truncation error to be estimated simultaneously, with a bit of extra computation.

Assume the current operating point is $t = t_k$ with estimate $u(t_k) \approx u_k$, and that h is the current default step size.

1. Take an Euler step $u_{k+1} = u_k + hf(t_k, u_k)$ of size h . In (3.26) both h and u_{k+1} are known, but neither C nor $u(t_k + h)$ are known.

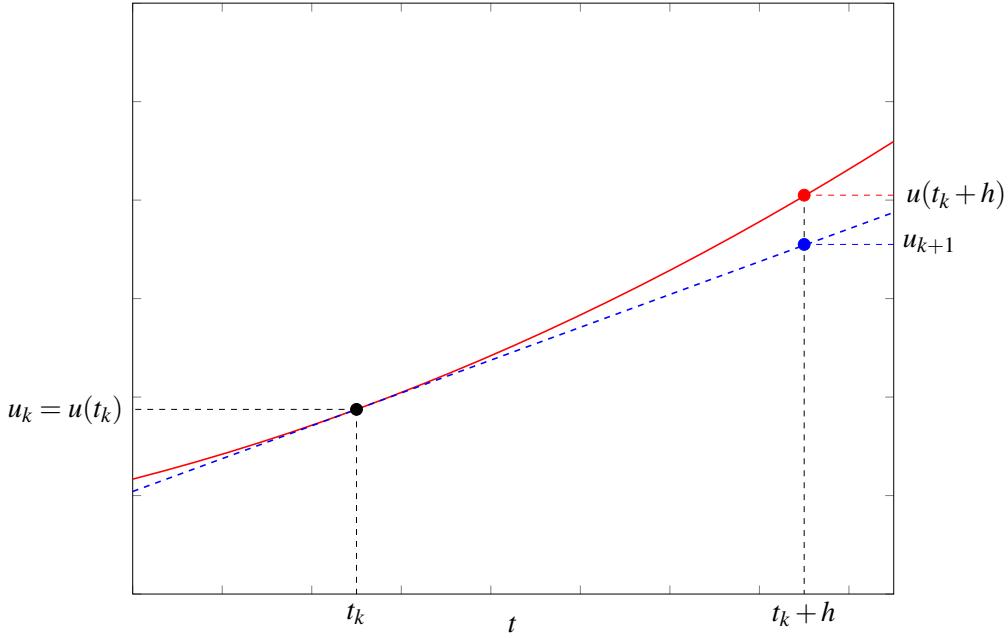


Figure 3.9: Graph of $y = u(t)$ (red, solid) and the linearization $y = L(t)$ (blue, dashed) of u at $t = t_k$, with points (t_k, u_k) (black dot), $(t_k + h, u(t_k + h))$ (red dot) and Euler step $(t_k + h, u_{k+1})$ (blue dot).

2. Starting at $t = t_k$ again, take two Euler steps, each of size $h/2$, as

$$u_{k+1/2} = u_k + \frac{h}{2} f(t_k, u_k), \quad (3.27)$$

$$\tilde{u}_{k+1} = u_{k+1/2} + \frac{h}{2} f(t_k + h/2, u_{k+1/2}), \quad (3.28)$$

to again step the solution to $t = t_k + h$. The quantity \tilde{u}_{k+1} is an estimate of $u(t_k + h)$, but presumably more accurate than u_{k+1} , since two steps of size $h/2$ ought to track the solution better than a single step of size h .

3. A bit of analysis yields the plausible conclusion that each of step (3.27) and step (3.28) introduces approximate truncation error $C(h/2)^2$, for a total of $2C(h/2)^2 = Ch^2/2$. Then we have

$$u(t_k + h) - \tilde{u}_{k+1} \approx Ch^2/2. \quad (3.29)$$

4. Equations (3.26) and (3.29) can be considered two (approximate) equations in unknowns C and $u(t_k + h)$. In particular, subtract (3.29) from (3.26) so the $u(t_k + h)$ terms cancel and a bit of algebra yields

$$|\text{LTE}| = |Ch^2| \approx 2|\tilde{u}_{k+1} - u_{k+1}|, \quad (3.30)$$

which is an estimate of $|\text{LTE}|$, the magnitude of the local truncation error.

The punchline is (3.30): by taking a single Euler step of size h to produce the estimate u_{k+1} and then repeating with two steps of size $h/2$ to produce the estimate \tilde{u}_{k+1} , (3.30) can be used to estimate the error Ch^2 made when taking a step size of h . We can then decide whether the magnitude of this error is acceptable. This can also be used to improve the estimate of $u(t_k + h)$.

■ **Example 3.9** Let $u(t)$ be a solution to $u'(t) = t + u(t)$, so $f(t, u) = t + u$. Suppose the solution is being marched out in time using Euler's method with current operating point $t_k = 1.5$ with $u_k = 5.95$. Let us estimate the truncation error that will be made with a step size of $h = 0.1$.

First, for a single step of size $h = 0.1$ the Euler estimate for $u(1.6)$ is

$$u_{k+1} = u_k + hf(t_k, u_k) = 5.95 + 0.1f(1.5, 5.95) = 6.695.$$

Alternatively, take two steps of size $h/2 = 0.05$, as

$$\begin{aligned} u_{k+1/2} &= u_k + \frac{h}{2}f(t_k, u_k) = 5.95 + 0.05f(1.5, 5.95) = 6.3225, \\ \tilde{u}_{k+1} &= u_{k+1/2} + \frac{h}{2}f(t_k + h/2, u_{k+1/2}) = 6.716125. \end{aligned} \quad (3.31)$$

From (3.30) the truncation error with step size h is approximately

$$|\text{LTE}| = |Ch^2| \approx 0.04225.$$

This error would be on top of any errors up to time $t = t_k$, and it may be compounded in later iterations. If the error meets whatever tolerance criterion that has been set, the step of size h is accepted, otherwise h decreased by some amount and the computation repeated. ■

Reading Exercise 3.3.3 Suppose the ODE $u'(t) = u^2(t) - t$ is being solved with Euler's method, with current iterate $t_k = 1$, $u_k = 0.7$, and step size $h = 0.2$. Estimate the magnitude of the truncation error that comes from taking a step of size $h = 0.2$.

Implementing Adaptive Step Sizing and Error Control

Let's consider an example to see how (3.30) might be used to inform a simple adaptive step sizing scheme, but with some caveats outlined below.

Suppose the ODE of interest is

$$u'(t) = \frac{1}{2}u(t) + t,$$

with $u(0) = 1$, to be solved out to time $T = 1$ using Euler's method. Suppose also that we want the final estimate of $u(1)$ produced by the solver to be accurate to within 10 percent of the correct value, a pretty big error tolerance, but this is for illustrative purposes. We might hope that by estimating and controlling the error in each iteration of Euler's method as in Example 3.9 this could be achieved. Unfortunately there is no general way to guarantee that the final estimate for $u(1)$ is accurate to within a specified tolerance by simply controlling the error at each step leading up to $t = 1$. The difficulty is that errors made in earlier steps can be magnified in later steps in a way that is difficult to predict or quantify. What can be done in practice is to impose a condition that the estimated truncation error in each step is less than some fraction of the current solution value, perhaps one percent, and hope this translates into something comparable in the estimate for $u(1)$.

Let's try this strategy. We will require that the local truncation error (3.30) at the k th iteration of Euler's method satisfies

$$|\text{LTE}| \leq \tau |u_k| \quad (3.32)$$

for some chosen tolerance τ , which here will be taken as $\tau = 0.01$.

Begin with an initial step size of $h = 0.25$ (an arbitrary choice). Set $u_0 = 1$ and $t_0 = 0$. A single Euler step produces the estimate

$$u_1 = u_0 + hf(t_0, u_0) = 1.125.$$

Two Euler steps of size $h/2 = 0.125$ produce the estimate \tilde{u}_1 :

$$\begin{aligned} u_{1/2} &= u_0 + \frac{h}{2}f(t_0, u_0) = 1.0625, \\ \tilde{u}_1 &= u_{1/2} + \frac{h}{2}f(t_0 + h/2, u_{1/2}) = 1.1445. \end{aligned}$$

Use (3.30) to estimate the local truncation error $\text{LTE} \approx Ch^2$ for the step of size $h = 0.25$ as

$$|\text{LTE}| = 2|\tilde{u}_1 - u_1| \approx 0.039.$$

The current iterate is $u_0 = 1$ and since $|\text{LTE}| \approx 0.039$, the inequality (3.32) is not satisfied with the given tolerance τ . The step of size $h = 0.25$ is therefore rejected.

The value of h must then be decreased. Let's cut h in half and try again. Repeat the step from time $t_0 = 0$ to $t_1 = h$ with $h = 0.125$ to obtain

$$u_1 = u_0 + hf(t_0, u_0) = 1.0625.$$

Two Euler's steps of size $h/2 = 0.00625$ produce estimate \tilde{u}_1 :

$$\begin{aligned} u_{1/2} &= u_0 + \frac{h}{2}f(t_0, u_0) = 1.03125, \\ \tilde{u}_1 &= u_{1/2} + \frac{h}{2}f(t_0 + h/2, u_{1/2}) = 1.0674. \end{aligned}$$

We estimate the local truncation error LTE for the step of size $h = 0.125$ by using (3.30) to find

$$|\text{LTE}| = 2|\tilde{u}_1 - u_1| = 0.00977.$$

In this case (3.32) is satisfied. The iterate $u_1 = 1.0674$ is accepted and $t_1 = 0.125$.

The next step of Euler's method is to extrapolate to $t_2 = t_1 + h$. We begin with the current value of $h = 0.125$ and continue this process. At each iteration the initial value of h used is the final adopted value from the previous iteration. The value of h might be increased (perhaps by a factor of 2) at some iteration if the estimate of the local truncation error is much less than the budgeted amount. This is an example of a numerical ODE solver that uses an **adaptive step size**: The estimated error is monitored at each iteration and the step size is adjusted to provide **error control**.

As noted earlier however, despite use of the imposed tolerance (3.32), there is no firm guarantee that the final answer is within any tolerance of the correct answer, since the errors are not simply additive. An error made at any stage may be amplified in later stages.

Practical ODE Solvers

Real adaptive stepsizing strategies employ many improvements over this simple scheme. Instead of taking steps of size h and then $h/2$ to estimate and control local truncation error, most combine two different ODE solvers to monitor and control error. One of the more popular approaches to adaptive stepsizing is the Runge-Kutta-Fehlberg 4/5 (RKF45) method, which uses the RK4 method discussed earlier and pairs it with another method that is fifth-order accurate. Together the estimates from these methods can be used in a manner similar to our Euler scheme above to estimate local truncation error. One of the goals is to make the scheme as efficient as possible by evaluating the function $f(t, u)$ as few times as possible. The RKF45 method is designed to do this.

As noted above, general purpose numerical ODE solvers don't typically try to control the error in the estimate of $u(T)$ at the final time $t = T$ (called the **global error**), but rather the error made at each step in the iteration, in a fashion similar to what we've done. Matlab's `ode45` numerical ODE solver is a typical general purpose modern solver for problems of the form $u' = f(t, u)$. This solver accepts an argument “`RelTol`” that controls the step size by dictating the maximum allowed local truncation error relative to the current value u_k by enforcing a bound similar to (3.32). Additionally, `ode45` accepts an argument “`AbsTol`” that enforces a bound $|\text{LTE}| \leq \varepsilon$ for some tolerance ε . If at any iteration the tolerance criteria are not met, the step size can be adjusted according to some strategy. Other software packages have similar options. See Exercise 3.3.3.

For much more information on the numerical solution of ODEs, see [76].

3.3.3 Exercises

Exercise 3.3.1 Apply the RK4 method (3.22) to each initial value problem, by hand (but use a calculator), using the indicated step size h and number of steps N . Carry computations to at least four significant figures. Then compute the value of the true (analytical) solution at time $T = t_0 + Nh$ and compare.

- (a) $u'(t) = u(t) + 3$, $u(0) = 1$, step size $h = 0.5$, $N = 2$ steps.
- (b) $u'(t) = -u(t) + 3t$, $u(0) = 2$, step size $h = 0.5$, $N = 2$ steps.
- (c) $u'(t) = 1/u(t)$, $u(0) = 2$, step size $h = 0.5$, $N = 2$ steps.
- (d) $u'(t) = tu(t)$, $u(1) = 3$, step size $h = 0.25$, $N = 2$ steps.

Exercise 3.3.2 Apply the RK4 method using whatever technology you have available to the given initial value problems with the indicated step sizes h to estimate $u(T)$ for the given value of T . Compare these estimates to the true value of $u(T)$ obtained from an analytical solution.

- (a) $u'(t) = 1 - u(t)/3$, $u(0) = 2$. Estimate $u(5)$ using step sizes $h = 1, 0.1, 0.01$.
- (b) $u'(t) = te^{-u(t)}$, $u(0) = 1$. Estimate $u(3)$ using step sizes $h = 1, 0.1, 0.01$.
- (c) $u'(t) = u^2(t)$, $u(0) = 2$. Estimate $u(0.5)$ using step sizes $h = 0.5, 0.1, 0.01, 0.001$.
- (d) $u'(t) = 1/u(t)$, $u(0) = 2$. Estimate $u(4)$ using step sizes $h = 1.0, 0.1, 0.01$.

Exercise 3.3.3 Consider the logistic ODE (3.23) with initial condition $u(0) = 0.5$. To ten significant figures, the solution at $t = 1$ is $u(1) = 9.971345698$.

- (a) Estimate $u(1)$ by using Euler's method with steps of size $h = 0.1$, and compute the error.
- (b) Estimate $u(1)$ by using the improved Euler's method with steps of size $h = 0.1$, and compute the error.
- (c) Estimate $u(1)$ by using the RK4 method with steps of size $h = 0.1$, and compute the error.
- (d) Estimate $u(1)$ by using whatever numerical ODE solver you have available, without specifying the method, then compute the error. Most software, e.g., Mathematica, Maple, Matlab, and Sage, will use a good ODE solver with adaptive stepsizing for error control.
- (e) Explore the error tolerances in the solver you are using. For example, the Maple `dsolve` command, when solving numerically using the usual RKF45 method, accepts arguments `abserr` (default value 1.0×10^{-7}) and `relerr` (default value 1.0×10^{-6}) that can be used to control error. In Matlab the `ode45` command accepts arguments `AbsTol` (default value 1.0×10^{-6}) and `RelTol` (default value 1.0×10^{-3}). In Mathematica the `NDSolve` command has arguments `AccuracyGoal` and `PrecisionGoal`.

Exercise 3.3.4 (Compare the results here to Exercises 3.1.5 and 3.2.4.) Apply the RK4 method to the ODE $u'(t) = u^2(t)$ with $u(0) = 1$, to estimate $u(2)$ using step sizes $h = 1, 0.1, 0.01$, and 0.001 . Explain what's going on. Hint: compute the analytical solution using separation of variables. Then recall Definition 2.4.1 and the notion of the maximum domain of a solution from Section 2.4.2.

Exercise 3.3.5 (Compare the results here to Exercises 3.1.6 and 3.2.5.) Consider the linear

ODE

$$u'(t) = u(t) - \sin(t) + \cos(t).$$

- (a) Find a general solution.
- (b) Find the solution with initial condition $u(0) = 0$.
- (c) Sketch a direction field on the range $0 \leq t \leq 10, -5 \leq u \leq 5$, and superimpose the solution with $u(0) = 0$ on top of it.
- (d) Apply the RK4 method with step sizes $h = 1, 0.1, 0.01, 0.001$ with initial condition $u(0) = 0$ to estimate $u(10)$. Explain the poor estimates for $u(10)$ in light of the direction field from (c) and general solution from (a). Hint: what happens if the RK4 method ever strays off of the analytical solution curve? It might be helpful to plot the RK4 iterates for $h = 0.001$.

Exercise 3.3.6 (Compare to Exercises 3.1.7 and 3.2.6.) This problem illustrates that if the step size is too large, the RK4 method (like Euler's method) isn't just inaccurate—it may actually blow up, even if the true solution to the ODE decays.

Consider the differential equation $u'(t) = -10u(t)$ with $u(0) = 1$.

- (a) Find the analytic solution to this initial value problem, and show that it decays to zero as $t \rightarrow \infty$.
- (b) Apply the RK4 method with step size $h = 0.1$ to estimate $u(5)$.
- (c) Apply the RK4 method with step size $h = 0.2$ to estimate $u(5)$.
- (d) Apply the RK4 method with step size $h = 1$ to estimate $u(5)$.
- (e) Suppose we apply the RK4 method with step size h . Experiment to find how large h can be before the estimate solution $u(t)$ no longer decays to zero.

Exercise 3.3.7 Suppose we are numerically solving $u' = tu^2(t) + \sin(t)$ with current point $t_k = 0.5, u_k = 1.0$, and step size $h = 0.1$ using Euler's method. We wish to take a step that introduces a local truncation error less than 10^{-2} .

- (a) Use (3.11) to take a single Euler step, to produce an estimate u_{k+1} of $u(t_k + h)$.
- (b) Take two Euler steps of size $h/2$ to produce \tilde{u}_{k+1} , an estimate of $u(t_k + h)$.
- (c) Use (3.30) to estimate the local truncation error. Is it within the accepted tolerance? If not, would $h/2$ work?

3.4 Parameter Estimation

3.4.1 Hill-Keller Revisited

Recall the Hill-Keller model developed in Chapter 1 to describe the motion of a sprinter along a track. The motivation was to model Usain Bolt's performance in the 2008 Olympics, as detailed by the data in Table 1.1. The Hill-Keller ODE is

$$v'(t) = P - kv(t), \quad (3.33)$$

with initial condition $v(t_0) = 0$. Based on the data in Table 1.1, $t_0 = 0.165$ for Bolt's 2008 Olympic race. The choice $P = 11$ meters per second squared in (3.33) was made based on estimates from physiological lab data for world-class sprinters. However, the constant k (dimension: reciprocal time) is unknown; how can a reasonable value for k be determined? Estimating k would allow us to

find Bolt's velocity and position, and make a quantitative comparison of the model's prediction to Bolt's data.

Let's start by solving (3.33) with $P = 11$ and $v(0.165) = 0$, leaving k as an unspecified constant. The solution can be obtained by using separation of variables or the integrating factor technique and is

$$v(t) = \frac{11}{k} \left(1 - e^{-k(t-t_0)} \right), \quad (3.34)$$

where $t_0 = 0.165$. The data in Table 1.1 is not velocity, however, but rather Bolt's time as he passed the 10, 20, ..., 100 meter marks. In order to compare the model to the data we'll use $v(t)$ from (3.34) to compute Bolt's position $x(t)$ as a function of time. His position $x(t)$ satisfies $x'(t) = v(t)$ with $x(t_0) = 0$, for he starts at position $x = 0$, the starting line, at time t_0 . The function $x(t)$ can be computed, using z as the variable of integration, as

$$\begin{aligned} x(t) &= \int_{t_0}^t v(z) dz \\ &= \int_{t_0}^t \frac{11}{k} \left(1 - e^{-k(z-t_0)} \right) dz \\ &= \frac{11}{k} \left(z + \frac{e^{-k(z-t_0)}}{k} \right) \Big|_{z=t_0}^{z=t} \\ &= \frac{11}{k^2} \left(e^{-k(t-t_0)} - 1 + k(t-t_0) \right). \end{aligned} \quad (3.35)$$

The function $x(t)$ in (3.35) allows us to estimate k . One crude method would be to guess a value for k , graph $x(t)$, compare this graph to a plot of the data, and then adjust k to obtain a good visual fit. To illustrate, consider the choice $k = 1$ (units: reciprocal seconds) in (3.35). The graph of the resulting function $x(t)$ is shown in Figure 3.10 as the solid red curve, overlayed on the data from Table 1.1.

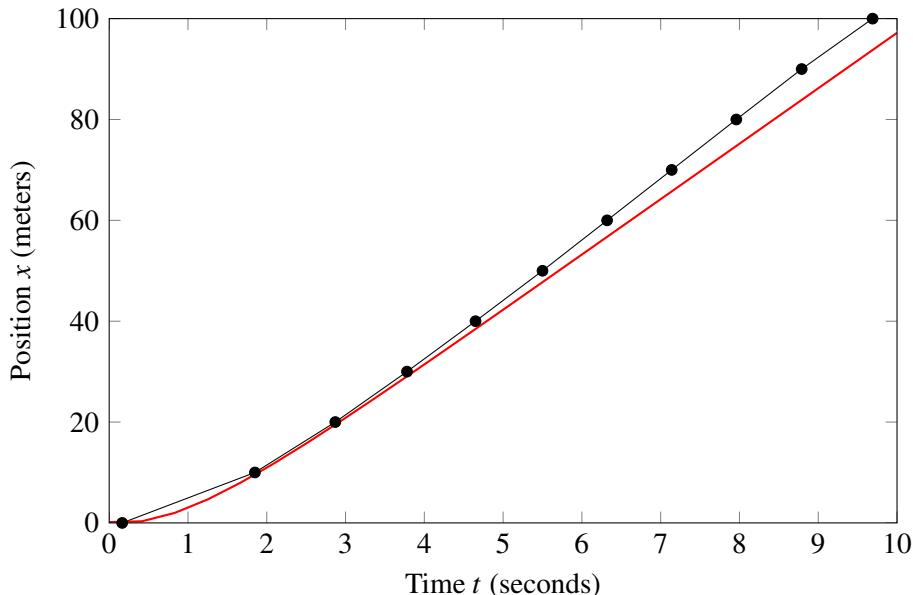


Figure 3.10: Hill-Keller data from Table 1.1 (shown as connected dots) and position $x(t)$ from (3.35) with $k = 1$ (solid red curve).

The fit looks reasonable, but is it really the best we can do? Further refinement of the value of k might give a slightly better fit, though what constitutes better is in the eye of the beholder, except in the unlikely event that we obtain perfect agreement with the data. It would be nice to have a more quantitative, objective way to obtain a best value for k .

3.4.2 Least-Squares Estimation

First, note that $x(t)$ as defined by (3.35) depends not only on t , but also on k . Let's indicate this dependence by writing

$$x(k, t) = \frac{11}{k^2} \left(e^{-k(t-t_0)} - 1 + k(t-t_0) \right). \quad (3.36)$$

Also, let the times at which Bolt's position is known be denoted by $t_0 = 0.165, t_1 = 1.85, \dots, t_{10} = 9.69$. Likewise let $x_0 = 0, x_1 = 10, \dots, x_{10} = 100$.

Suppose that some value of k , say $k = k^*$, actually yields a perfect fit to the data at each (time, distance) data point, so

$$\begin{aligned} x(k^*, t_1) - x_1 &= 0, \\ x(k^*, t_2) - x_2 &= 0, \\ &\vdots = 0, \\ x(k^*, t_{10}) - x_{10} &= 0. \end{aligned}$$

This will almost certainly not happen. Instead, we'll settle for a value of k that makes each of the quantities $x(k, t_j) - x_j$ small, on average. But this needs to be quantified, and there are any number ways to do this. One common approach is to use

$$s_j(k) = (x(k, t_j) - x_j)^2$$

as a measure of how well a given value of k yields a fit to the j th data point. Note that

- $s_j(k)$ is always nonnegative, i.e., $s_j(k) \geq 0$, and
- $s_j(k) = 0$ exactly when $x(k, t_j) = x_j$.

If $s_j(k^*) = 0$ for some k^* this means that $x(k^*, t_j) = x_j$, and so $k = k^*$ gives a perfect fit to the j th data point (but probably not the other data points). For example, you can check that $s_1(0.8988) = 0$, so the choice $k = 0.8988$ makes $x(k, t_1)$ exactly equal to $x_1 = 10$. But to make $s_2(k) = 0$ (so $x(k, t_2) = x_2$) requires $k = 0.9556$. Each data point likely needs a different value of k to obtain $s_j(k) = 0$.

Reading Exercise 3.4.1 What value of $k = k^*$ makes $s_3(k^*) = 0$ (so $x(3.78, t_3) = 30$)? For this value k^* , what is $s_1(k^*)$? What is $s_2(k^*)$?

The Sum of Squares

Since the s_j 's can't all be made to equal zero simultaneously, let's make them all as close to zero as possible in some overall sense. One way common way to do this is to seek a value of k that minimizes the quantity

$$S(k) = \sum_{j=1}^{10} s_j(k) = \sum_{j=1}^{10} (x(k, t_j) - x_j)^2.$$

The function $S(k)$ is called the **sum of squares** function for this problem. Because $S(k)$ is a sum of nonnegative terms, we have $S(k) \geq 0$ for any k . Moreover, if any quantity $x(k, t_j) - x_j \neq 0$ then $s_j(k) > 0$, and this adds a positive contribution to the value of $S(k)$, assuring that $S(k) > 0$. As

a result, $S(k) = 0$ is satisfied exactly when $x(k, t_j) - x_j = 0$ for all j , which occurs exactly when that value of k gives a perfect fit at each data point. As noted, it is unlikely that such a k exists. Minimizing the function $S(k)$ is a way to make the model fit the data at each point as well as possible in some overall sense.

Let's look at $S(k)$ for the Hill-Keller model and data. Writing out $S(k)$ explicitly by making use of (3.36) yields

$$\begin{aligned} S(k) &= \sum_{j=1}^{10} (x(k, t_j) - x_j)^2 \\ &= (x(k, t_1) - x_1)^2 + (x(k, t_2) - x_2)^2 + \cdots + (x(k, t_{10}) - x_{10})^2 \\ &= ((10 - 18.535/k - 11e^{-1.685k} - 1)/k^2)^2 \\ &\quad + ((20 - 29.755/k - 11e^{-2.705k} - 1)/k^2)^2 \\ &\quad + \cdots \\ &\quad + ((100 - 104.775/k - 11e^{-9.525k} - 1)/k^2)^2. \end{aligned} \tag{3.37}$$

Minimizing $S(k)$ is definitely a task for a computer. But since $S(k) \geq 0$ it is likely that S has an absolute or global minimum. This is easy to check in this case with a graph of $S(k)$, shown in Figure 3.11. Visually, the best value of k appears to be somewhere around $k = 0.93$.

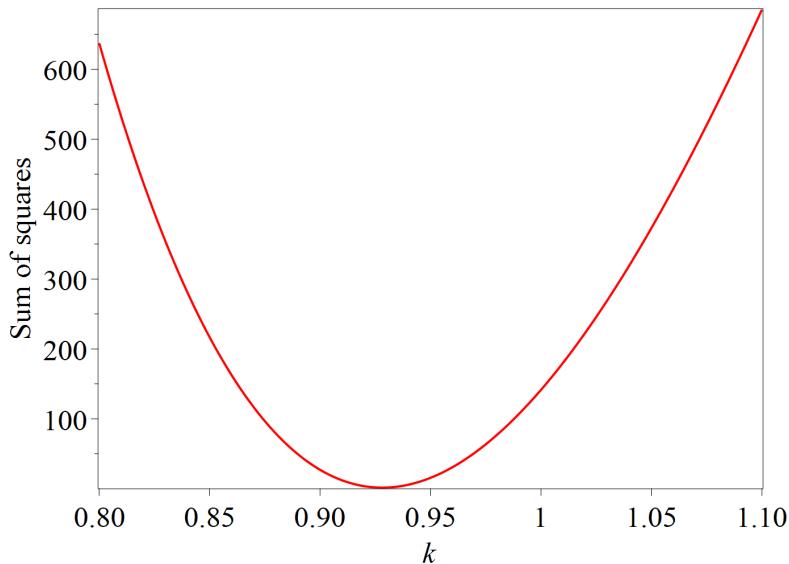


Figure 3.11: Plot of the sum of squares function $S(k)$ defined by (3.37).

Minimizing the Sum of Squares

One approach to actually computing the value of k that minimizes $S(k)$ is to use Calculus 1 techniques: compute $S'(k)$ and then solve $S'(k) = 0$. A computer algebra system such as Maple or Mathematica is a big help here. Even with S' in hand, solving $S'(k) = 0$ requires a numerical root-finding technique such as Newton's method. It can also be helpful to give the root-finding algorithm a good starting guess at the best value of k , or perhaps a range in which to seek k . Based on Figure 3.11, a starting guess of $k = 0.93$ looks good. Performing the computation yields an optimal value $k^* \approx 0.928$. The resulting graph of $x(0.928, t)$ overlayed on the data is shown in Figure 3.12. Compare this to Figure 3.10. For this example $S(0.928) \approx 1.544$, which quantifies the overall discrepancy between the actual data and the best-fit model. The value 1.544 is called the **residual sum of squares** or just **residual**.

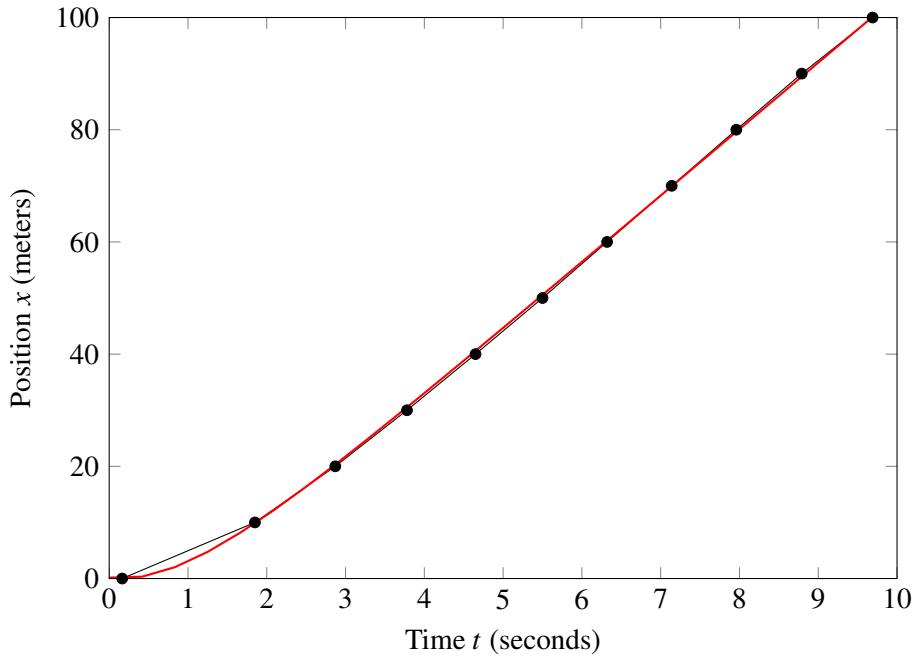


Figure 3.12: Hill-Keller data from Table 1.1 (shown as connected dots) and position $x(t)$ from (3.35) with $k = 0.928$ (solid red curve).

The process we used to estimate k is an example of **least-squares estimation**: we adjust an unknown parameter like k to obtain the best fit to the data, by minimizing a function like $S(k)$ that quantifies how well the model with parameter k fits the data. This fit is quantified with a sum of squares like (3.37). In this case the function $S(k)$ was easy to minimize, for it had a single, clear global (absolute) minimum, with no nearby local (relative) extrema. The situation won't always be so simple.

Reading Exercise 3.4.2 Suppose the ODE $u'(t) = ku(t)$ with $u(0) = 1.6$ models some physical situation and we have data points (t_j, u_j) for $1 \leq j \leq 3$ given by $(0.6, 2.1)$, $(1.1, 2.45)$, and $(1.4, 2.82)$. Use the solution $u(k, t) = 1.6e^{kt}$ to the ODE to form a sum of squares

$$S(k) = \sum_{j=1}^3 (u(k, t_j) - u_j)^2$$

and find the value $k = k^*$ that minimizes $S(k)$. Plot $u(k^*, t)$ for $0 \leq t \leq 1.5$ and compare to the data. Hint: to find k^* , start by graphing $S(k)$ for $k = 0$ to $k = 1$ or so.

3.4.3 Hill-Keller Again

The Hill-Keller model as presented in Chapter 1 contained a single unknown parameter k , which was estimated as $k \approx 0.928$ using a least-squares approach. But the ODE (1.1) also contains a parameter P , which was taken as $P = 11$ meters per second squared based on laboratory data. The parameter P might be interpreted as the maximum acceleration a runner is capable of from a standing start. It seems reasonable that the value of P may vary from runner to runner, or even over time if the runner's fitness varies. Why not try to estimate P from the data as well?

The Sum of Squares 2

The original Hill-Keller model was

$$v'(t) = P - kv(t), \quad (3.38)$$

with initial condition $v(t_0) = 0$, where $t_0 = 0.165$ and P and k are constants. The solution to (3.38) can be obtained via separation of variables or the integrating factor technique and is

$$v(t) = \frac{P}{k} \left(1 - e^{-k(t-t_0)} \right)$$

where $t_0 = 0.165$. As before position can be computed as

$$\begin{aligned} x(k, P, t) &= \int_{t_0}^t v(z) dz \\ &= \int_{t_0}^t \frac{P}{k} \left(1 - e^{-k(z-t_0)} \right) dz \\ &= \frac{P}{k} \left(z + \frac{e^{-k(z-t_0)}}{k} \right) \Big|_{z=t_0}^{z=t} \\ &= \frac{P}{k^2} \left(e^{-k(t-t_0)} - 1 + k(t-t_0) \right), \end{aligned} \tag{3.39}$$

where the notation $x(k, P, t)$ indicates the dependence of x on both P and k , as well as t .

As in the case when only k is to be estimated, we form a sum of squares function $S(k, P)$ as

$$\begin{aligned} S(k, P) &= \sum_{j=1}^{10} (x(k, P, t_j) - x_j)^2 \\ &= (x(k, P, t_1) - x_1)^2 + (x(k, P, t_2) - x_2)^2 + \cdots + (x(k, P, t_{10}) - x_{10})^2 \\ &= (10 - P(e^{-1.685k} - 1 + 1.685k)/k^2)^2 \\ &\quad + (20 - P(e^{-2.705k} - 1 + 2.705k)/k^2)^2 \\ &\quad + \cdots \\ &\quad + (100 - P(e^{-9.525k} - 1 + 9.525k)/k^2)^2. \end{aligned} \tag{3.40}$$

The function $S(k, P)$ quantifies the fit to the data given by any pair of parameters k and P in the model (3.39). The goal is to find that pair $(k, P) = (k^*, P^*)$ that minimizes $S(k, P)$.

Minimizing the Sum of Squares Part 2

It's helpful to start with a plot of $S(k, P)$. Since both k and P are independent variables here, the graph of $S(k, P)$ is a surface in three dimensional space, shown in the left panel of Figure 3.13. A contour plot of $S(k, P)$ is also shown in the right panel, with the same color scheme. The range for both plots is $0.8 \leq k \leq 1.1$, $9 \leq P \leq 13$. Note the valley in the graph of $S(k, P)$. The function $S(k, P)$ is almost constant along the bottom of this valley, at least compared to the much steeper sides. Neither the graph of $S(k, P)$ nor its contour plot make it easy to accurately gauge the location of the minimum.

In a situation like this, where the function has a large disparity in value from point to point, it can be helpful to graph $\ln(S(k, P))$ instead of $S(k, P)$ itself. Such a plot is shown in the left panel of Figure 3.14, and a contour plot of $\ln(S(k, P))$ is shown in the right panel. This makes it a bit easier to see the location of the minimum. Visual inspection suggests a minimum $k \approx 0.87$ and $P \approx 10.3$.

In order to find the minimum precisely, some computation is needed. A technique you learned in multivariable calculus is applicable here: form equations

$$\frac{\partial S}{\partial k} = 0 \text{ and } \frac{\partial S}{\partial P} = 0$$

to obtain two equations in two unknowns, k and P . These two equations are nonlinear and rather complicated, so of course we'll use the computer to form them and then solve them with a standard

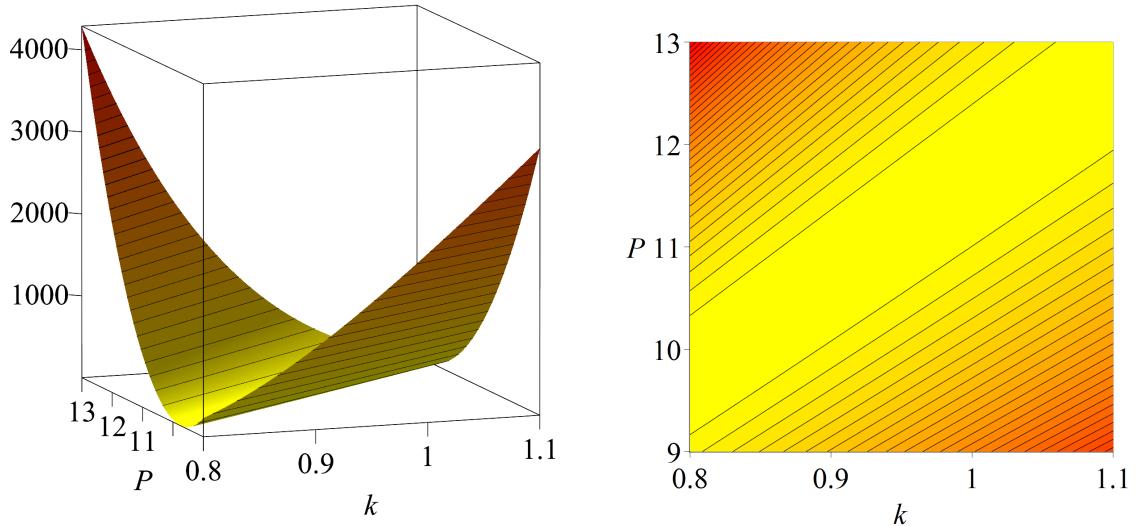


Figure 3.13: Left panel: graph of $S(k, P)$ defined by (3.40). Right panel: contour plot of $S(k, P)$.

method, such as Newton's method. In this instance the initial guess $k = 0.87$ and $P = 10.3$ leads to optimal choices $k^* = 0.865$ and $P^* = 10.38$. The value of $S(k^*, P^*)$ here is 0.775, a small improvement in the residual sum of squares (1.544) obtained by fitting only k . The resulting graph of $x(k^*, P^*, t)$ is shown in Figure 3.15, superimposed on the data. Compare this to Figures 3.10 and 3.12.

It seems that little has been gained in Figure 3.15, despite adding the second adjustable parameter P . This is a common phenomenon in parameter estimation: fitting additional parameters may make only minor improvement to the model's agreement with the data. Moreover, when a model has many free parameters to adjust, almost any data set can be fit, whether or not the model is any good.¹ In general one strives for the fewest number of parameters that provide a reasonable fit to the data. See the project “Shuttlecocks and the Akaike Information Criterion” in Section 3.5.4 for some interesting material on the **Akaike information criterion**, a technique for deciding that enough is enough when it comes to throwing more parameters into the process.

Reading Exercise 3.4.3 This is a variation on Reading Exercise 3.4.2. Again, consider the ODE $u'(t) = ku(t)$ with unknown initial condition $u(0) = A$ models some physical situation and we have data points (t_j, u_j) with $1 \leq j \leq 3$ given by $(0.6, 2.1)$, $(1.1, 2.45)$, and $(1.4, 2.82)$. Now both k and A are to be estimated from the data. Use the solution $u(k, A, t) = Ae^{kt}$ to the ODE to form a sum of squares

$$S(k, A) = \sum_{j=1}^3 (u(k, A, t_j) - u_j)^2$$

and find the pair $k = k^*, A = A^*$ that minimizes $S(k, A)$. Plot $u(k^*, A^*, t)$ for $0 \leq t \leq 1.5$ and

¹Freeman Dyson, on discussing his model for meson-proton interactions with physicist Enrico Fermi in 1954: “In desperation I asked Fermi whether he was not impressed by the agreement between our calculated numbers and his measured numbers. He replied, ‘How many arbitrary parameters did you use for your calculations?’ I thought for a moment about our cut-off procedures and said, ‘Four.’ He said, ‘I remember my friend Johnny von Neumann used to say, with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.’”—Freeman Dyson, “A meeting with Enrico Fermi.” *Nature* 427, 297 (2004). <https://doi.org/10.1038/427297a>

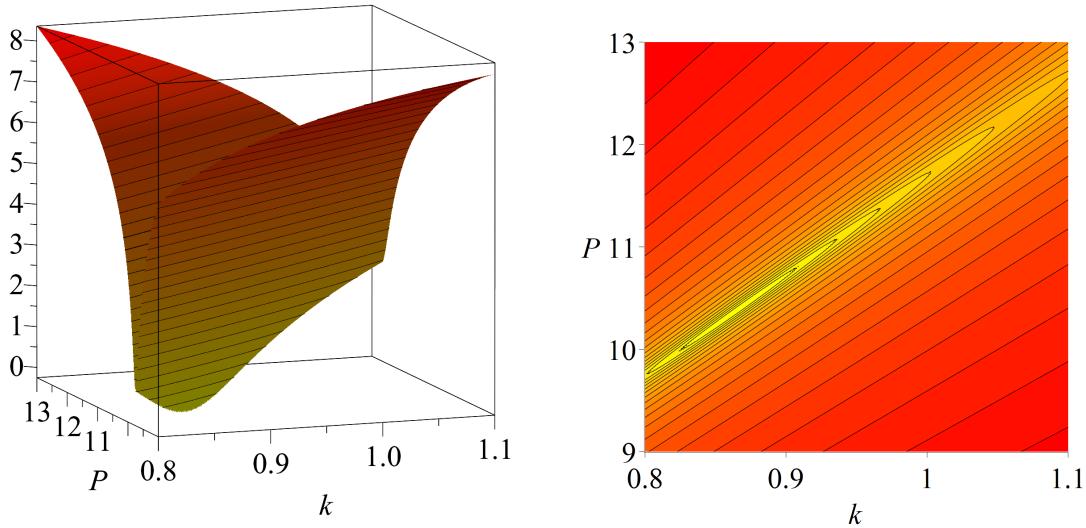


Figure 3.14: Left panel: graph of $\ln(S(k, P))$ defined by (3.40). Right panel: contour plot of $\ln(S(k, P))$.

compare it to the data. Hint: to find k^* and A^* , start by graphing the surface $z = S(k, A)$ for $0 \leq k \leq 1, 1 \leq A \leq 2$, or even $z = \ln(S(k, A))$.

3.4.4 Least Squares For ODE Parameter Estimation

The General Setting

The least-squares approach to parameter estimation is more generally applicable. Consider a first-order ODE for a function $u(t)$ and suppose the ODE involves unknown parameters $\alpha_1, \alpha_2, \dots, \alpha_m$. For notational convenience define $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)$, a vector with j th component α_j . The ODE can be expressed as

$$u'(t) = f(\boldsymbol{\alpha}, u(t))$$

with initial condition $u(t_0) = u_0$, where $f(\boldsymbol{\alpha}, u(t))$ indicates the dependence of the right side of this first-order ODE on the parameters $\alpha_1, \dots, \alpha_m$. The solution $u(t)$ itself will also depend on these parameters, so let us write $u(\boldsymbol{\alpha}, t)$ to indicate this dependence whenever convenient.

To estimate the components of $\boldsymbol{\alpha}$, we collect samples u_1, u_2, \dots, u_n of $u(\boldsymbol{\alpha}, t)$ at corresponding times t_1, t_2, \dots, t_n . The goal is to adjust the parameters α_i so that $u(\boldsymbol{\alpha}, t)$ agrees with the data as well as possible, and this is accomplished by minimizing the sum of squares function

$$S(\boldsymbol{\alpha}) = \sum_{j=1}^n (u(\boldsymbol{\alpha}, t_j) - u_j)^2. \quad (3.41)$$

In particular, we seek the value $\boldsymbol{\alpha}^* = (\alpha_1^*, \dots, \alpha_m^*)$ that minimizes S .

Remark 3.4.1 It may sometimes be convenient or advantageous to work with rescaled versions of the data and function $u(t)$ in (3.41), for example, by taking the logarithm of each. In this case we might use an alternate least-squares function of the form

$$\tilde{S}(\boldsymbol{\alpha}) = \sum_{j=1}^n (\ln(u(\boldsymbol{\alpha}, t_j)) - \ln(u_j))^2,$$

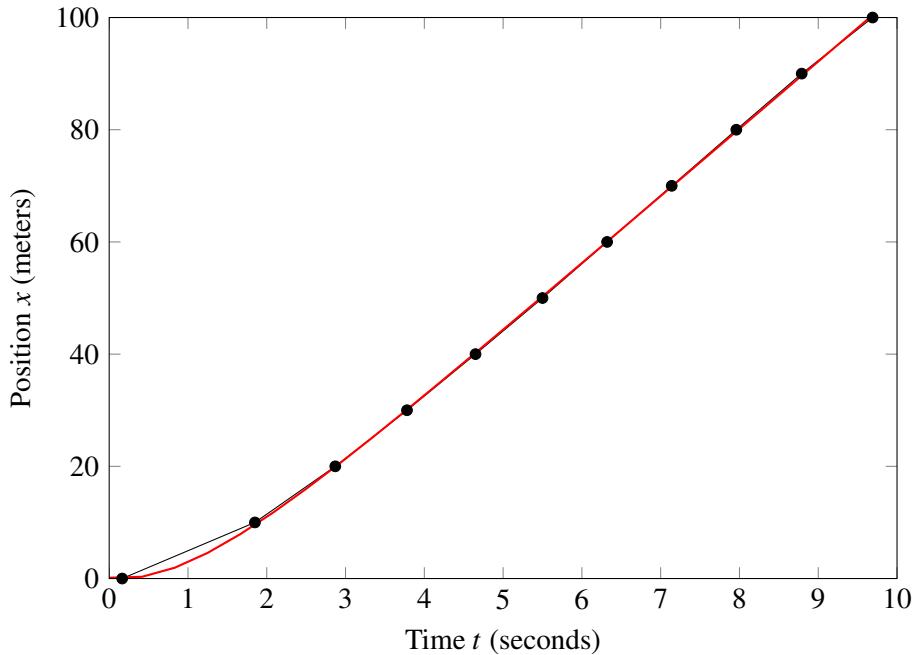


Figure 3.15: Hill-Keller data from Table 1.1 (shown as connected dots) and position $x(t)$ from (3.39) with $k = 0.928$ and $P = 10.38$ (solid red curve).

assuming here that u_j and u are positive. This is the approach that was used in the project “Chemical Kinetics” in Section 2.5. See also Exercises 3.4.9 and 3.4.10.

Minimizing the Sum of Squares Part 3

The elementary multivariable calculus approach to minimizing $S(\boldsymbol{\alpha})$ in (3.41) is to compute each partial derivative $\partial S / \partial \alpha_j$, and form m equations $\partial S / \partial \alpha_j = 0$, $1 \leq j \leq m$, for the m unknowns, $\alpha_1, \dots, \alpha_m$. This system of equations is then solved to find a critical point(s) $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$. In the Hill-Keller examples above this critical point was unique and corresponded to the minimizer for $S(\boldsymbol{\alpha})$. This was verified by plotting $S(\boldsymbol{\alpha})$, which was possible because there were only one or two parameters to estimate.

In practice this is not how S is minimized. The primary difficulty with this approach is that critical points for S need not be minima, but can also be maxima or higher dimensional saddle points. Finding and testing each critical point to determine what type it is can be laborious and impractical. What one does in practice is use algorithms specially designed for minimizing functions. This leads into the subject of **optimization**, which is concerned with theory and algorithms for the minimization of functions, and there are many algorithms for this purpose. Some algorithms are specially adapted for least-squares problems of the type we’ve considered. However, although standard minimization algorithms locate minima, they do not in general distinguish local minima from global minima. We may find a parameter estimate $\boldsymbol{\alpha}^*$ that is better than anything nearby, but not the overall best that can be obtained. The next section contains an example.

We will not go into optimization algorithms here. Most of the parameter estimation problems in this text will be light-duty and involve perhaps two or three parameters, at most. A combination of graphical and critical point methods will generally suffice. For information on solving least-squares minimization problems see [22].

3.4.5 A Cautionary Example

Least-squares estimation is not a panacea for determining model parameters from data. As a simple example, let's consider the logistic equation (1.10), $u'(t) = ru(t)(1 - u(t)/K)$, but in which the growth rate r is a function of t that varies periodically (perhaps due to some species-specific biological rhythm) as

$$r(t) = r_0 + A \sin(\omega t). \quad (3.42)$$

Here r_0 is a baseline growth rate, A is an amplitude, and ω dictates the frequency of the oscillation. In this case the logistic equation (1.10) becomes

$$u'(t) = r(t)u(t)(1 - u(t)/K), \quad (3.43)$$

with $r(t)$ given by (3.42). The ODE (3.43) involves parameters r_0, A, ω , and K . Any or all of these parameters might be considered as unknowns to be estimated from population data.

The ODE (3.43) with initial condition $u(0) = u_0$ can be solved analytically using separation of variables. The solution is

$$u(t) = \frac{K}{e^{-R(t)}(K/u_0 - 1) + 1}, \quad \text{where} \quad R(t) = \int r(t) dt \quad (3.44)$$

with the antiderivative chosen to satisfy $R(0) = 1$. Consider the case in which $K = 5, A = 0.7, u_0 = 1, r_0 = 0.2$, and $\omega = 1.5$. The graph of the solution $u(t)$ to (3.44) is shown in Figure 3.16 on the range $0 \leq t \leq 40$.

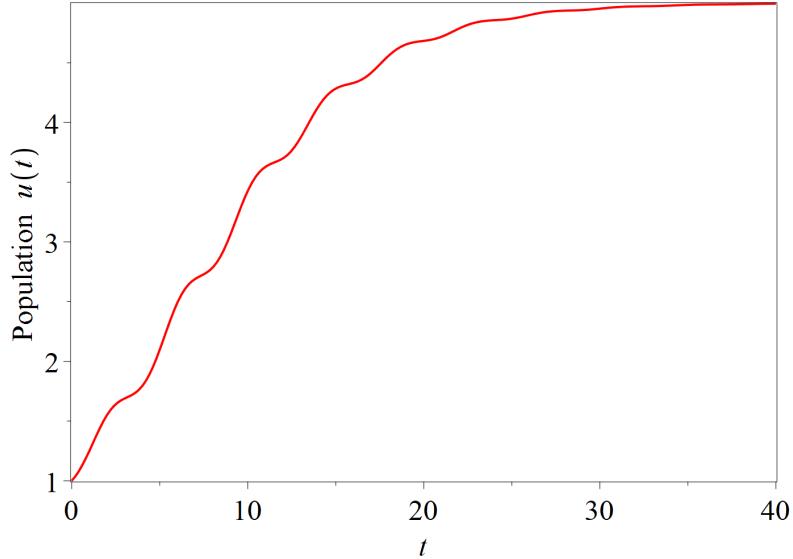


Figure 3.16: Function (3.44) satisfying the logistic equation (3.43) with time-varying growth rate, parameters $K = 5, A = 0.7, u_0 = 1, r_0 = 0.2$, and $\omega = 1.5$.

Suppose data is obtained by sampling $u(t)$ at times $t_0 = 0, t_1 = 2, \dots, t_{20} = 40$ to produce samples u_0, u_1, \dots, u_{20} . Let us consider the parameters K, A , and r_0 as known with only ω to be determined. To estimate ω , form the sum of squares

$$S(\omega) = \sum_{j=0}^{20} (u(\omega, t_j) - u_j)^2, \quad (3.45)$$

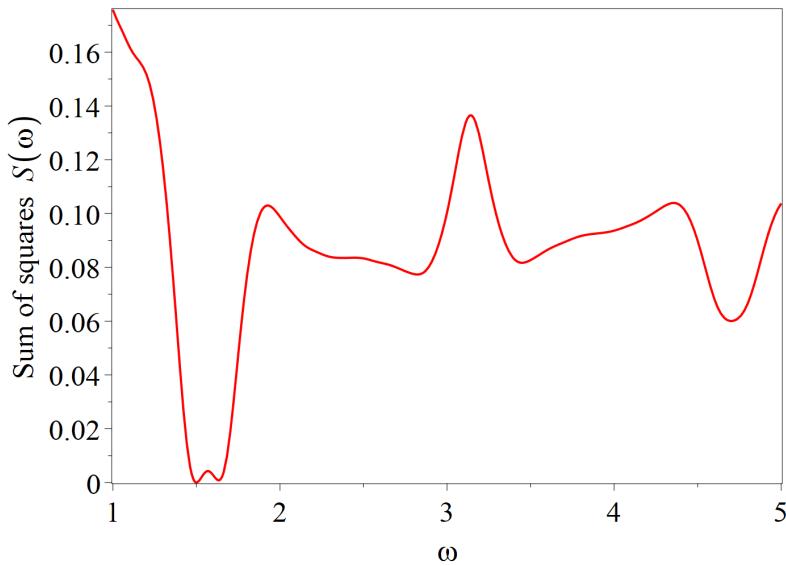


Figure 3.17: Sum of squares $S(\omega)$ defined by (3.45).

where $u(\omega, t)$ denotes the dependence of u on ω . Moreover, assume that the data is perfect, with no error or noise. A plot of the resulting sum of squares function $S(\omega)$ is shown in Figure 3.17.

Even in this ideal case, Figure 3.17 illustrates that the function $S(\omega)$ has many critical points. Solving $S'(\omega) = 0$ to locate critical points and then sorting through them for minima may be laborious. Even a modern optimization algorithm will very likely find a suboptimal local minimum, rather than the global minimum at $\omega = 1.5$. And this is just the noise-free case; noisy data makes things even more difficult.

The situation gets worse when two or more parameters are in play. More critical points and local minima are usually present, and in the case in which three or more parameters are to be estimated we even can't graph the sum of squares function. In situations such as these it pays to have a good estimate of what reasonable parameter values are, and then use them to inform the minimization algorithm's search.

3.4.6 Exercises

Exercise 3.4.1 Consider the set of data points (t_j, u_j) with $1 \leq j \leq 4$,

$$\{(0.1, 0.11), (0.6, 0.5), (1.1, 0.6), (1.4, 0.5)\}.$$

For each function u of the given form in (a)-(d), construct an appropriate least-squares function to fit this data and then minimize with respect to the specified parameters. In each case compute the residual, graph the best-fit function u , and compare to a plot of the data.

- (a) $u(a, t) = at$, parameter a .
- (b) $u(a, b, t) = at + b$, parameters a and b . Compare the residual to part (a); it should be smaller. Why?
- (c) $u(a, t) = at^2$, parameter a .
- (d) $u(a, b, c, t) = at^2 + bt + c$, parameters a, b , and c . Compare the residual to parts (b) and (c); it should be smaller than the residual for each of these. Why?

Exercise 3.4.2 In Section 2.5.2 a model for a first-order chemical reaction was presented, and led to the ODE $y'(t) = -ky(t)$ (this was equation (2.77) where t is time and $y(t)$ the concentration of the reactant). With initial condition $y(0) = y_0$ the solution is $y(t) = y_0 e^{-kt}$. In Table 3.5 is a subset of the data from Table 2.3 for the decomposition of hydrogen peroxide.

As detailed in Modeling Exercise 5.2.3 for “Chemical Kinetics” of Section 2.5, this reaction should be first-order, so that if $y(t)$ denotes the concentration $[H_2O_2]$ of H_2O_2 then taking the logarithm of both sides of $y(t) = y_0 e^{-kt}$ yields $\ln(y(t)) = \ln(y_0) - kt$. Since $y_0 = 1$ here, $\ln(y_0) = 0$ and so

$$\ln(y(t)) = -kt. \quad (3.46)$$

We will use (3.46) to adjust k to best fit the data.

Using whatever software you have available, form the sum of squares

$$S(k) = (\ln(0.78) + 300k)^2 + (\ln(0.37) + 1200k)^2 + (\ln(0.08) + 3000k)^2$$

that is appropriate to (3.46) (the first term, $(\ln(1) + 0k)^2$ corresponding to data point $(0, 1)$, is always zero). Plot $S(k)$ on the range $0 \leq k \leq 0.002$, then set $S'(k) = 0$ and solve to find the minimizing value k^* for k . What is the residual sum of squares? Also plot the linear function $-k^* \ln(t)$ and compare the graph to a plot of the data pairs (time, $\ln([H_2O_2])$) from Table 3.5.

Time (seconds)	$[H_2O_2]$ (mol/L)
0	1.00
300	0.78
1200	0.37
3000	0.08

Table 3.5: Subset of data from Table 2.3.

Exercise 3.4.3 Table 3.6 contains split data for Tori Bowie’s gold medal women’s 100-meter victory in the 2017 IAAF World Championships (see [21]). Use the procedure of Section 3.4.3 to find the best choices k^* and P^* for k and P in the Hill-Keller model for this data; note that the initial condition is $v(0.182) = 0$ and that you need to fit her position function $x(t)$ to the data (not velocity), as we did with Usain Bolt’s data in Section 3.4.3. Plot the function $x(t)$ with these parameters and compare to the data.

Time (seconds)	0.182	2.07	3.22	4.24	5.18
Position (meters)	0	10	20	30	40
Time (seconds)	6.11	7.04	7.98	8.93	9.88
Position (meters)	50	60	70	80	90
Time (seconds)	10.85				
Position (meters)	100				

Table 3.6: Race splits (seconds) every 10 meters for Tori Bowie’s gold medal run at the 2017 IAAF World Championship 100-meter race.

Exercise 3.4.4 Use (3.39) along with the optimal parameter values $k = 0.865$ and $P = 10.38$ for Usain Bolt found in Section 3.4.3 to predict how fast Bolt could run 200 meters, then compare your prediction to the current 200-meter world record. What might account for any significant discrepancy? Do the same to predict how fast Bolt could run a mile (1609.34 meters) or a marathon (42195 meters). Compare to the current world records. Why are the predictions so far off?

Exercise 3.4.5 Suppose we have data points (x_j, y_j) for $j = 1$ to $j = n$ and that we wish to fit a model in the form $y_j = mx_j$ (a straight line through the origin). To do this we form a sum of squares

$$S(m) = \sum_{j=1}^n (y_j - mx_j)^2$$

and then minimize with respect to m . It is desirable that this function should actually have a minimizer, and that the minimizer is unique. One property that a function $f(m)$ of a single variable can possess that guarantees a unique minimizer exists is that

$$f''(m) > 0 \quad \text{and} \quad \lim_{m \rightarrow \pm\infty} f(m) = \infty.$$

Verify that $S(m)$ has these properties if at least one of the $x_j \neq 0$, so any such least-squares problem has a unique solution. Hint: show that $S(m)$ can be expressed as

$$S(m) = \left(\sum_{j=1}^n x_j^2 \right)^2 m^2 - 2 \left(\sum_{j=1}^n x_j y_j \right) m + \sum_{j=1}^n y_j^2.$$

Exercise 3.4.6 Consider adjusting a parameter m to fit data points (x_j, y_j) for $1 \leq j \leq n$ where the model is $y_j = mx + b$ for some known constant b . Use the result of Exercise 3.4.5 to show that the sum of squares $S(m) = \sum_{j=1}^n (y_j - mx_j - b)^2$ has a unique minimizer for m . Hint: define $\tilde{y}_j = y_j - b$.

Exercise 3.4.7 Another approach to fitting parameters is called **L^1 -minimization**. In this technique rather than measure the discrepancy between a data point u_j and the solution $u(t_j)$ as $(u(t_j) - u_j)^2$, we use the absolute value $|u(t_j) - u_j|$. The analogue of (3.41) is

$$S(\boldsymbol{\alpha}) = \sum_{j=1}^n |u(\boldsymbol{\alpha}, t_j) - u_j| \tag{3.47}$$

where $\boldsymbol{\alpha}$ embodies the parameter(s) to be fit. We seek that value $\boldsymbol{\alpha}^* = (\alpha_1^*, \dots, \alpha_m^*)$ for the parameters that minimizes S in (3.47). This technique has certain advantages over least squares estimation; in particular, it can be more robust against large errors in the data u_j . The drawback is that $S(\boldsymbol{\alpha})$ as defined by (3.47) is not typically differentiable, so it's not straightforward to make use of derivative information in seeking a minimum. Nonetheless, numerical algorithms do exist for this type of minimization problem.

Redo Exercise 3.4.2 using L^1 -minimization. In particular, graph the function

$$S_1(k) = |\ln(0.78) + 300k| + |\ln(0.37) + 1200k| + |\ln(0.08) + 3000k|$$

on the range $0 \leq k \leq 0.002$ and visually identify the value of k that provides a minimum. Zooming in on promising k values may be helpful.

Exercise 3.4.8 This is an extension of Exercise 2.2.8. In Table 3.7 are some population data from a classic study [35], concerning the growth of a species of yeast. Let us model the growth of this yeast species using the ODE (1.10), with solution given by (1.11). The parameters we wish to estimate are the intrinsic growth rate r and the carrying capacity K . Based on the data in Table 3.7, we will take $u_0 = 9.6$ for the initial population.

- (a) Plot the data as (time, population) pairs. Does this seem to obey logistic growth? Can you estimate the carrying capacity K from a visual inspection of the data?
- (b) Let

$$u(r, K, t) = \frac{K}{1 + e^{-rt}(K/9.6 - 1)}$$

denote the solution to (1.11) (with $u_0 = 9.6$ assumed). Using whatever software you have available, write out the sum of squares

$$S(r, K) = \sum_{j=1}^{18} (u(r, K, j) - p_j)^2,$$

where j indexes time in hours and p_j denotes the measured population in millions at hour j from Table 3.7. (There are 18 data points.)

- (c) Minimize $S(r, K)$ as a function of r and K by setting

$$\frac{\partial S}{\partial r} = 0 \quad \text{and} \quad \frac{\partial S}{\partial K} = 0$$

and solving for r and K (you'll have to use a numerical method). It may be helpful to start the numerical solver with a good guess at r and K .

- (d) If $r = r^*$ and $K = K^*$ are your minimizing values for r and K , plot $u(r^*, K^*, t)$ and compare it to a plot of the data. Does this seem like a good model?

Time (hours)	0	1	2	3	4	5	6	7	8	9
Population (millions)	9.6	18.3	29.0	47.2	71.1	119.1	174.6	257.3	350.7	441.0

Time (hours)	10	11	12	13	14	15	16	17
Population (millions)	513.3	559.7	594.8	629.4	640.8	651.1	655.9	659.6

Table 3.7: Yeast data with population in millions, time in hours.

Exercise 3.4.9 This exercise is inspired by the SIMIODE project “Potato Cooling” [126]. A medium-sized potato was placed in a microwave oven for two minutes, removed, and then its

temperature was monitored using a cooking thermometer. The data for various times (minutes) after the potato was removed from the microwave oven are tabulated in Table 3.8. The ambient temperature of the room was 72 degrees Fahrenheit.

Recall Newton's law of cooling is quantified by the ODE (2.14), with solution $u(t) = A + (u_0 - A)e^{-kt}$ previously developed in (2.15). Here A is the ambient temperature, u_0 the initial temperature, and $k > 0$ a cooling constant. With $A = 72$ and $u_0 = 204$ it follows that $u(t) = 72 + 132e^{-kt}$ for the potato, with k to be estimated from the data. With time-temperature data pairs (t_j, u_j) for $1 \leq j \leq n$, we expect $u_j \approx 72 + 132e^{-kt_j}$ for the correct value of k .

In accord with (3.41) (where α there is just k here) and using the data from Table 3.8, form an appropriate sum of squares

$$S(k) = (72 + 132e^{-2k} - 193)^2 + \cdots + (72 + 132e^{-30k} - 130)^2$$

(noting that the initial term at time $t = 0$ is always zero). Then:

- Graph $S(k)$ for $0.01 \leq k \leq 0.05$ and visually identify the minimum.
- Find that value k^* that minimizes $S(k)$. What is the residual sum of squares?
- Plot the resulting function $u(t) = 72 + 132e^{-k^*t}$ and compare it to the data. Does this seem like a reasonable model?

Time (minutes)	0	2	4	8	10
Temperature (°F)	204	193	184	169	162
Time (minutes)	13	17	20	24	30
Temperature (°F)	156	149	143	138	130

Table 3.8: Potato temperature data with time in minutes and temperature in degrees Fahrenheit.

Exercise 3.4.10 This is a variation on Exercise 3.4.9. It may be helpful to review Remark 3.4.1. For the case $A = 72$ and $u_0 = 204$ in Newton's law of cooling, the data in Table 3.8 might be modeled as $u(t) = 72 + 132e^{-kt}$. With (time,temperature) data pairs (t_j, u_j) for $1 \leq j \leq n$ we expect $u_j \approx A + (u_0 - A)e^{-kt_j}$ for the correct value of k , or $u_j - A \approx (u_0 - A)e^{-kt_j}$ for $1 \leq j \leq n$. Take the logarithm of both sides of this last equation and rearrange to find (assuming $u_0 > A$)

$$\ln(u_j - A) \approx \ln(u_0 - A) - kt_j,$$

$1 \leq j \leq n$. We can then seek an optimal value of k by minimizing an alternate sum of squares function

$$\tilde{S}(k) = \sum_{j=1}^n (\ln(u_j - A) - \ln(u_0 - A) + kt_j)^2.$$

In the present case this is

$$\tilde{S}(k) \approx (2k - 0.087)^2 + \cdots + (30k - 0.822)^2.$$

Using whatever software you have available, form the function $\tilde{S}(k)$ and then:

- Plot $\tilde{S}(k)$ for $0 \leq k \leq 0.05$ and visually identify the minimum.
- Find the minimizer k^* by solving $\tilde{S}'(k) = 0$.

- (c) Plot the resulting function $u(t) = 72 + 132e^{-k^*t}$ and compare it to the data. Does this seem like a reasonable model? Compare this estimate of k to that obtained in Exercise 3.4.9.
- (d) What advantage might minimizing $\tilde{S}(k)$ have over minimizing $S(k)$ as we did in Exercise 3.4.9?

Exercise 3.4.11

- (a) Newton's law of cooling assumes that the rate of temperature change is proportional to $u - A$, the difference between the object's current temperature and the ambient temperature. A more general and flexible model might posit a general functional relationship between this rate of change and the temperature difference, of the form

$$u'(t) = -F(u(t) - A) \quad (3.48)$$

for some function F , where $F(0) = 0$ and F is strictly increasing. Sketch a phase portrait for this ODE under these assumptions. Why is $F(0) = 0$ reasonable? Why should we require that F be strictly increasing?

- (b) One variation on Newton's law of cooling is to take

$$F(v) = \begin{cases} -k|v|^r, & v < 0 \\ k|v|^r, & v \geq 0 \end{cases}$$

in (3.48) for some positive real numbers k and r . The usual Newton's law of cooling corresponds to $r = 1$. Verify that F satisfies $F(0) = 0$ and that F is strictly increasing.

- (c) With F as in part (b) the ODE (3.48) becomes

$$u'(t) = \begin{cases} -k|u(t) - A|^r, & v < 0 \\ k|u(t) - A|^r, & v \geq 0 \end{cases} \quad (3.49)$$

However, we are interested in the case in which $u(t) > A$ at all times (our potato started and stayed above the ambient room temperature). In this case the ODE (3.49) becomes

$$u'(t) = -k(u(t) - A)^r. \quad (3.50)$$

Verify that if $r \neq 1$ the solution to (3.50) with $u(0) = u_0 > A$ is given by

$$u(t) = A + ((u_0 - A)^{1-r} + k(r-1)t)^{1/(1-r)}. \quad (3.51)$$

- (d) For the potato data $A = 72$ and $u_0 = 204$. Consider k and r as parameters to be estimated from the data in Table 3.8. Form an appropriate sum of squares $S(k, r)$ using (3.51). Plotting this function reveals little—it's very difficult to tell where a minimum is. Instead, try plotting $\ln(S(k, r))$ on the range $0 \leq k \leq 0.0004, 2 \leq r \leq 2.5$ (it looks like there are many local minima in this region).
- (e) Set $\frac{\partial S}{\partial k} = 0$ and $\frac{\partial S}{\partial r} = 0$ and solve for k and r in the range $0 \leq k \leq 0.0004, 2 \leq r \leq 2.5$ (you should find a solution). Use this optimal value for k and r in (3.51) to plot the model that best fits the data. Does it seem reasonable? Compare the value of r to the choice $r = 1$ used in the standard Newton's law of cooling. How does the residual sum of squares here compare to that of Exercise 3.4.9?

3.5 Modeling Projects

In this section we offer four modeling opportunities, based on projects from the SIMIODE website [9] and book website [8]. All of these projects involve parameter estimation, in conjunction with ODE models—sometimes only one parameter, sometimes several. In each the parameter estimation can be done informally (guess and check/plot) or with a more formal least-squares procedure.

3.5.1 Project: Sublimation of Carbon Dioxide

This modeling project is based on the SIMIODE Modeling Scenario “Sublimation of Carbon Dioxide” [121].

Sublimation

Matter can exist in a number of states: plasma, gas, liquid, and solid. When a solid becomes liquid, the object melts. When a liquid becomes a gas, the matter evaporates. When a solid transitions directly to a gas we say that it **sublimates**. A readily available example of sublimation occurs when solid dry ice (solid carbon dioxide or CO_2) becomes gaseous.

A solid block of dry ice at room temperature will slowly disappear as it sublimates into its gaseous state, and the mass of the block decreases over time. A reasonable and interesting question to ask is, “At what rate does the mass decrease as the dry ice sublimates?” A related question is, “What does the rate of change of the mass depend upon?”

An Experiment

Consider the apparatus depicted in Figure 3.18. The apparatus consists of a scale on which is mounted a ring stand to hold a solid block of CO_2 away from the base, so that the mass of the sublimating dry ice can be measured. This configuration will not allow condensed moisture created by the falling cold sublimated CO_2 to collect on any surface involved in the scale apparatus, a mistake one of the authors made in his first attempt at designing such an apparatus.

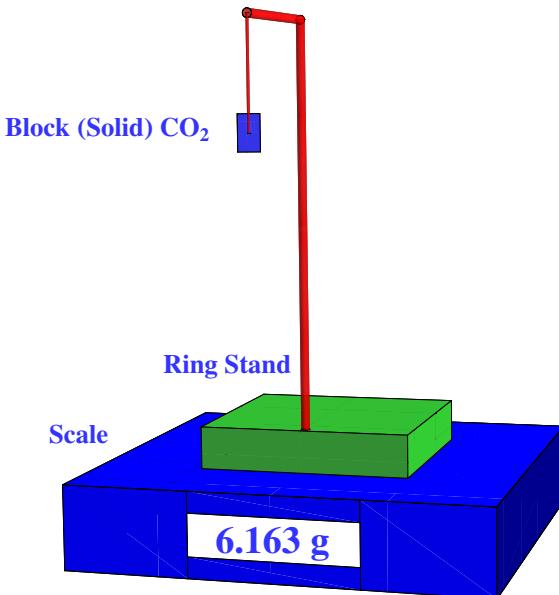


Figure 3.18: Apparatus for measuring the mass of a sublimating block of dry ice.

This apparatus was used to measure the mass of a block of dry ice over a period of one hour. The data is presented in Table 3.9. The data was collected in a room at a temperature of 19.3°C

using a small box-shaped piece of dry ice approximately $1\text{ cm} \times 1\text{ cm} \times 2\text{ cm}$. A plot of this data is shown in Figure 3.19.

Time (seconds)	Mass (g)	Time (seconds)	Mass (g)
0	25.525	1800	13.553
120	24.512	1930	12.910
240	23.524	2040	12.331
360	22.639	2170	11.689
480	21.765	2280	11.188
600	20.890	2410	10.566
720	20.043	2530	10.043
840	19.221	2690	9.377
960	18.431	2780	9.011
1080	17.677	2880	8.616
1200	16.936	3060	7.945
1320	16.220	3220	7.404
1440	15.548	3380	6.877
1570	14.828	3480	6.593
1680	14.213	3600	6.244

Table 3.9: Data collected by students Masood Makkar and Paul Werner (3 December 1992) on successful run for mass of dry ice (g) as a function of time (s).

The goal of this project is to model the data in Table 3.9. In particular, let $m(t)$ denote the mass of the dry ice block in grams at time t (seconds). We seek a differential equation that reasonably models the evolution of $m(t)$ over time, in the form

$$m'(t) = F(t, m(t)) \quad (3.52)$$

with initial condition $m(0) = m_0$.

Modeling Sublimation

Modeling Exercise 5.1.1 Consider the form of the function $F(t, m)$ in (3.52). What form should it take? Some things to think about: Should it be autonomous? If so, what fixed point(s) or equilibrium solution(s) should it have? How should m' behave if m is larger? Note that in this setting we only care about $m \geq 0$. Sublimation occurs only at the surface of the dry ice block—is that relevant? And of course, consider the graph in Figure 3.19. In any case, the ODE you come up with should certainly have at least one adjustable parameter that can be estimated from the data at hand, either by visual fitting or a least-squares approach.

Strive to balance the conflict between a complicated model that captures all facets of the process at hand and a model that embraces the “KISS” philosophy (Keep It Simple, Stupid.) It might be beneficial, on a first pass, to come up with an analytically solvable ODE.

Modeling Exercise 5.1.2 Solve the ODE you came up with in Modeling Exercise 5.1.1 with an appropriate initial condition. The solution should contain whatever undetermined parameters you introduced in that Modeling Exercise.

Modeling Exercise 5.1.3 You can take a guess-and-check approach to fit the parameters in your ODE/solution (guess, plot the ODE solution, compare to the data) or form an appropriate sum of squares and then minimize to produce estimates. In either case, plot the solution to the ODE with your estimated parameters and compare to the data. Comment. Does it seem reasonable? How might it be improved? If you see a way to improve the model, do it, and explain.

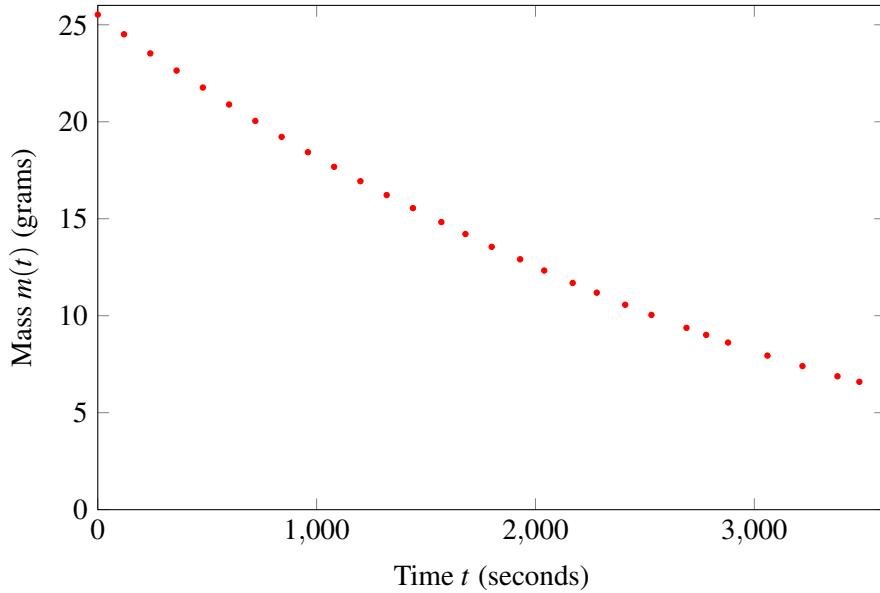


Figure 3.19: Sublimation data from Table 3.9.

3.5.2 Project: Fish Harvesting Revisited

Recall the Atlantic cod population logistic growth model with harvesting from Section 1.3, based on the SIMIODE project “Fishery Harvesting” [42]. We developed an ODE (1.12), reproduced here for convenience:

$$u'(t) = ru(t) \left(1 - \frac{u(t)}{K}\right) - h(t)u(t). \quad (3.53)$$

Here $u(t)$ represents the population of Atlantic cod (as measured in metric tons of biomass), r is the intrinsic growth rate of the species, K is the carrying capacity, and $h(t)$ is the population harvesting rate on a percentage basis relative to $u(t)$; we previously presumed h is constant, but now allow it to be a function of time. The ODE also had an initial condition $u(0) = u_0$.

Parameter Estimation

The goal is to estimate the parameters K and r from the data in Table 1.2, reproduced here as Table 3.10 for convenience. The last data point from Table 1.2 is omitted, since the harvest rate at that time was not known. In this version of the table all quantities are indexed from $j = 0$, which corresponds to the year 1978.

One difficulty is that (3.53) is not obviously solvable in closed-form for a general function $h(t)$. Without an analytical solution, how are we to form a sum of squares to determine how well any given choice for r and K fit the data? One approach is this: instead of solving the ODE and comparing the solution to the raw data, use the data to approximate the ODE and compare the result to (3.53).

Specifically, suppose that as in Table 3.10, we have data points u_j at times $t = t_j$, for $0 \leq j \leq n$.

- For each time t_j for $0 \leq j \leq n-1$ with data u_j , construct an approximation $u'_j \approx u'(t_j)$ as

$$u'_j = \frac{u_{j+1} - u_j}{\Delta t_j} \quad (3.54)$$

where Δt_j is the time interval between data points j and $j+1$. With the data from Table 1.2, the appropriate choice is $\Delta t_j = 1$ for all j . The quantity u'_j is called a **finite difference** approximation to $u'(t_j)$.

Year j	u_j	h_j	Year	u_j	h_j	Year	u_j	h_j
0	72,148	0.18847	10	68,702	0.23154	20	20,196	0.18953
1	73,793	0.14974	11	61,191	0.20860	21	25,776	0.17011
2	74,082	0.21921	12	49,599	0.33565	22	23,796	0.15660
3	92,912	0.17678	13	46,266	0.29534	23	19,240	0.28179
4	82,323	0.28203	14	34,877	0.33185	24	16,495	0.25287
5	59,073	0.34528	15	28,827	0.35039	25	12,167	0.25542
6	59,920	0.20655	16	21,980	0.28270	26	21,104	0.08103
7	48,789	0.33819	17	17,463	0.19928	27	18,871	0.08739
8	70,638	0.14724	18	18,057	0.18781	28	21,241	0.08195
9	67,462	0.19757	19	22,681	0.19357	29	22,962	0.10518

Table 3.10: Annual (1978-2007) values of Atlantic cod biomass in metric tons, u_j , and harvest rate, h_j , in Georges Bank from [130].

2. Form a discrete version of the ODE (3.53) as

$$\frac{u_{j+1} - u_j}{\Delta t_j} = ru_j(1 - u_j/K) - h_j u_j$$

for $0 \leq j \leq n - 1$, where r and K are to be determined, and the harvest rates h_j are tabulated in Table 3.10. The $u'(t)$ term in (3.54) at time $t = t_j$ has been replaced by u'_j , while $u(t_j)$ and $h(t_j)$ have been replaced by u_j and h_j , respectively.

Of course (3.54) is not likely to hold for any choice of r and K , but a sum of squares function

$$S(r, K) = \sum_{j=0}^{n-1} \left[\frac{u_{j+1} - u_j}{\Delta t_j} - (ru_j(1 - u_j/K) - h_j u_j) \right]^2 \quad (3.55)$$

can be formed to measure how well the discrete version of the ODE is satisfied for any given choice of r and K .

3. Minimize $S(r, K)$ to produce estimates for r and K .

Steps 1 to 3 above allow us to do an end run around the ODE solution process.

For the fish data in Table 3.10, there are $n = 30$ data points for the u_j (years 1978 to 2007) and 30 data points for h_j (1978 to 2007). Let us estimate r and K for (3.53) in this setting.

Modeling Exercise 5.2.1 Using whatever software you have available, form the sum of squares function $S(r, K)$ according to (3.55).

Modeling Exercise 5.2.2 Minimize $S(r, K)$. A plot of $S(r, K)$ on the range $4 \times 10^4 \leq K \leq 3 \times 10^5$, $0.1 \leq r \leq 0.5$ is a good start. It may be helpful to plot $\ln(S(r, K))$ as well.

Modeling Exercise 5.2.3 Let r^* and K^* denote the least-squares estimates for r and K . A function for $h(t)$ isn't given explicitly, so the ODE (3.54) can't be solved analytically, but we can use our estimates for r and K along with the data h_j for the harvesting rates to solve the ODE (3.54) numerically using Euler's method with step size 1. Specifically, define $U_0 = u_0 = 72148$ and then define

$$U_{j+1} = U_j + r^* U_j (1 - U_j/K^*) - h_j U_j$$

for $j = 0$ to $j = n - 2$. The result is an Euler estimate of the solution to (3.54) using step size 1 with the least-squares estimates for r and K , that uses the sampled values of the harvesting rate $h(t)$ at times $t = t_j$.

Plot the solution defined by the pairs (t_j, U_j) for $0 \leq j \leq n - 1$ (note $t_j = j$ here, if 1978 is year 0.) Compare to a plot of the data from Table 1.2. Comment on the fidelity of the model. If the fit isn't perfect, what might be improved?

3.5.3 Project: The Mathematics of Marriage

This modeling project is based on the SIMIODE modeling scenarios “Mathematics of Marriage” [70], “At What Age Do People Get Married?” [116], and a model developed in [61]. We explore the process of entry into marriage by individuals. In particular, we wish to model the fraction of people in a given sociological group who are married, as a function of age.

The Model

Consider some of the societal factors that cause people to marry. In this project a model for this process will be built based upon the following assumptions:

1. Social pressure: As the fraction of married people in an age group increases over time, people in that age group may feel more pressure to get married.
2. Age: The chances for marriage decline as one gets older.

Modeling Exercise 5.3.1 Do these assumptions seem reasonable? Write down several other factors that may affect the probability of an individual in a given age group marrying.

The model to be constructed will consist of a differential equation in which the dependent variable is the fraction of individuals in a cohort already married. In this work, the term **cohort** refers to the group of men or women who were born in a specific time period. For example, the women in the US born between 1970 and 1974 may be considered as a cohort. The independent variable in this model will be time t , measured in years.

Let us assume that the cohort of interest contains n people and that this number does not change with time. Let $m(t)$ be the number of people in that cohort who are already married at time t and let $P(t)$ denote the fraction of people in the cohort who are married at time t , so

$$P(t) = m(t)/n. \quad (3.56)$$

We will develop a differential equation for $P(t)$.

Consider a short time interval from time t to time $t + dt$. There is a certain probability that an unmarried person in the cohort will marry in this time period. We assume this is the same for each unmarried person, and that this probability approximately obeys the relation that:

$$\text{the probability of an individual marrying in time interval } (t, t + dt) = p(t) dt \quad (3.57)$$

for some function $p(t)$. On a short time interval of length dt the probability is thus approximately proportional to dt .

If each unmarried person in the cohort behaves independently of the others (one person getting married does not change the probability of another marrying) then we expect that in a short time interval t to $t + dt$ the increase dm in the number of married persons is, from (3.57), approximately

$$dm = \underbrace{(n - m(t))}_{\text{unmarried persons}} \times \underbrace{p(t) dt}_{\text{probability of any individual marrying}}$$

Dividing both sides above by dt and taking the limit as $dt \rightarrow 0$ yields

$$\frac{dm}{dt} = (n - m(t))p(t). \quad (3.58)$$

Divide both sides of (3.58) by n and use (3.56) to obtain

$$\frac{dP}{dt} = (1 - P(t))p(t). \quad (3.59)$$

This is an ODE for $P(t)$, but some choice for the function $p(t)$ in (3.57) is needed. This is where assumptions (1) and (2) come into play.

Let us assume that

$$p(t) = q(t)P(t) \quad (3.60)$$

where

$$q(t) = Ab^t \quad (3.61)$$

for constants A and b with $0 < A < 1$ and $0 < b < 1$. Here A is the initial ($t = 0$) average marriage potential of the cohort and b is a deterioration term. Note that $0 < q(t) < 1$ and $q(t)$ strictly decreases to 0 from the initial value $q(0) = A$. Together (3.60) and (3.61) capture assumptions (1) and (2) above: $p(t)$, the probability of an individual marrying between t and $t + dt$, is proportional to the fraction of married persons in the cohort, and that this probability decreases with time.

If we substitute $q(t)$ as defined in (3.61) into (3.60) we obtain $p(t) = Ab^t P(t)$. Substituting this into (3.59) yields an ODE

$$\frac{dP}{dt} = Ab^t P(t)(1 - P(t)), \quad (3.62)$$

for $P(t)$, the fraction of married persons in the cohort as a function of time. This is the model to use in what follows. We'll use $t = 0$ to refer to the lowest age of members of the cohort (in the data sets below, 20 years old).

Modeling Exercise 5.3.2 Review the derivation of (3.62) and explicitly list some assumptions that were made, especially any assumptions you see that were not stated explicitly. Do they seem like a reasonable first approximation? (Think about how you might change them too, as you'll have a chance to do so later.)

Modeling Tip 3.5.1 Now that we have an ODE (3.62) to model the situation, the next step is to solve this ODE. But it's always a good idea to do a quick sanity check to see if the model might yield predictions in accord with the situation being modeled (or won't), and this can sometimes be done without solving the ODE. See Modeling Exercise 5.3.3 below.

Modeling Exercise 5.3.3 Since P is a fraction (or proportion), it should always be the case that $0 \leq P(t) \leq 1$, and in this exercise we show this to be true if the initial condition satisfies $0 \leq P(0) \leq 1$. To show this, first verify that (3.62) satisfies the conditions of the existence-uniqueness theorem (Theorem 2.4.1) on the entire tP plane, so the ODE possesses a unique solution for any initial condition. Then verify that the solution to (3.62) with initial condition $P(0) = 0$ is $P(t) = 0$ for all $t > 0$. Verify that the solution with $P(0) = 1$ is $P(t) = 1$ for all $t > 0$. Why does this imply that any solution with $0 < P(0) < 1$ will satisfy $0 < P(t) < 1$ for all $t > 0$?

A slightly messy separation of variables can be used to solve (3.62). With a bit of simplifying algebra and with the assumptions that A, b , and P all lie in the interval $(0, 1)$, we obtain the solution

$$P(t) = \frac{P_0}{P_0 + e^{-\frac{A(b^t-1)}{\ln(b)}}(1-P_0)} \quad (3.63)$$

to (3.62) with $P(0) = P_0$.

Modeling Exercise 5.3.4 Verify that $P(t)$ as given by (3.63) satisfies (3.62) with $P(0) = P_0$.

Age	20	25	30	35	40	45	50
1940-44	21.1	66.1	83.1	88.8	91.2	92.7	94
1945-49	22.3	65.5	80.1	86.1	89.3	91.3	92.5

Table 3.11: Cumulative Marriage Rates for Men Ever Married.

Parameter Estimation and Comparison to Data

Table 3.11 gives some data for two cohorts, US men born between 1940 and 1944, and US men born between 1945 and 1949. This data is from [70] and [73]. For each cohort at each age, the table has the percentage of men who have been married.

Let's focus on the 1940-44 cohort, and take $t = 0$ to correspond to age 20. In this case an appropriate initial condition is $P_0 = 0.211$, and note that the percentages in the table should be converted into proportions for our model. With $P(0) = P_0 = 0.211$ (3.63) becomes

$$P(t) = \frac{0.211}{0.211 + 0.789e^{-\frac{A(b^t - 1)}{\ln(b)}}}. \quad (3.64)$$

The function $P(t)$ defined by (3.64) can be used in (3.64) to form a sum of squares for fitting A and b :

$$S(A, b) = \sum_{j=1}^7 (P(5j - 5) - R_j)^2, \quad (3.65)$$

where R_j is the entry for the 1940-44 cohort for age $5j + 5$.

It can be difficult to locate a global minimum for A and b , but fortunately in this case a plot is illuminating. The left panel of Figure 3.20 shows a graph of $\ln(S(A, b))$ on the domain $0 < A < 1, 0 < b < 1$, and the right panel shows a contour plot for $\ln(S(A, b))$.

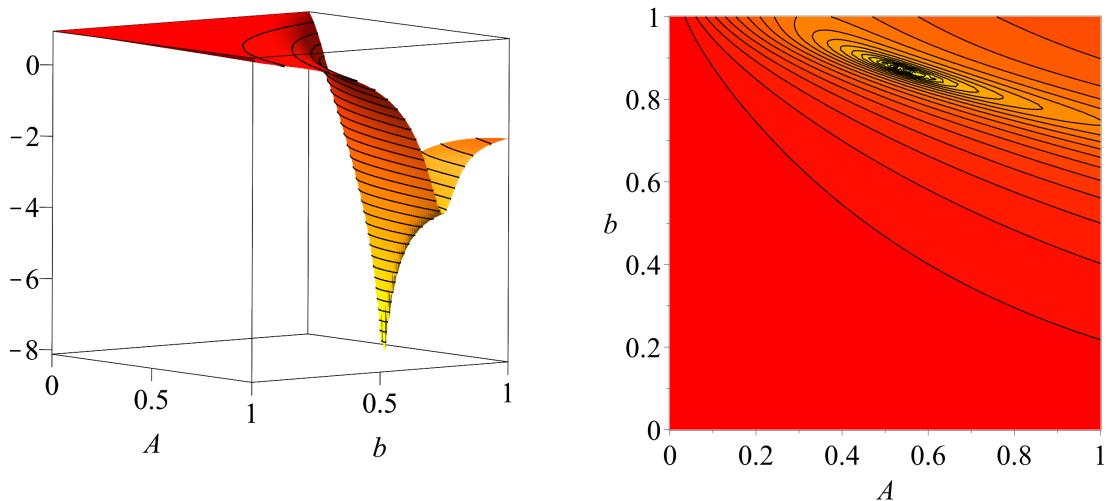


Figure 3.20: Left panel: Graph of $\ln(S(A, b))$ defined by (3.65). Right panel: contour plot for $\ln(S(A, b))$.

Modeling Exercise 5.3.5 Based on Figure 3.20 (especially the contour plot) you can make a good visual estimate of the minimizing values $A = A^*$ and $b = b^*$. Use this to find the precise values that minimize $S(A, b)$. Then use these values in (3.65) to compute the residual sum of squares and in (3.64) to plot the function $P(t)$ that best fits this data. Compare a plot of $P(t)$ to a plot of the data.

To obtain P_0 in (3.63) we used the initial data point $(0, 0.211)$. This forces the solution $P(t)$ given by (3.63) to go exactly through the data point $(0, 0.211)$. But why should this data point be treated specially? In a real sense any of the data points can be considered as the initial condition if we allow the solution to flow backward in time. It is often the case that better results can be obtained by letting P_0 float as an undetermined parameter, and obtain its optimal value as part of a minimizing process. Thus all data points are put on an equal footing.

Modeling Exercise 5.3.6 Let us estimate P_0 along with A and b . Define a new sum of squares as

$$S(A, b, P_0) = \sum_{j=1}^7 (P(5j - 5) - R_j)^2 \quad (3.66)$$

where $P(t)$ is given by (3.63) (which leaves P_0 undefined). Minimize this sum of squares as a function of P_0, A , and b . To make the computation easier, use $P_0 = 0.211$ and the previously determined values of A and b in Modeling Exercise 5.3.4 as initial guesses. Does this change the previous values much? Use these values in (3.66) to compute the residual sum of squares, and in (3.63) to plot the function $P(t)$ that best fits this data. Compare this plot to a plot of the data. Based on your parameters, what percentage of this cohort will eventually marry?

Modeling Exercise 5.3.7 Form and minimize an appropriate sum of squares, analogous to that of Reading Exercise 5.3.6, of the form

$$S(A, b, P_0) = \sum_{j=1}^7 (P(5j - 5) - R_j)^2$$

but for men in the 1945-1949 cohort. Based on your estimated parameters, what percentage of this cohort will eventually marry?

Modeling Exercise 5.3.8 Table 3.12 contains data for the 1940-44 and 1945-49 female cohorts. Form and minimize an appropriate sum of squares of the form

$$S(A, b, P_0) = \sum_{j=1}^7 (P(5j - 5) - R_j)^2$$

for each of these data sets. Based on your parameters, what percentage of each cohort will eventually marry? Are there marked difference between the men and women in these cohorts?

Age	20	25	30	35	40	45	50
1940-44	48.1	78.2	86.8	89.7	91.4	92.5	93.2
1945-49	43.1	76.9	85	88.4	90.2	91.5	92.2

Table 3.12: Cumulative Marriage Rates for Women Ever Married.

Note that the procedure we've developed is not mere curve fitting, but is based on the two basic assumptions stated above. See [70] for more data and ideas for analyzing this type of data.

3.5.4 Project: Shuttlecocks and the Akaike Information Criterion

This problem is based on the SIMIODE modeling project [123] and extends Exercise 2.2.9.

Table 2.2 in Section 2.2.6 contains data for how long it takes a shuttlecock (the projectile used in badminton) to fall a given distance when dropped; the data is from [97]. In this project the goal is to use this data to examine several different models for the force of air resistance on the shuttlecock as it falls. The models to be examined are:

- (a) No air resistance.
- (b) Air resistance proportional to speed.
- (c) Air resistance proportional to the square of speed.
- (d) Air resistance proportional to a more general quadratic function of speed.
- (e) Air resistance proportional to an r th power of speed.

We'll formulate and solve an ODE model for each possibility and compare how well each fits the data. To decide which model is best we'll invoke the Akaike information criterion.

General Models for Air Resistance

Consider an object of mass m falling straight down under the influence of gravitational force. As in Section 2.2.1, downward will be taken as the positive coordinate direction. Let $v(t)$ denote the velocity of the object, so $v > 0$ corresponds to a falling object. The force of gravity on the object is $F_g = mg$ with $g > 0$ as gravitational acceleration and m as the object's mass.

The force F_r of air resistance on the object is what is of interest here. The magnitude of this force is assumed to be a function of the object's speed $|v|$ and opposed to the direction of motion. Our interest is the case in which $v > 0$, so here $|v| = v$. We will take this resistance force as

$$F_r = -F_0(v) \quad (3.67)$$

for some function F_0 . The minus sign in front of F_0 will be used to incorporate opposition to the direction of motion. The function F_0 should be chosen so that $F_0(0) = 0$ (if the object is stationary, there is no force due to air resistance), and $F_0(v) > 0$ if $v > 0$, so $-F_0(v)$ is an upward force.

In conjunction with Newton's second law of motion (here, $ma = mv' = F_g - F_0(v)$, if gravity and air resistance are the only relevant forces), it follows that $mv'(t) = mg - F_0(v(t))$. Divide by m to obtain

$$v'(t) = g - \frac{F_0(v(t))}{m}$$

or

$$v'(t) = g - F(v(t)) \quad (3.68)$$

where $F(v) = F_0(v)/m$; the constant m has been absorbed into the definition of F .

The goal is to determine what type of function for F best models air resistance in this situation by making use of the data in Table 2.2. The choices for air resistance models listed above lead to possibilities listed in Table 3.13.

We consider each possibility in turn. The shuttlecock data was taken in Villanova, Pennsylvania, altitude 120 meters above sea level, longitude 75.3492 degrees west, latitude 40.0376 degrees north, where $g \approx 9.80136$ meters per second squared, according to the Local Gravity Calculator at <https://www.sensorsone.com/local-gravity-calculator/>. Thus gravitational acceleration g in (3.68) will be considered known.

Fitting the Air Resistance Models

Modeling Exercise 5.4.1 Solve the ODE (3.68) with $F(v) = 0$ and initial data $v(0) = 0$. Use this to show that the position of the shuttlecock would be $x(t) = gt^2/2$ (assuming $x(0) = 0$). Note that

Model	$F(v)$
(a) No air resistance	$F(v) = 0$
(b) Linear air resistance	$F(v) = kv$ for some positive constant k
(c) Pure quadratic resistance	$F(v) = kv^2$ for some positive constant k
(d) General quadratic resistance	$F(v) = k_1v + k_2v^2$ for positive constants k_1 and k_2
(e) General power law resistance	$F(v) = kv^r$ for positive constants with k and r

Table 3.13: Some models for air resistance.

there are no parameters to estimate here, since $g = 9.80136$ is known. Compute the sum of squares

$$S = \sum_{j=1}^{17} (x(t_j) - x_j)^2,$$

where (t_j, x_j) denotes the j th (time, distance) pair from Table 2.2. Also, plot $x(t)$ along with the data and comment: how well does a no air resistance model fit the data?

Modeling Exercise 5.4.2 Consider the ODE (3.68) with $F(v) = kv$ and initial data $v(0) = 0$. Show that in this case the shuttlecock position is $x(t) = \frac{g}{k^2}(kt + e^{-kt} - 1)$ (again assuming $x(0) = 0$). Here the parameter k is to be estimated (we still assume $g = 9.80136$). Form the sum of squares

$$S(k) = \sum_{j=1}^{17} (x(t_j) - x_j)^2,$$

where (t_j, x_j) denotes the j th (time, distance) pair from Table 2.2. Find that value $k = k^*$ that minimizes $S(k)$, and compute the residual. Also, plot $x(t)$ using this optimal k^* , along with the data and comment: how well does a linear air resistance model fit the data?

Modeling Exercise 5.4.3 Repeat Modeling Exercise 5.4.2 using $F(v) = kv^2$. Note that in this case the ODE (3.68) effectively becomes (2.37), but with k replacing k/m . As such, $v(t)$ is given by (2.48), but with k/m there replaced by just k here. Use this to show that (with $x(0) = 0$) we have

$$x(t) = \frac{\ln(\cosh(t\sqrt{kg}))}{k}.$$

Consulting Exercise 2.2.9 may be useful. Form the sum of squares and find the optimal value of k . Compute the residual sum of squares.

Modeling Exercise 5.4.4 In the case that $F(v) = k_1v + k_2v^2$, a slightly unpleasant separation of variables (better yet, a computer algebra system) shows that the solution to ODE (3.68) with $v(0) = 0$ is

$$v(t) = \frac{\alpha \tanh(\alpha t/2 + \operatorname{arctanh}(k_1/\alpha)) - k_1}{2k_2},$$

where $\alpha = \sqrt{k_1^2 + 4k_2g}$. Then the object position $x(t)$ can be computed as $x(t) = \int_0^t v(\tau) d\tau$, which leads to

$$x(t) = -\frac{k_1}{2k_2}t + \frac{\ln(\cosh(\alpha t/2 + \operatorname{arctanh}(k_1/\alpha)))}{k_2} + \frac{\ln(1 - k_1^2/\alpha^2)}{k_2}.$$

Form an appropriate sum of squares (with $\alpha = \sqrt{k_1^2 + 4k_2g}$ and $g = 9.80136$)

$$S(k_1, k_2) = \sum_{j=1}^{17} (x(t_j) - x_j)^2$$

and then minimize S in the variables k_1 and k_2 . Confine your attention to nonnegative values for k_1 and k_2 ; a combination of graphing and analytical computation is a good approach. Compute the residual sum of squares, and plot $x(t)$ with these optimal k_1 and k_2 values, along with the data.

The Case $F(v) = v^r$

The case in which $F(v) = v^r$ in the ODE (3.68) yields

$$v'(t) = g - kv^r(t) \quad (3.69)$$

along with $v(0) = 0$ and presents a special complication: this ODE is not analytically solvable for $v(t)$. Of course this makes finding the position $x(t)$ correspondingly difficult, and complicates the task of finding the optimal values for k and r . One straightforward approach is to proceed numerically, as follows. For a given choice of k and r :

1. Solve (3.69) numerically to compute $v(\tau_i)$ at equispaced times $\tau_i = T/N$ with $0 \leq i \leq N$, where $T = 1.873$ (the largest time for which we have position data) and N is reasonably large, e.g., $N = 1000$.
2. Use any numerical integration rule, e.g., the trapezoidal rule, along with the solution values $v(\tau_i)$ from step 1, to compute $x(t_j)$, where the t_j are the times at which we have distance data in Table 2.2.

The above steps allow us to compute the residual sum of squares $S(k, r)$ for any choice of k and r . We must then find the optimal values for k and r without the use of derivatives, since we have no obvious way to compute $\frac{\partial S}{\partial k}$ and $\frac{\partial S}{\partial r}$.

Though conceptually straightforward, this computation involves a fair amount of programming. See the appropriate scripts for Maple, Matlab, Mathematica, and Sage at the book website [8]. For the sake of brevity, the resulting optimal estimates are $k = 0.205$ and $r = 2.02$. The residual sum of squares is $S(0.205, 2.02) \approx 1.01 \times 10^{-2}$. A plot of the solution $x(t)$ stemming from $v' = g - kv^r$ with the optimal parameters superimposed on the data from Table 2.2 is shown in Figure 3.21.

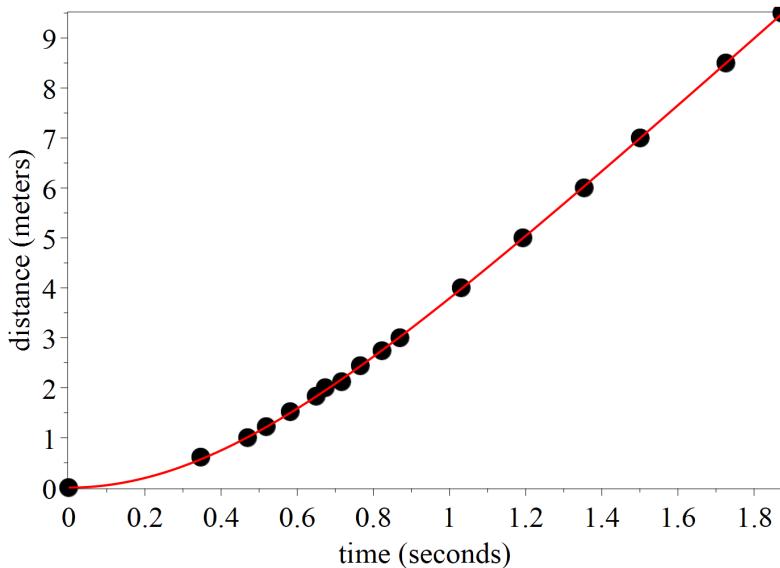


Figure 3.21: Estimated position $x(t)$ from model $v' = g - kv^r$ with optimal values $k \approx 0.205$ and $r \approx 2.02$ (solid red curve) and data from Table 2.2 (black dots).

Modeling Exercise 5.4.5 Complete Table 3.14 by computing the residual sum of squares for each air resistance model in Table 3.13. The results for model (e), $F(v) = kv^r$, are already filled in.

Model	$F(v) = 0$	$F(v) = kv$	$F(v) = kv^2$	$F(v) = k_1v + k_2v^2$	$F(v) = kv^r$
Residual sum of squares					1.01×10^{-2}

Table 3.14: Residual sum of squares for each air resistance model for a falling shuttlecock.

Modeling Exercise 5.4.6 We can perform a quick sanity check on our computations. Let RSS_a denote the residual sum of squares for model (a) in Table 3.13, the no air resistance model. Let RSS_b to RSS_e denote the residual sum of squares for the other models (b) to (e), respectively. Make sure the residual sum of squares for the models (a) to (e) in Table 3.13 satisfy each of the following inequalities:

$$RSS_d \leq RSS_b \leq RSS_a$$

$$RSS_d \leq RSS_c \leq RSS_a$$

$$RSS_e \leq RSS_b$$

$$RSS_e \leq RSS_c$$

Explain why these relations should be expected. Hint: the first you can obtain with two parameters and $F(v) = k_1v + k_2v^2$ ought to be at least as good as what you can obtain with just one parameter in $F(v) = kv^2$ or $F(v) = kv$.

Model Selection and Akaike Information Criterion

Which model is the best of the five considered? A glance at the model $F(v) = 0$, specifically the resulting graph of $x(t)$ superimposed on the data, should convince you that this model is not worth further consideration, at least not in competition with the others. It has a much, much larger residual sum of squares, corresponding to a poor fit to the data.

The model $F(v) = kv$ is much superior to $F(v) = 0$ and has a much smaller residual sum of squares, but again, an examination of $x(t)$ and the data in comparison to the other models and their residual sum of squares (the next largest of which is still 50 times smaller) takes this model out of the running.

But the other three models have residual sum of squares that are all comparable in size, and in each case a graph of $x(t)$ overlayed on the data shows almost perfect agreement. On what basis might we choose one model over another? One approach is known as the **Akaike information criterion** (AIC). Each model above contains a certain number P of explicit undefined parameters: $P = 0$ parameters in the case $F(v) = 0$, $P = 1$ parameter when $F(v) = kv$ and $F(v) = kv^2$, and $P = 2$ parameters when $F(v) = k_1v + k_2v^2$ and $F(v) = kv^r$. The **AIC figure of merit** for a model with P parameters used to fit N data points is

$$AIC = 2(P+1) + \frac{2P(P+1)}{N-P-1} + N \ln(RSS/N) \quad (3.70)$$

where RSS denotes the residual sum of squares after the least-squares parameter values have been found. Equation (3.70) is the form that the AIC takes when the residuals (the quantities $x(t_j) - x_j$ used in computing the RSS, where $x(t)$ is the predicted value in the best-fit model) are normally distributed independent random variables. This is an assumption that should be tested, although we have not done that here. If the residuals have a different statistical distribution then the AIC figure of merit will take a different form. The idea is that a smaller value of AIC indicates a superior model. Notice that an increase in P increases AIC and penalizes a model with more parameters, while a decrease in RSS toward zero decreases the value of AIC, as this indicates a better fit to the data. When presented with a number of different models from which to choose, the value of the

AIC figure of merit is one way to select a model that does the best job of explaining the data with the fewest parameters.

The AIC was developed by Japanese statistician Hirotugu Akaike in the early 1970s and is based on ideas from information theory. Like all statistical procedures, its validity and interpretation require that certain assumptions are met. An examination of these assumptions would take us too far afield here, but they are reasonable in this setting for this data set. Additional information and examples for this topic can be found in [79], and the original paper by Akaike is [15].

Modeling Exercise 5.4.7 Use the data from Table 3.14 to compute the AIC figure for the models $F(v) = kv^2$, $F(v) = k_1v + k_2v^2$, and $F(v) = kv^r$. Models (c), (d), and (e) should turn out to have the same residual sum of squares, yet the AIC for model (c) is lower, indicating it is the preferred model. Why?

4. Second-Order Equations

In this chapter we examine second-order ordinary differential equations. These types of equations govern vibration and many other periodic phenomena. The mathematics involved is essential to modeling and understanding many mechanical, electrical, and other types of physical phenomena.

4.1 Vibration and the Harmonic Oscillator

4.1.1 The 2010 Chilean Earthquake

On February 27, 2010 at 3:34 a.m., an earthquake of magnitude 8.8 struck Chile, one of the strongest quakes ever recorded. The shaking lasted for more than two minutes and when it was over at least 500 people had been killed. Damage was estimated to be in excess of \$30,000,000,000. Although it was little consolation to the loved ones of those who died, this death toll was considered to be relatively low, given the size of the quake and the population of the country. In contrast, a much less powerful magnitude 7.0 earthquake struck Haiti the previous month, on January 10, 2010, and yet is estimated to have killed between 100,000 and 250,000 people.

One of the primary factors for the lower death toll in the Chilean earthquake are the strict building codes that exist in that country and their stringent enforcement: see [109]. These building codes were adopted after a massive quake of magnitude 9.5 there in 1960, the strongest earthquake ever recorded. For ten years after a building's construction, builders in Chile are held liable for any damages that result from not adhering to these codes. The high death toll in Haiti has been attributed in large part to the poor construction of many buildings and lack of an enforced code that mandates structures be resistant to earthquakes. To see the difference modern engineering can make during an earthquake, see the informative video at [5].

A detailed understanding of how mechanical objects vibrate is at the heart of engineering structures that can endure this kind of abuse. Such an understanding allows engineers to design buildings that can respond to earthquakes without collapsing or endangering occupants. Much of the relevant mathematics falls into the realm of second-order differential equations. This mathematics governs not only mechanical vibration, but many electromagnetic and other physical phenomena.

4.1.2 The Harmonic Oscillator

Springs, Dashpots, and Masses

Consider the left panel in Figure 4.1 (adapted from [83]) in which a highly simplified two-dimensional model of a building is presented. It consists of a single story, with the roof depicted as a point mass m supported by vertical walls. When the mass is displaced from its equilibrium position (it is swaying to the right in the left panel of Figure 4.1), the walls exert a force that opposes this displacement, and the magnitude of the force is proportional to the magnitude of the displacement. A simplified abstraction of the situation is shown in the right panel of Figure 4.1, in which the roof mass is the cart on wheels and the walls' restoring force is embodied by the spring.

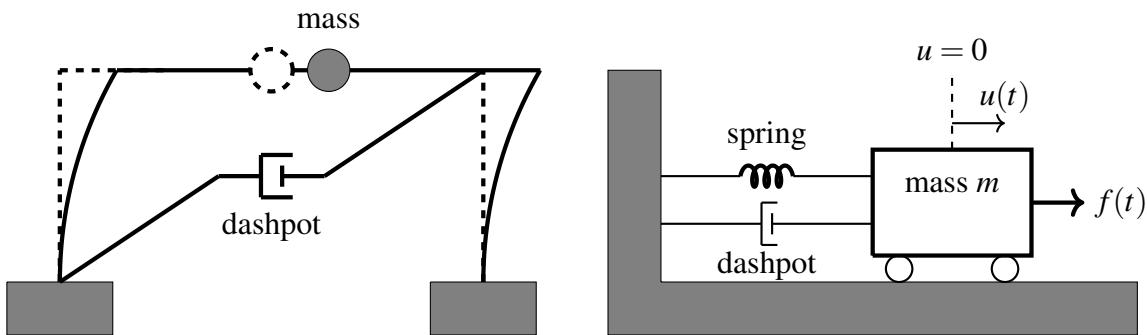


Figure 4.1: Left panel: a simplified one-story building. Right panel: a mass-spring-damper abstraction of such a building.

Motion of the roof is also opposed by frictional forces with magnitude in proportion to the speed of the mass; this is embodied in the right panel by the **dashpot** (or **damper**). The dashpot here need not be an actual device, but rather it is a hypothetical entity that represents the frictional forces present. An earthquake might be modeled as an additional externally applied force $f(t)$ on the mass. Our interest here is in the lateral back and forth motion of the mass m as a function of time.

Let $u(t)$ denote the horizontal displacement at time t of the mass in the right panel from its **equilibrium position**. This is the position in which the spring (attached to the vertical wall) exerts no force on the cart, so the spring is at its natural or rest length. Take $u > 0$ to indicate displacement to the right. We will assume the spring and dashpot have negligible mass.

The assumption that the spring exerts a force in proportion to and opposed to the displacement u is known as **Hooke's law**. Hooke first published his law in 1676 as a Latin anagram, and later explicitly as *ut tensio, sic vis* ("as the extension, the force"). Hooke's law is quantified as

$$F_{\text{spring}} = -ku \quad (4.1)$$

for some nonnegative constant k , known as the **spring constant**. The minus sign indicates the force is opposed to the displacement (recall Modeling Tip 1.1.1). Larger values for k quantify stiffer springs. The constant k has the physical dimension of force per length ($[k] = MT^{-2}$), with units of newtons per meter in SI units. Hooke's law is accurate for modest elongations of the spring, so-called **elastic deformation**. If the spring is stretched too far it undergoes **plastic deformation**, which alters the spring's mechanical properties and damages the spring. Hooke's law is not valid in that case.

The dashpot acts as a frictional or damping force of the form

$$F_{\text{damping}} = -cu' \quad (4.2)$$

for some nonnegative constant c , so the damping force is always opposed to the direction of motion. The linear relation between force and velocity in (4.2) is known as **viscous damping**. The constant

c has dimensions of force per velocity ($[c] = MT^{-1}$), or units of newtons per meter per second in SI units. It may also be the case that an additional external time-dependent force $f(t)$ acts on the mass, as indicated in the right panel of Figure 4.1.

Reading Exercise 4.1.1 Use Newton's second law of motion, $F = ma$, along with (4.1) and (4.2), to write down a relation between m , c , k , u , u' , u'' , and $f(t)$. Here a is the horizontal acceleration of the mass m and F denotes the net horizontal force on m , the sum of the spring and dashpot forces and the external force $f(t)$.

Forced and Unforced Harmonic Oscillators

The relation you find in Reading Exercise 4.1.1 should be equivalent to

$$mu''(t) + cu'(t) + ku(t) = f(t). \quad (4.3)$$

The second-order ODE (4.3) is one of the most important and common types of ordinary differential equations for modeling mechanical and electromagnetic phenomena. It is the equation of the **forced harmonic oscillator**, or **driven harmonic oscillator**. The highest derivative of the unknown $u(t)$ that appears in (4.4) is $u''(t)$, so this ODE is second-order. The equation is also linear, since the left side of (4.3) is linear in the variable u . The equation is **constant-coefficient** since m , c , and k are constants.

A common special case is that in which there is no external force $f(t)$, so only the spring and dashpot exert force on m . In this case (4.3) becomes

$$mu''(t) + cu'(t) + ku(t) = 0. \quad (4.4)$$

The ODE (4.4) is the equation of the **unforced harmonic oscillator**. This ODE is **homogeneous**, since the right side is zero, while (4.3) is **nonhomogeneous** if $f(t)$ is not the zero function. The focus in the next section, Section 4.2, is the structure and behavior of solutions to the unforced harmonic oscillator equation (4.4). We'll consider the nonhomogeneous version (4.3) in Section 4.3.

Consider (4.4) in the case in which $c = 0$, so the system has no frictional forces or damping. The only force acting on the mass is the spring, and (4.4) then becomes

$$mu''(t) + ku(t) = 0. \quad (4.5)$$

This is the **undamped harmonic oscillator** or **pure harmonic oscillator**. In this case common sense suggests that if set in motion, the mass in the right panel of Figure 4.1 should oscillate and never stop.

Reading Exercise 4.1.2

- Consider (4.5) when $m = 1$ kg and $k = 1$ newton per meter, so the ODE is $u''(t) + u(t) = 0$. Show that in this case $u(t) = A \cos(t) + B \sin(t)$ provides a solution for any choice of A and B .
- Show that in the general case for (4.5) the function $u(t) = A \cos(\omega t) + B \sin(\omega t)$ with $\omega = \sqrt{k/m}$ provides a solution for any choice of A and B . What is the dimension of ω ?

Reading Exercise 4.1.3 Suppose that for the undamped spring-mass system of Reading Exercise 4.1.2 where $m = 1$ kg and $k = 1$ newton per meter, the mass is pulled to initial position $u(0) = 0.5$ meters and released with no initial velocity, so $u'(0) = 0$. After time $t = 0$ no external forces act on the mass, just the spring. Use the results of Reading Exercise 4.1.2 to find a function $u(t)$ that satisfies both (4.5) and these initial conditions.

Remark 4.1.1 For brevity in the examples that follow, we won't explicitly state the units on the various constants and variables, unless the problem is of an applied nature.

Reading Exercise 4.1.4 Consider the harmonic oscillator when $m = 1$, $c = 2$, and $k = 26$. Since $c > 0$ this is a **damped harmonic oscillator**. Suppose the mass is displaced to position $u(0) = 1$ and released with zero initial velocity, so $u'(0) = 0$. Verify that

$$u(t) = e^{-t} \cos(5t) + e^{-t} \sin(5t)/5$$

satisfies (4.4) in this case with the appropriate initial conditions. Graph the solution for $0 \leq t \leq 5$. Does it make intuitive sense?

4.1.3 Initial Conditions

As you might suspect after working Reading Exercises 4.1.2-4.1.4, in order to find a specific solution to (4.4), two **initial conditions** are required, typically of the form $u(0) = u_0$ and $u'(0) = v_0$. Here u_0 can be interpreted physically as the initial displacement of the mass and v_0 as the initial velocity imparted to the mass when it is released. We previously encountered the necessity of specifying both initial position and initial velocity for second-order ODEs, if not explicitly, in the Hill-Keller ODE $v'(t) = 11 - kv(t)$. This ODE was first solved for $v(t)$ using an initial condition $v(t_0) = 0$, but the ultimate goal was the sprinter's position $x(t)$, which was found using $x'(t) = v(t)$ and the initial position $x(t_0) = 0$. If the Hill-Keller ODE had been cast in terms of $x(t)$ from the start it would have been a second-order ODE $x''(t) = 11 - kx'(t)$, with initial conditions $x(t_0) = 0$ and $x'(t_0) = 0$.

In Section 4.2 we'll consider how to solve (4.4) with any desired initial conditions, and in Section 4.3 we'll consider the more general equation (4.3). But for now let's look at some other physical situations that lead to equations of the form (4.4) or (4.3).

4.1.4 More Applications of Spring-Mass Models

Bicycle Shock Absorbers

■ **Example 4.1** The front shock absorber of a typical mountain bike (see Figure 4.2) may be modeled as a spring-dashpot system. A typical value for the spring constant might be $k = 15000$ newtons per meter with a damping constant $c = 1700$ newtons per meter per second; we'll explore this model more in later examples and the Modeling Projects in Section 4.6.

The mass in this system consists of the rider's mass and the bike's mass, less the mass of the wheels since they are not suspended by the shock absorber. Suppose the rider has a mass of 80 kg, the bike (less wheels) has a mass of 12 kg, and that half of this total mass is supported by the front shock absorber, so the effective supported mass is $m = (80 + 12)/2 = 46$ kg. Assume that the only other force acting on the rider is gravity. If the front wheel is in contact with the ground, it may be considered fixed or immovable, and if $u(t)$ denotes the vertical displacement of the front shock from equilibrium then (4.3) becomes

$$46u''(t) + 1700u'(t) + 15000u(t) = -450.8, \quad (4.6)$$

where $f(t) = -mg = 450.8$ with $g = 9.8$, half the weight of the bike and rider. The ODE (4.6) is nonhomogeneous and of the form of (4.3). ■

Reading Exercise 4.1.5 Suppose the rider in Example 4.1 has been pedaling on level ground for some distance, so it's reasonable to assume that $u(t)$ has settled to a constant or equilibrium value, $u(t) = u_{eq}$ for some constant u_{eq} . To find u_{eq} , substitute $u(t) = u_{eq}$ into the ODE (4.6) and use the fact that $u'(t) = u''(t) = 0$ here. How far does the shock absorber compress under the rider's weight? A typical bike front shock has a range of motion of about 140 mm before bottoming out, and it is recommended that the rider's weight alone should compress the shock 20 to 30 percent of the shock's range of motion. Is this recommendation satisfied here?



Figure 4.2: Front shock absorber on a mountain bike.

Vibration Isolation

■ **Example 4.2 Vibration isolation** is an important area of mechanical engineering. The goal of vibration isolation is to shield one part of a system from mechanical vibrations induced by another part of the system. For example, we may wish to prevent the vibrations caused by a large air conditioner from shaking the supporting floor. Another application is a **vibration isolation table**, often found in settings where sensitive equipment or experiments must be shielded from environmental disturbances. These tables are common in optical laboratories, or for supporting electron microscopes, and even for supporting patients during eye surgery. Vibration isolation can be done in many different ways, depending on the application. Some systems are active, with sensors and powered actuators to sense and counter vibrations, a scenario you can explore in the modeling project “Vibration Isolation Table Shakedown” in Section 5.7. Other techniques for vibration isolation are passive and consist of coil springs, air springs, dashpots, and rubber pads.

As an idealized example of a vibration isolation table, consider a rectangular tabletop of mass m supported on a single column. This column acts to isolate the tabletop from ground vibration. In reality the table would have more than one leg, but this simple model will illustrate the general principles. Suppose the leg acts as a spring-dashpot system, as illustrated by the cylindrical support in the left panel of Figure 4.3 or its spring-dashpot counterpart in the right panel. This leg supports the tabletop of mass m , the rectangular slab in either panel. Let c and k denote the damping and spring constants, respectively. Suppose the floor on which the leg rests moves vertically as a function of time t with displacement $d(t)$. What motion will the tabletop experience?

To determine this, consider the right panel in Figure 4.3 in which the vertical motion of the ground is depicted as the wavy curve. With y as a vertical coordinate and upward as the positive direction $y > 0$, the ground motion is described by $y = d(t)$; here $d(t)$ need not be periodic. Let L_0 denote the natural (rest) length of the spring and let $u(t)$ denote the displacement of the spring from its natural length. If $y(t)$ is the altitude of the tabletop mass m above $y = 0$ then, from Figure 4.3,

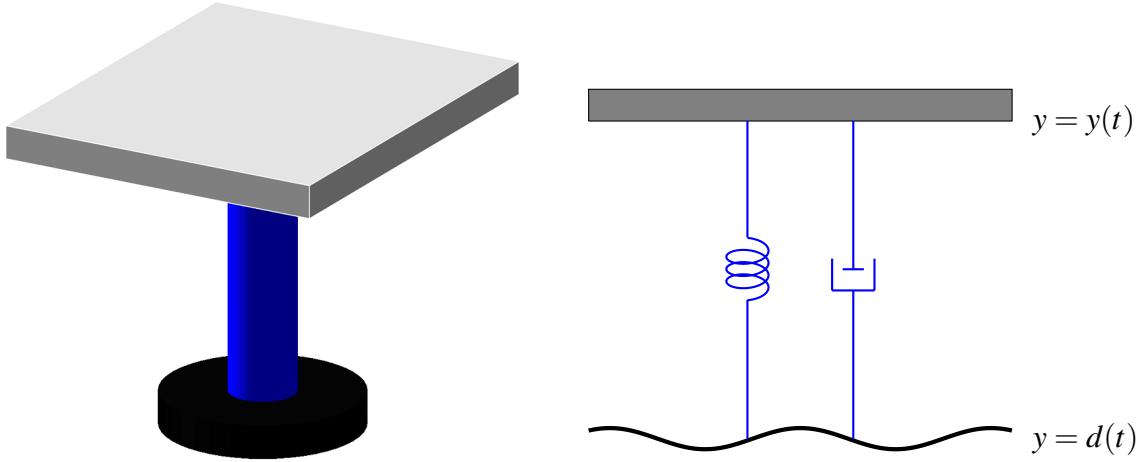


Figure 4.3: Left panel: vibration isolation table. Right panel: spring-mass-damper model of the table with possible vertical ground motion $d(t)$.

it's easy to see that the length $L(t)$ of the spring at time t is

$$L(t) = y(t) - d(t). \quad (4.7)$$

The displacement $u(t)$ of the spring from its natural length is then

$$u(t) = L(t) - L_0 = y(t) - d(t) - L_0.$$

The force of the spring on the mass m is $F_{spring} = -ku(t)$, or

$$F_{spring} = -ku(t) = -k(y(t) - d(t) - L_0).$$

Assume the damping force is proportional to the rate at which the spring-damper system is lengthening (or contracting). From (4.7) this rate is $L'(t) = y'(t) - d'(t)$, and so the corresponding force on m is

$$F_{damping} = -cL'(t) = -c(y'(t) - d'(t)).$$

The vertical acceleration of the tabletop mass is $y''(t)$. If gravity is the only other force acting on the tabletop then from Newton's second law of motion,

$$\begin{aligned} my''(t) &= F_{spring} + F_{damping} + F_{gravity} \\ &= -k(y(t) - d(t) - L_0) - c(y'(t) - d'(t)) - mg, \end{aligned} \quad (4.8)$$

with $g > 0$ (hence the explicit minus sign in front of mg on the right in (4.8)). The ODE (4.8) can be rearranged to

$$my''(t) + cy'(t) + ky(t) = k(d(t) + L_0) + cd'(t) - mg, \quad (4.9)$$

a linear, constant-coefficient, nonhomogeneous, second-order ODE for $y(t)$, the vertical displacement of the tabletop. Note that all terms on the right side of (4.9) are known or given. This equation fits the mold of (4.3). ■

Reading Exercise 4.1.6 Suppose $d(t) = 0$ (the ground is motionless). Find an equilibrium solution $y(t) = y_{eq}$ to (4.9). What is the physical interpretation of this solution?

RLC Circuits

■ **Example 4.3** Figure 4.4 shows a single-loop resistor-inductor-capacitor (RLC) circuit that contains a voltage source $V(t)$, a resistor R , an inductor L , and a capacitor C . The situation is similar to that of Example 2.5, but now there is an inductor in the loop. Appendix C contains a more detailed explanation of the basic laws that govern these types of circuits. For an ideal inductor the current-voltage relationship is

$$V(t) = LI'(t),$$

where $V(t)$ is the voltage across it and $I(t)$ is the current through the inductor. The constant L is the **inductance** of the inductor. In the SI system the unit of inductance is the **henry**.

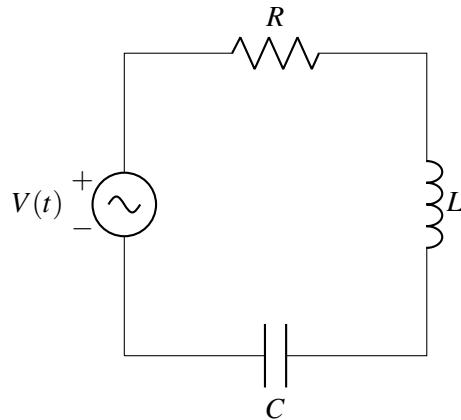


Figure 4.4: Single-loop RLC series circuit.

We can use reasoning similar to that of Example 2.5 to find an ODE that governs the charge $q(t)$ on the capacitor at any time, and from this we can find $I(t)$, the current through the loop and the voltage across any component. With the same conventions as in Example 2.5, start at the negative side of the voltage source and step around the RLC loop. The voltage rise over the source is $V(t)$, the voltage drop across the resistor is $RI(t)$, the voltage drop across the inductor is $LI'(t)$, and the voltage drop across the capacitor is $q(t)/C$, at which point we have returned to the minus, or ground, side of the source. From Kirchhoff's voltage law these voltage changes must sum to zero so that

$$V(t) - RI(t) - LI'(t) - q(t)/C = 0. \quad (4.10)$$

The changing charge on the positive capacitor plate is due to the charge entering through the wire, and so $I(t) = q'(t)$. From this it follows that $I'(t) = q''(t)$, so that (4.10) can be written as

$$Lq''(t) + Rq'(t) + q(t)/C = V(t). \quad (4.11)$$

This is a linear, second-order, nonhomogeneous ODE for $q(t)$, the charge on the capacitor. Typical initial conditions might be $q(0) = q_0$ and $I(0) = q'(0) = I_0$ (often $q(0) = 0$ and $I(0) = 0$, such as when a switch in the circuit is closed at time $t = 0$).

Note that equation (4.11) and the spring-mass-damper equation (4.3) are effectively *identical*, despite the completely different physical situations that each models. In a circuit the voltage source $V(t)$ is a bit like a force in the mechanical system; inductors act like masses that oppose changes in current, the resistor R is like viscous damping, and the capacitor acts like a spring. However, capacitance C is comparable to $1/k$, the reciprocal of the spring constant. For a spring the quantity $1/k$ is called the **compliance** of the spring. ■

Reading Exercise 4.1.7 Consider (4.11) in the case that $R = 0$ and $V(t) = \cos(\omega t)$. Verify that

$$q(t) = \frac{C}{1 - LC\omega^2} \cos(\omega t)$$

is a solution to (4.11). Note that $q(t)$ is periodic and has the same frequency as $V(t)$. How does the amplitude of $q(t)$ behave as ω approaches $1/\sqrt{LC}$?

4.1.5 Exercises

Exercise 4.1.1 Consider a mass m acted on by two springs and dashpots as shown in Figure 4.5. Assume the spring and dashpot on the left have constants k_1 and c_1 , respectively, and those on the right have constants k_2 and c_2 . Assume that the displacement $u = 0$ corresponds to a point at which both springs exert no force on the mass. Assume also that no other forces act on the mass.

Use Newton's second law of motion to find a second-order, linear, homogeneous, constant-coefficient ODE satisfied by $u(t)$, the displacement of the mass from its equilibrium position.

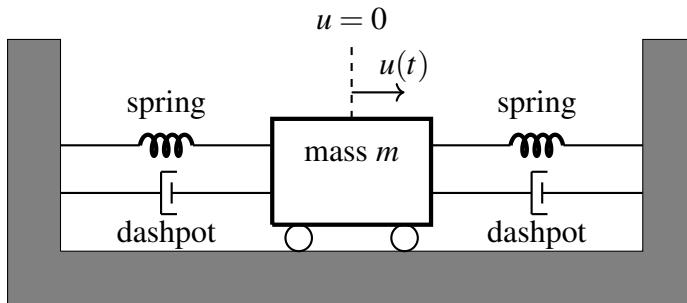


Figure 4.5: Mass acted on by two springs and dashpots.

Exercise 4.1.2 Viscous damping (4.2) is only one model for the frictional forces experienced by mechanical systems. In this exercise we explore several other common models. Let us first define the function

$$\operatorname{sgn}(z) = \begin{cases} 1, & z > 0 \\ -1, & z < 0 \\ 0, & z = 0 \end{cases}$$

that returns the sign of its argument. Consider a spring-mass-damper system with mass m and spring constant k , in which $u(t)$ denotes the displacement of the mass from the equilibrium position of the spring.

- (a) Instead of viscous damping we may consider a model in which the force of friction opposes the motion of the mass in proportion to the square of the mass's speed. Verify that the choice $F_{\text{friction}} = -c \operatorname{sgn}(u'(t))(u'(t))^2$ satisfies this condition by considering both possibilities $u'(t) > 0$ and $u'(t) < 0$. Then follow the modeling that led to (4.3) to derive a second-order ODE that governs the spring-mass-damper system with forcing function $f(t)$.
- (b) Another alternative damping model is **Coulomb damping**. In this model the force of friction has constant magnitude F , but is always opposed to the direction of the mass motion (unless the mass is motionless, in which case the frictional force is zero). The

constant F may depend on many factors, but we'll take it to be some unspecified constant. Formulate an appropriate model for F_{friction} and use it to write down a corresponding second-order ODE for $u(t)$. Hint: F_{friction} depends on u' and involves the sgn function.

Exercise 4.1.3 Consider a building modeled as a simple spring-mass-damper system as in Figure 4.1. Let us suppose that the building's mass (mostly the supported roof) is $m = 5000$ kg. The walls exert a restoring force modeled by a spring constant $k = 5 \times 10^5$ newtons per meter, and a damping constant $c = 2 \times 10^4$ newtons per meter per second.

- Write out the appropriate ODE to model this building, with $u(t)$ as the horizontal displacement of the roof mass.
- Suppose the building is perturbed to initial position $u(0) = 0.01$ meters with zero initial velocity. Verify that in this case $u(t)$ is given by

$$u(t) = \frac{\sqrt{6}e^{-2t}}{1200} \sin(4\sqrt{6}t) + \frac{e^{-2t}}{100} \cos(4\sqrt{6}t).$$

Plot $u(t)$ on the range $0 \leq t \leq 5$ seconds.

- What is the period of the building's (damped) motion? According to [16], a typical period for a building's oscillation is in the range 0.1 to 2 seconds.
- Compute and plot $u''(t)$, the building's acceleration. Where is it at a maximum? How many g 's of acceleration is this? (Use $g = 9.8$ meters per second squared.) According to [16], poorly constructed buildings may experience damage from accelerations of only 0.1 g .
- Suppose the structure is undamped, so $c = 0$ (but with the same m and k). Write out the appropriate ODE and find a solution of the form $u(t) = u_0 \cos(\omega t)$ with initial conditions $u(0) = 0.01$ and $u'(0) = 0$, by adjusting u_0 and ω . Plot the solution on $0 \leq t \leq 10$.

Exercise 4.1.4 Consider a cylindrical buoy floating in the water (assume the water has a calm, flat surface), as depicted in Figure 4.6. In the figure the water surface is indicated by the light blue horizontal plane. We use $y(t)$ to indicate the position of the bottom of the buoy relative to the water surface at time t , with $y < 0$ as the downward direction (so $y(t) < 0$ when the bottom of the buoy is submerged). Our goal in this exercise is to derive a differential equation satisfied by $y(t)$. We'll use Newton's second law of motion along with an accounting of the net force acting on the buoy. We assume the buoy maintains a constant vertical orientation and never pops out of the water, nor is it ever fully submerged.

Assume the only forces acting on the buoy are gravity and the buoyant force of the water (no friction). We use m to denote the total mass of the buoy, A to denote the buoy's cross-sectional area, ρ for the density of water, and $g > 0$ for gravitational acceleration.

- According to Archimedes' principle, the upward buoyant force of the water on the buoy equals the weight of the water that is displaced by the submerged portion of the buoy. If the bottom of the buoy is submerged at position $y(t) < 0$, what is the volume of water displaced? What is the mass of the water displaced? What is the weight of the water displaced? This is the upward buoyant force, F_{buoyancy} . Hint: it involves m, g, ρ, A and $y(t)$. Given that $y(t) < 0$, make sure this force is upward.
- What is F_{gravity} , the force due to gravity on the buoy? Make sure this force is downward (and recall we are using $g > 0$).

- (c) Use Newton's second law of motion to show that $y(t)$ satisfies the constant-coefficient, nonhomogeneous, second-order ODE

$$y''(t) + \frac{\rho g A}{m} y(t) = -g.$$

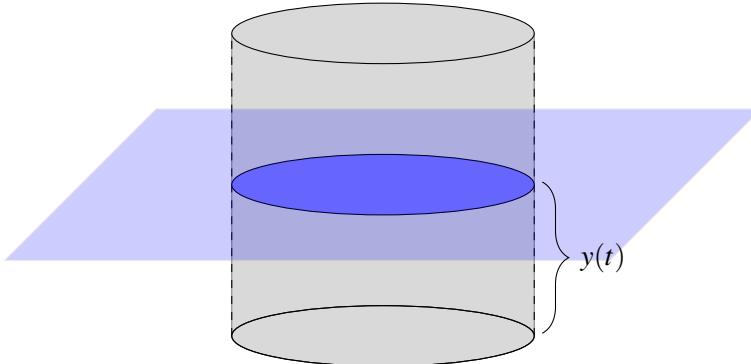


Figure 4.6: Cylindrical buoy with cross-sectional area A . The plane indicates the water surface.

Exercise 4.1.5 Consider an RLC circuit with $L = 10^{-3}$ henries, $R = 10$ ohms, $C = 10^{-4}$ farads, and voltage source $V(t) = 3$ volts. Let $q(t)$ denote the charge on the capacitor. Find an equilibrium solution $q(t) = q^*$ (q^* a constant) to the ODE (4.11). What is the resulting current in the circuit for this equilibrium solution?

Exercise 4.1.6 Sometimes one can glean information from an ODE even without solving it.

Newton's universal law of gravitation states that the force of gravitational attraction between point masses m and M is given by

$$F(r) = \frac{GMm}{r^2},$$

where r is the distance between the objects and $G \approx 6.674 \times 10^{-11}$ (SI units) is the gravitational constant. This law also applies to spherical masses of uniform density, where r is the distance between the bodies' centers of mass.

Let M denote the mass of the earth, $M \approx 5.972 \times 10^{24}$ kg, and let $m \ll M$ be the mass of some object, like a satellite. Suppose this object moves on a trajectory that is purely radial relative to the earth. That is, the object's position is purely a function of r , where r is the distance from the center of the earth to the center of the object.

- (a) Suppose an object is launched from the earth's surface, $r = R \approx 6.37 \times 10^6$ meters from the center of the earth, with initial radial velocity v_0 , and that it moves only radially with respect to the earth. Let $r(t)$ denote the object's distance from the center of the earth. Use Newton's law of gravitation and Newton's second law to show that if the force of the earth's gravity is the only force acting on the object, then $r(t)$ satisfies

$$r''(t) = -\frac{GM}{r^2(t)}. \quad (4.12)$$

What are the initial conditions for this second-order ODE?

- (b) Despite its apparent simplicity, the ODE (4.12) is not solvable in any simple analytical form. Try some numerical experiments: solve the ODE numerically with whatever software you have available, with $v_0 = 100, 1000, 10000$, and 100000 meters per second. Do the solutions behave as you expect?
- (c) You should find in (b) that for small initial velocities, the object falls back to earth, but for sufficiently large initial velocities the object escapes to infinity. We can compute this escape velocity by using conservation of energy, which requires that the total energy of the object, kinetic plus potential, remains constant.

To make use of the conservation of energy, first multiply both sides of (4.12) by $r'(t)$ and integrate from $t = 0$ to $t = T$ (here T is some unspecified time with $T > 0$.) Show that this yields

$$\frac{1}{2}((r'(T))^2 - (r'(0))^2) = GM \left(\frac{1}{r(T)} - \frac{1}{R} \right).$$

Then multiply through by the object's mass m and find

$$\frac{m(r'(T))^2}{2} + GMm \left(\frac{1}{R} - \frac{1}{r(T)} \right) = \frac{mv_0^2}{2}. \quad (4.13)$$

Let us take $r = R$ as our normalization for zero potential energy (when the object is at the earth's surface) and note that $m(r'(0))^2/2$ or $mv_0^2/2$ is the object's kinetic energy. Show that the left side of (4.13) can be interpreted as the total energy of the object at time $t = T$, kinetic energy plus potential energy.

- (d) Suppose the object is launched with initial speed v_0 in such a way that $r'(T) \rightarrow 0$ and $r(T) \rightarrow \infty$ as $T \rightarrow \infty$ (thus the object has just enough velocity at launch to escape earth's gravity, but no more). Use this to find the escape velocity v_0 in terms of G, M , and R . Show that the formula is dimensionally consistent.

4.2 The Harmonic Oscillator

In this section we'll examine how to solve the homogeneous harmonic oscillator ODE (4.4). We'll look at the effect of damping on the solution both mathematically and physically, and we'll gain insight into the behavior of systems governed by (4.4). We start with some concrete examples.

4.2.1 Solving the Harmonic Oscillator ODE: Examples

An Overdamped Example

The following example illustrates how the solution process works in most cases. Consider a spring-mass system with mass $m = 1$, spring constant $k = 3$, and damping constant $c = 4$. In this case (4.4) becomes

$$u''(t) + 4u'(t) + 3u(t) = 0. \quad (4.14)$$

This is an **overdamped** system, a term that will be made precise shortly. Think of a spring-mass system immersed in a very heavy, viscous oil, and consider what a solution with initial position $u(0) = 1$ and initial velocity $u'(0) = 0$ would look like. A solution might appear as the function graphed in Figure 4.7, in which the mass slowly oozes back to its equilibrium position at $u = 0$ without oscillating. The graph looks a bit like that of a decaying exponential, especially when t is large. This is in contrast to the results obtained in Reading Exercises 4.1.2-4.1.4, where solutions to (4.4) were composed of sines and cosines or the product of an exponential with sines and cosines. Is there a common algebraic structure to all these solutions?

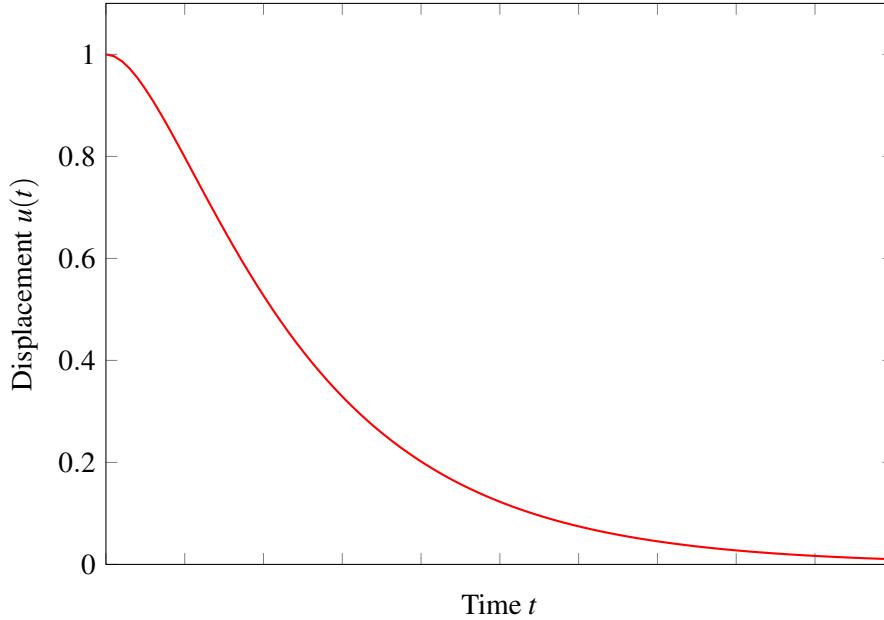


Figure 4.7: Motion $u(t)$ of a heavily damped spring-mass system.

Sines and cosines are really exponentials in disguise. If you've studied complex arithmetic you've seen this fact in the form of **Euler's formula**, which states that $e^{i\theta} = \cos(\theta) + i\sin(\theta)$. If not, don't worry, the material is presented in Appendix A. In any case, the above discussion strongly suggests that solutions to (4.4) might be obtained in the form $u(t) = e^{rt}$ for a proper choice of r .

Guessing Solutions: The Characteristic Equation

For the specific equation (4.14) we will look for a solution of the form $u(t) = e^{rt}$ for an appropriate choice of r by substituting $u(t) = e^{rt}$ into (4.14) and seeing what happens. The basic rules for differentiation yield:

$$\begin{aligned} u(t) &= e^{rt}, \\ u'(t) &= re^{rt}, \\ u''(t) &= r^2e^{rt}. \end{aligned}$$

Insert this information into (4.14) and collect terms to obtain

$$e^{rt}(r^2 + 4r + 3) = 0, \quad (4.15)$$

which must be satisfied identically in t , that is, the left side must be the zero function. Note that e^{rt} is never 0, so we can divide both sides of (4.15) by e^{rt} to obtain

$$r^2 + 4r + 3 = 0. \quad (4.16)$$

Equations (4.15) and (4.16) have exactly the same solutions for r . Thus a necessary and sufficient condition for $u(t) = e^{rt}$ to satisfy (4.14) is that r satisfy equation (4.16), a quadratic equation.

Equation (4.16) is called the **characteristic equation** or **auxiliary equation** for (4.14). The roots to (4.16) are easily found from the quadratic formula or by factoring as $r^2 + 4r + 3 = (r+1)(r+3) = 0$, and are $r = -1$ and $r = -3$. As a result each of the functions

$$u(t) = e^{-t} \quad \text{and} \quad u(t) = e^{-3t} \quad (4.17)$$

provides a solution to (4.14).

Remark 4.2.1 In the analysis above we guessed that there might be solutions to (4.14) of the very specific form $u(t) = e^{rt}$, then successfully adjusted r to make our guess work. This is a very common approach in mathematics, physics, and engineering. More generally, when confronted with an ODE that we don't know how to solve, we use mathematical intuition, physical intuition, and sometimes just plain desperation to make an educated guess at what solutions might look like, try them in the ODE, then adjust as necessary. Such an educated guess is often referred to as an **ansatz**.

Reading Exercise 4.2.1 Look up the definition of the word “ansatz”. What is the literal translation from German?

Linearity and Superposition

From (4.17) we see that there are at least two distinct solutions to (4.14). These solutions can be used to construct infinitely many more solutions; the key is to use the linearity of (4.14), an indispensable asset here. Specifically, if $u_1(t)$ and $u_2(t)$ are solutions to (4.14) then any linear combination of the form

$$u(t) = c_1 u_1(t) + c_2 u_2(t) \quad (4.18)$$

is also a solution. This is easy to verify: with u as in (4.18), a bit of algebra shows that

$$\begin{aligned} u'' + 4u' + 3u &= (c_1 u_1 + c_2 u_2)'' + 4(c_1 u_1 + c_2 u_2)' + 3(c_1 u_1 + c_2 u_2) \\ &= c_1 \underbrace{(u_1'' + 4u_1' + 3u_1)}_0 + c_2 \underbrace{(u_2'' + 4u_2' + 3u_2)}_0 \\ &= 0, \end{aligned} \quad (4.19)$$

so $u(t)$ also satisfies (4.14), for any choice of c_1 and c_2 . The computation leading to (4.19) is an example of the **principle of superposition** or simply, **superposition**, in which arbitrary linear combinations of solutions to a linear ODE are again a solution. The principle of superposition is one of the most fundamental tools available for analyzing linear ODEs.

Constructing A General Solution and Obtaining Initial Data

With $u_1(t) = e^{-t}$ and $u_2(t) = e^{-3t}$ from (4.17), the principle of superposition shows that any function $u(t)$ of the form

$$u(t) = c_1 e^{-t} + c_2 e^{-3t} \quad (4.20)$$

satisfies (4.14), for any choice of constants c_1 and c_2 . The function $u(t)$ defined by (4.20) is called a **general solution** to (4.14). We say “a” general solution rather than “the” general solution, since a general solution may assume different forms; recall Remark 1.4.1.

The function $u(t)$ in (4.20) is called a general solution for good reason: any solution to (4.14) is uniquely determined by initial data $u(0) = u_0$ and $u'(0) = v_0$ (more generally, $u(t_0) = u_0$ and $u'(t_0) = v_0$), and any such initial data can be obtained from (4.20) by choosing c_1 and c_2 appropriately. To illustrate, note that $u(0) = u_0$ and $u'(0) = v_0$ yield equations

$$\begin{aligned} u(0) &= c_1 e^{-0} + c_2 e^{-3 \cdot 0} = c_1 + c_2 = u_0, \\ u'(0) &= -c_1 e^{-0} - 3c_2 e^{-3 \cdot 0} = -c_1 - 3c_2 = v_0. \end{aligned}$$

For any choice of u_0 and v_0 , the two equations $c_1 + c_2 = u_0$ and $-c_1 - 3c_2 = v_0$ can be solved uniquely for c_1 and c_2 , as $c_1 = (3u_0 + v_0)/2$ and $c_2 = -(u_0 + v_0)/2$.

■ **Example 4.4** Let us find the solution to (4.14) with initial conditions $u(0) = 1$ and $u'(0) = 0$. With a general solution of the form (4.20), $u(0) = 1$ implies that $c_1 + c_2 = 1$, while $u'(0) = 0$ implies that $-c_1 - 3c_2 = 0$. The solution to these two equations is $c_1 = 3/2$ and $c_2 = -1/2$, and so the solution to (4.14) with the desired initial conditions is

$$u(t) = \frac{3}{2}e^{-t} - \frac{1}{2}e^{-3t}.$$

This is the function graphed in Figure 4.7, on the interval $0 \leq t \leq 5$ (although we left the labels off the time axis in that figure). ■

Reading Exercise 4.2.2 Find the solution to (4.14) with initial conditions $u(0) = 2$ and $u'(0) = 4$. Plot the solution on the interval $0 \leq t \leq 5$.

■ **Example 4.5** A building in the configuration depicted in Figure 4.1 has roof mass $m = 10^4$ kg, damping $c = 50000$ newton-seconds per meter, and spring constant $k = 40000$ newtons per meter. Let's compute the displacement of the roof mass if a sudden shock, impact, or wind gust sets the roof in motion with initial velocity $u'(0) = 0.25$ meters per second and initial displacement $u(0) = 0$. The governing ODE is

$$10000u''(t) + 50000u'(t) + 40000u(t) = 0, \quad (4.21)$$

with initial conditions $u(0) = 0$ and $u'(0) = 1/4$.

To begin, seek solutions of the form $u(t) = e^{rt}$. Substituting this ansatz into the ODE and dividing through by e^{rt} yields the characteristic equation

$$10000r^2 + 50000r + 40000 = 0.$$

Note the easy correspondence between the coefficients of the characteristic equation of (4.21) with those of (4.21) itself. The characteristic equation factors as $10000(r+1)(r+4) = 0$, so the roots are $r = -1$ and $r = -4$. Thus both e^{-t} and e^{-4t} are solutions to the ODE. The same argument that led to (4.19) (the linearity of the ODE and superposition) shows that anything of the form

$$u(t) = c_1e^{-t} + c_2e^{-4t}$$

also satisfies $10000u''(t) + 50000u'(t) + 40000u(t) = 0$; this is a general solution.

The condition $u(0) = 0$ requires $c_1 + c_2 = 0$, while $u'(t) = -2c_1e^{-2t} - 4c_2e^{-4t}$ requires $u'(0) = -2c_1 - 4c_2 = 1/4$. The simultaneous solution to $c_1 + c_2 = 0$ and $-2c_1 - 4c_2 = 1/2$ is $c_1 = 1/12$ and $c_2 = -1/12$. The solution to the ODE with the desired initial conditions is then

$$u(t) = \frac{e^{-t}}{12} - \frac{e^{-4t}}{12}.$$

This function is graphed in Figure 4.8. Like the solution to $u'' + 4u' + 3u = 0$, this function decays to zero. This is the behavior of a system with a large amount of damping. ■

Reading Exercise 4.2.3 The initial conditions for an ODE need not be given at time $t = 0$, though this is common. Solve the ODE (4.14) using the general solution (4.20), but with initial conditions $u(1) = -1$ and $u'(1) = 6$.

4.2.2 Solving Second-Order Linear ODEs: The General Case

The solution procedure in the previous section illustrates the typical method for solving ODEs of the form (4.4) with initial conditions $u(0) = u_0$ and $u'(0) = v_0$. We find two solutions $u_1(t)$ and $u_2(t)$ to the ODE and then use them to construct a general solution that can be used to find a solution with specified initial data. We now define this more formally.

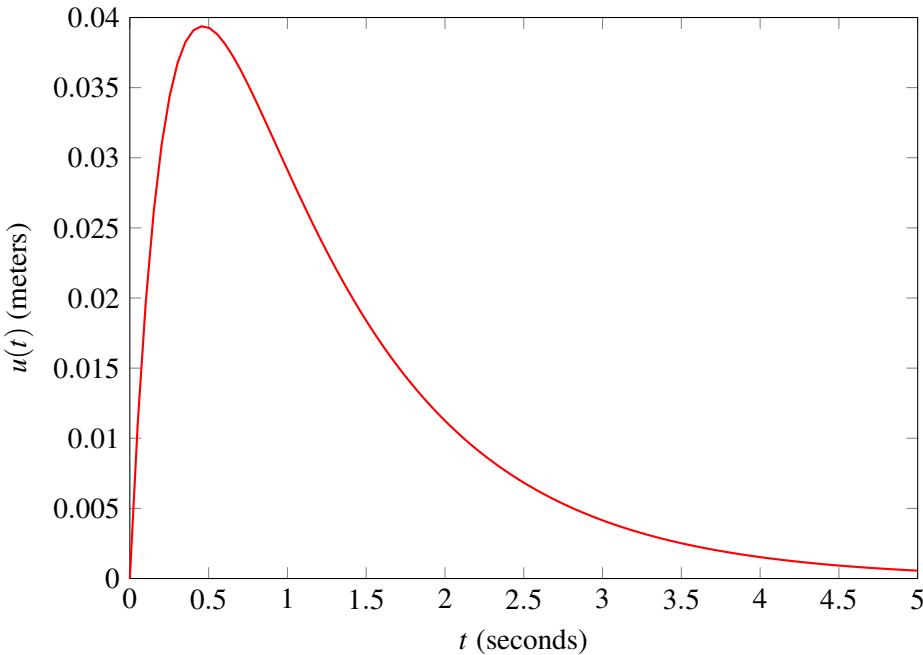


Figure 4.8: Graph of the solution to $10000u''(t) + 50000u'(t) + 40000u(t) = 0$ with $u(0) = 0$ and $u'(0) = 1/4$.

Definition 4.2.1 A **general solution** $u(t)$ to $mu''(t) + cu'(t) + ku(t) = 0$ is a function of the form

$$u(t) = c_1u_1(t) + c_2u_2(t), \quad (4.22)$$

where u_1 and u_2 both satisfy $mu''(t) + cu'(t) + ku(t) = 0$, and with the property that for any choice of t_0 , u_0 , and v_0 , the constants c_1 and c_2 in (4.22) can be adjusted so that $u(t)$ satisfies the conditions $u(t_0) = u_0$ and $u'(t_0) = v_0$.

The functions $u_1(t)$ and $u_2(t)$ in the general solution (4.22) are called **basis functions**, and u_1 and u_2 together form a **basis** for the set of solutions to the ODE, or are termed a **fundamental set of solutions** for the ODE. They are two functions out of which all solutions to the ODE can be constructed by superposition. As suggested by Reading Exercise 4.2.3, if a general solution of the form (4.22) works when initial conditions are given time $t = t_0$, then the general solution will work for any other initial time, though we will not prove this right now. It will be clear for the various ODEs we encounter, however.

Reading Exercise 4.2.4 Show that if $u_1(t)$ and $u_2(t)$ both satisfy $mu''(t) + cu'(t) + ku(t) = 0$, then so does the function $u(t) = c_1u_1(t) + c_2u_2(t)$ for any choice of c_1 and c_2 . Hint: mimic the computation in (4.19).

Reading Exercise 4.2.5 Show that if two solutions $u_1(t)$ and $u_2(t)$ to $mu''(t) + cu'(t) + ku(t) = 0$ satisfy $u_1(t) = \alpha u_2(t)$ for some constant α (so u_1 is a scalar multiple of u_2), then the pair $u_1(t)$ and $u_2(t)$ cannot be used in (4.22) to construct a general solution to the ODE, that is, they do not form a basis. Hint: show that for initial data u_0 and v_0 , the equations $c_1u_1(t_0) + c_2u_2(t_0) = u_0$ and $c_1u'_1(t_0) + c_2u'_2(t_0) = v_0$ are not solvable for c_1 and c_2 unless u_0 and v_0 happen to satisfy $u'_1(t_0)u_0 = u'_2(t_0)v_0$.

The result of Reading Exercise 4.2.5 shows that if u_1 and u_2 form a basis for the set of solutions

to $mu''(t) + cu'(t) + ku(t) = 0$ then neither function can be a scalar multiple of the other. In this case the functions u_1 and u_2 are said to be **linearly independent**.

Solution Procedure for Second-Order ODEs

To find a general solution to $mu'' + cu' + ku = 0$:

1. Find the roots r_1 and r_2 to the characteristic equation

$$mr^2 + cr + k = 0 \quad (4.23)$$

obtained from trying an ansatz $u(t) = e^{rt}$ in the ODE $mu'' + cu' + ku = 0$. From the quadratic formula these roots are, in no particular order,

$$\begin{aligned} r_1 &= -\frac{c}{2m} + \frac{\sqrt{c^2 - 4mk}}{2m} \\ r_2 &= -\frac{c}{2m} - \frac{\sqrt{c^2 - 4mk}}{2m}. \end{aligned} \quad (4.24)$$

2. If $r_1 \neq r_2$, form a general solution

$$u(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t}. \quad (4.25)$$

The function $u(t)$ is a solution to (4.4) for any choice of c_1 and c_2 , as shown in Reading Exercise 4.2.4. If $r_1 = r_2$ (so the characteristic equation has a double root, meaning $c^2 - 4mk = 0$) then (4.25) is not a general solution. We will consider this case shortly.

3. The initial conditions $u(0) = u_0$ and $u'(0) = v_0$ lead to equations

$$c_1 + c_2 = u_0$$

$$r_1 c_1 + r_2 c_2 = v_0.$$

If $r_1 \neq r_2$ the unique solution to these two equations is

$$\begin{aligned} c_1 &= \frac{v_0 - r_2 u_0}{r_1 - r_2} \\ c_2 &= \frac{r_1 u_0 - v_0}{r_1 - r_2} \end{aligned} \quad (4.26)$$

Then the solution to $mu'' + cu' + ku = 0$ with initial data $u(0) = u_0$ and $u'(0) = v_0$ is

$$u(t) = \frac{v_0 - r_2 u_0}{r_1 - r_2} e^{r_1 t} + \frac{r_1 u_0 - v_0}{r_1 - r_2} e^{r_2 t}.$$

The Role of Damping

The procedure of (4.23)-(4.26) works perfectly well for solving $mu'' + cu' + ku = 0$ when m, c , and k are any real or complex numbers, as long as the solutions to the characteristic equation are distinct. All the operations remain algebraically correct.

However, essentially all of our work will focus on the physically relevant situation in which m, c , and k are real with $m > 0$, $k > 0$, and $c \geq 0$. In this case solving $mu'' + cu' + ku = 0$ breaks into a few mathematically and physically distinct cases, depending on the nature of the roots (4.24) to the characteristic equation. The cases are

- **The Overdamped Case:** This occurs when $c^2 - 4mk > 0$ (that is, $c > 2\sqrt{mk}$, so the damping is sufficiently large). In this case $\sqrt{c^2 - 4mk}$ is a positive real number and the roots r_1 and r_2 of the characteristic equation are real and distinct, as is easily seen from (4.24). Examples 4.4 and 4.5 were both overdamped.

Moreover, both roots in this case must be negative. To see why, note that r_2 in (4.24) is clearly negative since $-c$ and $-\sqrt{c^2 - 4mk}$ are both negative, and m is positive. To see why r_1 is negative, start with

$$c^2 > c^2 - 4mk,$$

which is true since both m and k are positive. Take the square root of both sides above and find $c > \sqrt{c^2 - 4mk}$ (using $\sqrt{c^2} = c$ since $c \geq 0$). Multiply both sides of $c > \sqrt{c^2 - 4mk}$ by -1 to conclude $-c < -\sqrt{c^2 - 4mk}$ or equivalently, $-c + \sqrt{c^2 - 4mk} < 0$. This last inequality shows that the numerator of $r_1 = (-c + \sqrt{c^2 - 4mk})/(2m)$ in (4.24) is negative, hence $r_1 < 0$ since $m > 0$.

Since r_1 and r_2 are both negative, any solution $u(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t}$ consists of a sum of decaying exponentials. The solution decays to zero and does not oscillate, as illustrated in Figures 4.7 and 4.8. In fact, the solution in the overdamped case can cross the horizontal axis at most once; see Exercise 4.2.11.

- **The Underdamped Case:** This occurs when $c^2 - 4mk < 0$ and $c > 0$ (some damping is present; think of a stiff spring and a large mass in air). In this case $\sqrt{c^2 - 4mk}$ is an imaginary number and the roots r_1 and r_2 of the characteristic equation are distinct complex numbers. Moreover, r_1 and r_2 are conjugate to one another; see Appendix A. As will be shown in the next section, solutions in this case oscillate and decay to zero.
- **The Undamped Case:** This occurs when $c = 0$ and can be considered a special case of an underdamped system. Here both roots to the characteristic equation are purely imaginary and given by $\pm \sqrt{-4mk}/(2m)$ or equivalently, $\pm i\sqrt{k/m}$. Solutions in this case oscillate periodically and the oscillations never diminish in amplitude.
- **The Critically Damped Case:** This is the razor's edge between overdamped and underdamped, and occurs when $c^2 - 4mk = 0$. In this case the characteristic equation has a double root $r_1 = r_2 = -c/(2m)$; the double root is real since c and m are real.

Although the solution procedure (4.23)-(4.26) works in the underdamped and undamped settings, there is mathematical and physical insight to be gained by a more careful examination of these two cases.

4.2.3 The Underdamped and Undamped Cases

As mentioned in the last section, the solution procedure (4.23)-(4.26) works perfectly well if the roots to the characteristic equation are complex. Let's start by considering a few examples.

■ **Example 4.6** Consider the harmonic oscillator with $c = 0$, the undamped harmonic oscillator as quantified by equation (4.5). With $m = k = 1$ this ODE becomes $u''(t) + u(t) = 0$, which was examined in Reading Exercise 4.1.2. The characteristic equation (4.23) here is

$$r^2 + 1 = 0$$

with roots $r = i$ and $r = -i$, where $i = \sqrt{-1}$. As dictated by (4.25) a general solution in this case is

$$u(t) = c_1 e^{it} + c_2 e^{-it}, \quad (4.27)$$

where $e^{it} = \cos(t) + i \sin(t)$ and $e^{-it} = \cos(t) - i \sin(t)$ from Euler's formula. See Appendix A for more on Euler's formula.

As a specific example let's consider initial conditions $u(0) = 1$ and $u'(0) = 0$. From the general solution (4.27) the condition $u(0) = 1$ yields $u(0) = c_1 + c_2$ and from $u'(t) = ic_1 e^{it} - ic_2 e^{-it}$ it follows that $u'(0) = i(c_1 - c_2)$. The initial conditions thus dictate that $c_1 + c_2 = 1$ and $i(c_1 - c_2) = 0$.

The solution to these two equations is $c_1 = c_2 = 1/2$, and from (4.27) the solution with the desired initial conditions is then

$$u(t) = \frac{1}{2}e^{it} + \frac{1}{2}e^{-it}.$$

It may seem a bit perplexing, however: the original ODE $u''(t) + u(t) = 0$ and initial conditions $u(0) = 1$ and $u'(0) = 0$ contain no complex numbers; why are there i 's in the solution?

The answer is, there aren't (or rather, there don't have to be). Invoking Euler's formula shows that

$$\begin{aligned} u(t) &= \frac{1}{2}e^{it} + \frac{1}{2}e^{-it} \\ &= \frac{1}{2}(\cos(t) + i\sin(t)) + \frac{1}{2}(\cos(t) - i\sin(t)) \\ &= \cos(t). \end{aligned}$$

The complex numbers appear as roots of the characteristic equation $r^2 + 1 = 0$, facilitate the solution process, and then vanish. Notice that without damping to dissipate the system's energy, the mass oscillates at constant amplitude and never stops. ■

Reading Exercise 4.2.6 Verify that $u(t) = \cos(t)$ satisfies $u'' + u = 0$ with $u(0) = 1$ and $u'(0) = 0$.

Let's next look at a slightly more involved example.

■ **Example 4.7** Let us reconsider the model of Example 4.5, a building with roof mass $m = 10^4$ kg and spring constant $k = 40000$ newtons per meter, but now with damping $c = 20000$ newton-seconds per meter. Again we will compute the displacement of the roof mass if a sudden shock, impact, or wind gust sets the roof in motion with initial velocity $u'(0) = 0.25$ meters per second and initial displacement $u(0) = 0$. The governing ODE is now

$$10000u''(t) + 20000u'(t) + 40000u(t) = 0 \quad (4.28)$$

with initial conditions $u(0) = 0$ and $u'(0) = 1/4$.

To begin, seek solutions of the form $u(t) = e^{rt}$. Substituting this ansatz into the ODE and dividing through by e^{rt} yields the characteristic equation

$$10000r^2 + 20000r + 40000 = 0.$$

The characteristic equation can be written as $10000(r^2 + 2r + 4) = 0$ and has the same solutions as $r^2 + 2r + 4 = 0$, namely $r_1 = -1 + i\sqrt{3}$ and $r_2 = -1 - i\sqrt{3}$; note these roots are complex, indicating the system is underdamped, and the roots are complex conjugates. Both $e^{r_1 t}$ and $e^{r_2 t}$ are solutions to the ODE, and the linearity of the ODE (4.28) and the principle of superposition yield a general solution

$$u(t) = c_1 e^{(-1+i\sqrt{3})t} + c_2 e^{(-1-i\sqrt{3})t}.$$

The initial condition $u(0) = 0$ requires $c_1 + c_2 = 0$, while $u'(t) = c_1 r_1 e^{r_1 t} + c_2 r_2 e^{r_2 t}$ requires $u'(0) = r_1 c_1 + r_2 c_2 = 1/4$, or $(-1 + i\sqrt{3})c_1 + (-1 - i\sqrt{3})c_2 = 1/4$. The simultaneous solution to $c_1 + c_2 = 0$, $(-1 + i\sqrt{3})c_1 + (-1 - i\sqrt{3})c_2 = 1/2$ is $c_1 = -i\sqrt{3}/24$ and $c_2 = i\sqrt{3}/24$ (note c_1 and c_2 are also conjugate to each other.) The full solution to (4.28) is therefore

$$u(t) = -\frac{i\sqrt{3}}{24}e^{(-1+i\sqrt{3})t} + \frac{i\sqrt{3}}{24}e^{(-1-i\sqrt{3})t}.$$

However, as in Example 4.6, the solution contains complex numbers, even though none appeared in the statement of the problem. Moreover, it's clear that $u(t)$, the position of the roof mass, should be a real-valued function of t .

Again, as in Example 4.6, the solution is indeed real-valued. Note that $e^{(-1+i\sqrt{3})t} = e^{-t+it\sqrt{3}} = e^{-t}e^{it\sqrt{3}}$ and similarly $e^{(-1-i\sqrt{3})t} = e^{-t-it\sqrt{3}} = e^{-t}e^{-it\sqrt{3}}$. Euler's formula and a bit of algebra then yield

$$\begin{aligned} u(t) &= -\frac{i\sqrt{3}}{24}e^{(-1+i\sqrt{3})t} + \frac{i\sqrt{3}}{24}e^{(-1-i\sqrt{3})t} \\ &= -\frac{i\sqrt{3}}{24}e^{-t}e^{it\sqrt{3}} + \frac{i\sqrt{3}}{24}e^{-t}e^{-it\sqrt{3}} \\ &= -\frac{i\sqrt{3}}{24}e^{-t}(\cos(t\sqrt{3}) + i\sin(t\sqrt{3})) + \frac{i\sqrt{3}}{24}e^{-t}(\cos(t\sqrt{3}) - i\sin(t\sqrt{3})) \\ &= e^{-t}\left(\left(-\frac{i\sqrt{3}}{24}\right)(\cos(t\sqrt{3}) + i\sin(t\sqrt{3})) + \left(\frac{i\sqrt{3}}{24}\right)(\cos(t\sqrt{3}) - i\sin(t\sqrt{3}))\right) \\ &= \frac{\sqrt{3}}{12}e^{-t}\sin(t\sqrt{3}). \end{aligned}$$

The last line follows from multiplying out the previous line; all the imaginary cross-terms cancel. Thus the solution $u(t) = \frac{\sqrt{3}}{12}e^{-t}\sin(t\sqrt{3})$ is real-valued.

A plot of the solution is shown in Figure 4.9. The solution oscillates, crossing the horizontal axis infinitely many times, although the presence of the e^{-t} decay quickly diminishes the amplitude of the solution. This is typical behavior for an underdamped system with nonzero damping. ■

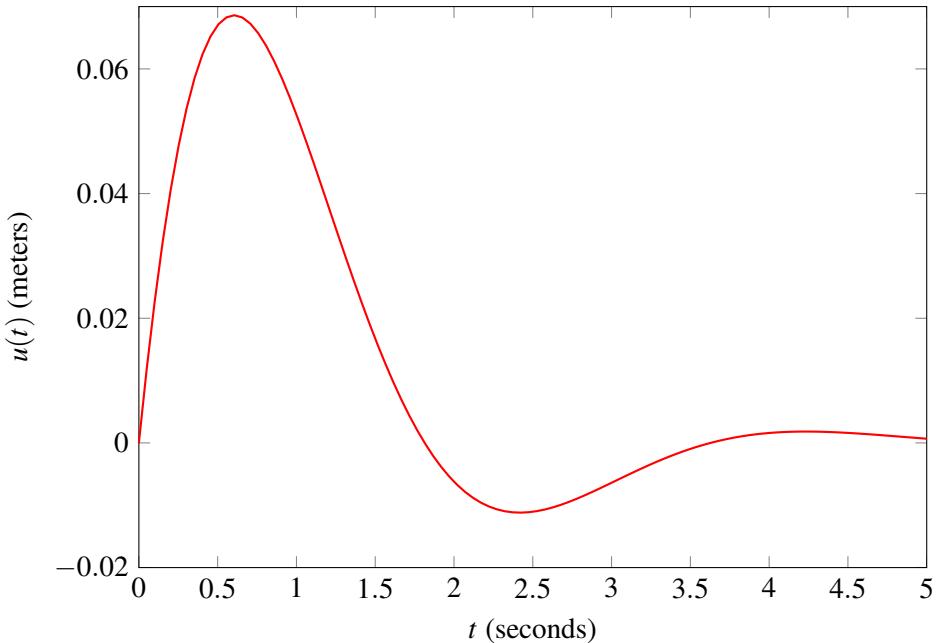


Figure 4.9: Graph of the solution to $10000u''(t) + 20000u'(t) + 40000u(t) = 0$ with $u(0) = 0$ and $u'(0) = 1/4$.

Here's an interesting observation: In the last example the characteristic equation had complex-conjugate roots $-1 \pm i\sqrt{3}$, and the final real-valued solution contained terms e^{-t} , $\sin(t\sqrt{3})$, and $\cos(t\sqrt{3})$. This is no coincidence, as we will soon show.

4.2.4 The General Underdamped Case

Complex Roots

Examples 4.6 and 4.7 illustrate that the solution procedure (4.23)-(4.26) still works if the roots to the characteristic equation are complex, and despite all the complex-valued arithmetic in those examples, the solutions turned out to be real-valued. Let's examine the situation more generally.

Consider the harmonic oscillator ODE (4.4), namely $mu'' + cu' + ku = 0$. The characteristic equation is $mr^2 + cr + k = 0$ with roots given by (4.24). The underdamped (or undamped) case occurs precisely when $c^2 - 4mk < 0$. Let's use the fact that

$$\sqrt{c^2 - 4mk} = i\sqrt{4mk - c^2}$$

and note that $\sqrt{4mk - c^2}$ is a positive real number since $4mk - c^2 > 0$. The roots to the characteristic equation in (4.24) can then be expressed as

$$\begin{aligned} r_1 &= -\frac{c}{2m} + i\frac{\sqrt{4mk - c^2}}{2m} \\ r_2 &= -\frac{c}{2m} - i\frac{\sqrt{4mk - c^2}}{2m}. \end{aligned}$$

For notational convenience define

$$\begin{aligned} \alpha &= \frac{c}{2m} \\ \omega &= \frac{\sqrt{4mk - c^2}}{2m}. \end{aligned} \tag{4.29}$$

Both α and ω are real-valued quantities with $\omega > 0$ and, since it was assumed that $c \geq 0$ it follows that $\alpha \geq 0$. You can check that both α and ω have the dimension T^{-1} , reciprocal time, although each plays a very different role in the solution. The roots to the characteristic equation (4.24) can then be expressed as

$$r_1 = -\alpha + i\omega \quad \text{and} \quad r_2 = -\alpha - i\omega. \tag{4.30}$$

The general solution $u(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t}$ to the harmonic oscillator (4.4) previously constructed can now be expressed as

$$u(t) = c_1 e^{(-\alpha+i\omega)t} + c_2 e^{(-\alpha-i\omega)t}. \tag{4.31}$$

Any initial conditions can be obtained using (4.31) with suitable (and unique) choices for c_1 and c_2 , as given in (4.26).

However, with a little additional work most of the complex arithmetic can be avoided.

A Real-Valued General Solution

Let's write (4.31) in a slightly more insightful form by using Euler's formula and bit of algebra. This alternate form also has the advantage of doing an end run around the complex numbers in the solution process and immediately making obvious the oscillatory nature of the solutions. First, note that

$$\begin{aligned} e^{(-\alpha+i\omega)t} &= e^{-\alpha t + i\omega t} \\ &= e^{-\alpha t} e^{i\omega t} \\ &= e^{-\alpha t} (\cos(\omega t) + i \sin(\omega t)) \\ &= e^{-\alpha t} \cos(\omega t) + i e^{-\alpha t} \sin(\omega t). \end{aligned}$$

A similar computation shows that $e^{(-\alpha+i\omega)t} = e^{-\alpha t} \cos(\omega t) - ie^{-\alpha t} \sin(\omega t)$. Armed with this knowledge, rewrite u in (4.31) as

$$\begin{aligned} u(t) &= c_1 e^{(-\alpha+i\omega)t} + c_2 e^{(-\alpha-i\omega)t} \\ &= c_1 e^{-\alpha t} \cos(\omega t) + c_1 i e^{-\alpha t} \sin(\omega t) + c_2 e^{-\alpha t} \cos(\omega t) - c_2 i e^{-\alpha t} \sin(\omega t) \\ &= \underbrace{(c_1 + c_2)}_{d_1} e^{-\alpha t} \cos(\omega t) + \underbrace{i(c_1 - c_2)}_{d_2} e^{-\alpha t} \sin(\omega t) \end{aligned} \quad (4.32)$$

where c_1 and c_2 are arbitrary constants. Define constants $d_1 = c_1 + c_2$ and $d_2 = i(c_1 - c_2)$, so a general solution (4.32) can also be expressed as

$$u(t) = d_1 e^{-\alpha t} \cos(\omega t) + d_2 e^{-\alpha t} \sin(\omega t). \quad (4.33)$$

This is a **real-valued general solution** for $mu'' + cu' + ku = 0$ in the underdamped case.

In conjunction with (4.29), (4.33) provides an alternate form of a general solution to $mu'' + cu' + ku = 0$, since any solution that can be expressed using (4.25) or (4.31) can be expressed using (4.33) with an appropriate choice of d_1 and d_2 , namely $d_1 = c_1 + c_2$ and $d_2 = i(c_1 - c_2)$. Conversely, any function that can be expressed using (4.33) can also be expressed via (4.25) using an appropriate choice for c_1 and c_2 , namely $c_1 = (d_1 - id_2)/2$ and $c_2 = (d_1 + id_2)/2$, obtained by solving $d_1 = c_1 + c_2$ and $d_2 = i(c_1 - c_2)$ for c_1 and c_2 .

■ **Example 4.8** Suppose the characteristic equation for a spring-mass-damper harmonic oscillator has $-3 + 5i$ as a root. It follows immediately that $-3 - 5i$ is the other root (as long as m, c , and k are real), and so a general solution to the corresponding ODE is

$$u(t) = c_1 e^{(-3+5i)t} + c_2 e^{(-3-5i)t}.$$

Based on (4.33) it follows that

$$u(t) = d_1 e^{-3t} \cos(5t) + d_2 e^{-3t} \sin(5t)$$

is also a general solution. All of this can be deduced from the fact that $-3 + 5i$ is a root of the characteristic equation. ■

■ **Example 4.9** Let us write out a general solution to

$$2u''(t) + 4u'(t) + 20u(t) = 0$$

using both (4.25) and (4.33) and use each to find a specific solution with initial conditions $u(0) = 1$ and $u'(0) = 2$.

First, the characteristic equation is $2r^2 + 4r + 20 = 0$ and has roots $r = -1 \pm 3i$. A complex-valued general solution is therefore

$$u(t) = c_1 e^{(-1+3i)t} + c_2 e^{(-1-3i)t}.$$

With this general solution the initial conditions dictate $c_1 + c_2 = 1$ and $(-1 + 3i)c_1 + (-1 - 3i)c_2 = 3$, with solution $c_1 = 1/2 - i/2, c_2 = 1/2 + i/2$. Thus the solution with these initial conditions is

$$u(t) = \left(\frac{1}{2} - \frac{i}{2}\right) e^{(-1+3i)t} + \left(\frac{1}{2} + \frac{i}{2}\right) e^{(-1-3i)t}. \quad (4.34)$$

But if we instead work with (4.33) and (4.29), then $\alpha = 1$ and $\omega = 3$; note $-\alpha = -1$ is exactly the real part of the roots of the characteristic equation while $\omega = 3$ is the absolute value of the imaginary part. Then (4.33) becomes the real-valued general solution

$$u(t) = d_1 e^{-t} \cos(3t) + d_2 e^{-t} \sin(3t).$$

The initial condition $u(0) = 1$ dictates $d_1 = 1$. Compute $u'(t) = d_1(-e^{-t} \cos(3t) - 3e^{-t} \sin(3t)) + d_2(-e^{-t} \sin(3t) + 3e^{-t} \cos(3t))$, and then $u'(0) = 3$ becomes $-1 + 3d_2 = 2$, so that $d_2 = 1$. This yields the solution

$$u(t) = e^{-t} \cos(3t) + e^{-t} \sin(3t). \quad (4.35)$$

Applying Euler's formula to (4.34) yields exactly (4.35). ■

Reading Exercise 4.2.7 Write out a complex-valued general solution (4.25) to $3u''(t) + 18u'(t) + 75u(t) = 0$ and use it to find the solution with $u(0) = 0$ and $u'(0) = 4$. Repeat using a general solution in the form (4.33). Verify that the two specific solutions with these initial conditions are the same.

Observations on the Real-Valued Solution (4.33)

The real-valued general solution (4.33) also provides some useful insights in the case that the system is underdamped.

1. The presence of the $\sin(\omega t)$ and $\cos(\omega t)$ terms indicate that the solution is oscillatory, with radial frequency ω .
2. If $c > 0$ (indicating the presence of some damping) then $-\alpha = -\frac{c}{2m} < 0$ and the solution decays due to the presence of $e^{-\alpha t}$ in (4.33).
3. If $c = 0$ (indicating the absence of any damping) then $\alpha = 0$ and the solution is periodic, of the form $u(t) = d_1 \cos(\omega t) + d_2 \sin(\omega t)$. The solution has period $2\pi/\omega$, where $\omega = \sqrt{k/m}$.

In the underdamped or undamped case the system vibrates at a frequency of ω radians per second.

Definition 4.2.2 For an underdamped or undamped system the quantity $\omega = \frac{\sqrt{4mk-c^2}}{2m}$ in (4.29) is known as the **natural frequency** of the spring-mass system (in radians per unit time). When $c = 0$ the natural frequency is $\omega = \sqrt{k/m}$ radians per unit time.

4.2.5 The Critically Damped Case

The general solutions presented above fail in the critically damped case, where the characteristic equation has a double root. Let's consider an example.

■ **Example 4.10** Consider the second-order ODE

$$u''(t) + 4u'(t) + 4u(t) = 0,$$

corresponding to a spring-mass system with $m = 1, c = 4, k = 4$. The characteristic equation is $r^2 + 4r + 4 = 0$, which factors as $(r + 2)^2 = 0$. This quadratic equation has a double root at $r = -2$. Using (4.25) in an attempt to produce a general solution leads to $u(t) = c_1 e^{-2t} + c_2 t e^{-2t}$. You should smell a rat at this point, for the two pieces of the solution are both e^{-2t} , copies of each other. Using this (incorrect) general solution to produce a solution with initial conditions $u(0) = u_0$ and $u'(0) = v_0$ yields equations $c_1 + c_2 = u_0$ and $-2c_1 - 2c_2 = v_0$. These equations are dependent and not solvable unless $v_0 = -2u_0$, which need not be the case.

However, the function $u(t) = c_1 e^{-2t} + c_2 t e^{-2t}$ is indeed a solution to the ODE for any choice of c_1 and c_2 , but it is not a general solution, since c_1 and c_2 cannot be adjusted to obtain arbitrary initial conditions. ■

Producing a Second Solution

If the procedure for finding a general solution fails in the case of a double root, what are we to do? Let's reconsider Example 4.10. Computing the roots of the characteristic equation does indeed produce a solution e^{-2t} (or any multiple thereof), but we need a second independent solution to

use in a superposition with e^{-2t} to form a true general solution that can accommodate any initial conditions.

Here is a technique of some versatility. The function $u_1(t) = ce^{-2t}$ is a solution to $u''(t) + 4u'(t) + 4u(t) = 0$, as deduced in Example 4.10. Consider the possibility of constructing another solution by replacing the constant c in $u(t) = c^{-2t}$ by a function of t , say $c(t)$. That is, seek another solution $u_2(t)$ of the form

$$u_2(t) = c(t)e^{-2t}. \quad (4.36)$$

Inserting $u(t) = u_2(t)$ into the ODE $u''(t) + 4u'(t) + 4u(t) = 0$ and simplifying produces many cancellations and yields

$$c''(t)e^{-2t} = 0.$$

Since e^{-2t} is never zero it follows that (4.36) provides a solution to $u''(t) + 4u'(t) + 4u(t) = 0$ if $c''(t) = 0$, or if $c(t) = At + B$ for any constants A and B . Taking $A = 0$ leads us back to solutions Be^{-2t} that are multiples of e^{-2t} , but if $A \neq 0$, we get something different. In particular, let's take $A = 1$ and $B = 0$ to construct the solution $u_2(t) = te^{-2t}$.

Reading Exercise 4.2.8 Verify that $u_2(t) = te^{-2t}$ satisfies $u''(t) + 4u'(t) + 4u(t) = 0$. What initial conditions $u_2(0)$ and $u'_2(0)$ does this solution satisfy?

The function $u_2(t)$ is not a scalar multiple of $u_1(t)$, and in fact

$$u(t) = c_1u_1(t) + c_2u_2(t) = c_1e^{-2t} + c_2te^{-2t} \quad (4.37)$$

provides a general solution to $u''(t) + 4u'(t) + 4u(t) = 0$ that can be used to find a solution with any specified initial conditions. To see this, consider arbitrary initial conditions $u(0) = u_0$ and $u'(0) = v_0$. The condition $u(0) = u_0$ in (4.37) dictates $c_1 = u_0$, since $u_2(0) = 0$. Use (4.37) to compute $u'(t) = -2c_1e^{-2t} + c_2(-2te^{-2t} + e^{-2t})$ and then $u'(0) = v_0$ forces $-2c_1 + c_2 = v_0$, so $c_2 = v_0 + 2c_1 = v_0 + 2u_0$. Thus any desired initial conditions can be obtained by using (4.37), so (4.37) is a general solution.

■ **Example 4.11** Let us solve the ODE $u''(t) + 4u'(t) + 4u(t) = 0$ of Example 4.10, with initial conditions $u(0) = 1$ and $u'(0) = -1$. A general solution to this ODE is given by (4.37) and the initial conditions require $u(0) = c_1 = 1$ and $u'(0) = -2c_1 + c_2 = -1$ with solution $c_1 = 1$ and $c_2 = 1$. The solution is thus

$$u(t) = e^{-2t} + te^{-2t}$$

and is graphed in Figure 4.10. This graph looks very much like the overdamped case; it's hard to determine whether a system is critically damped by looking at a solution graph. ■

A General Solution in the Critically Damped Case

The above procedure is called **reduction of order** and it works more generally. Suppose a system governed by $mu'' + cu' + ku = 0$ is critically damped, so the characteristic equation $mr^2 + cr + k = 0$ has a double root, say at $r = -\alpha$. Then $mr^2 + cr + k$ factors as $m(r + \alpha)^2$, and so

$$mr^2 + cr + k = m(r + \alpha)^2 = mr^2 + 2m\alpha r + m\alpha^2.$$

The coefficients for r and the constant term must match, and so $c = 2m\alpha$ and $k = m\alpha^2$. Thus the harmonic oscillator ODE in this case can also be written as

$$mu''(t) + 2m\alpha u'(t) + m\alpha^2 u(t) = 0. \quad (4.38)$$

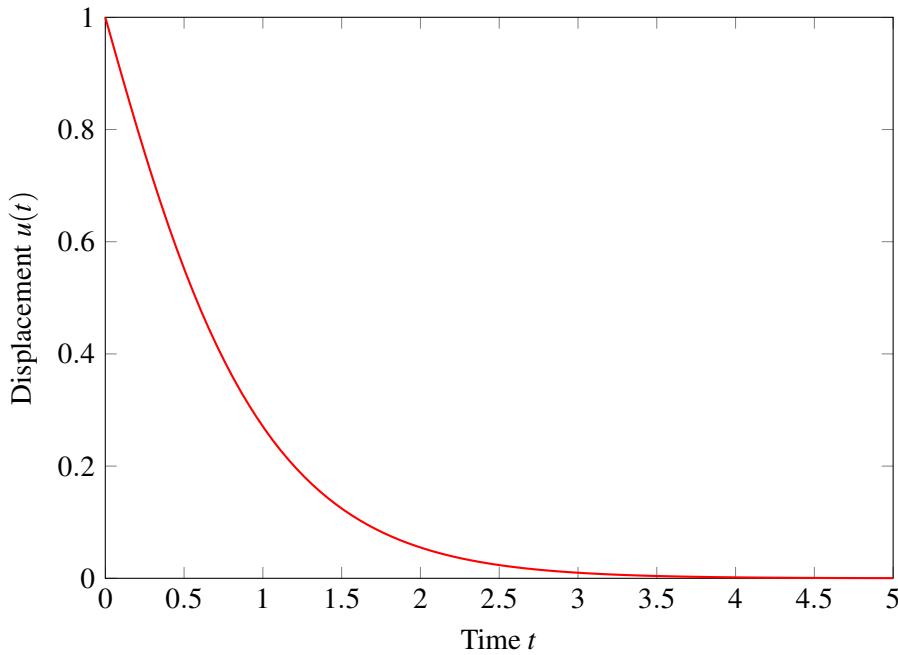


Figure 4.10: Graph of the solution to $u''(t) + 4u'(t) + 4u(t) = 0$ with $u(0) = 1$ and $u'(0) = -1$.

We know from the characteristic equation that $u_1(t) = e^{-\alpha t}$ satisfies the ODE. Inserting $u_2(t) = c(t)e^{-\alpha t}$ into (4.38) and simplifying produces, after fortuitous cancellations, the equation $me^{-\alpha t}c''(t) = 0$. Since $m > 0$ and $e^{-\alpha t} > 0$, it follows that $c''(t) = 0$, so $c(t) = At + B$ for some A and B . In particular, take $A = 1$ and $B = 0$ to find that $u_2(t) = te^{-\alpha t}$ is also a solution, where α is the root of the characteristic equation. You can then easily verify that

$$u(t) = c_1 e^{-\alpha t} + c_2 t e^{-\alpha t} \quad (4.39)$$

is a general solution to the ODE in the critically damped case: any initial conditions can be obtained with an appropriate choice of c_1 and c_2 .

Reading Exercise 4.2.9 Verify that substituting $u_2(t) = c(t)e^{-\alpha t}$ into (4.38) yields $me^{-\alpha t}c''(t) = 0$. Then check that $u(t)$ as given by (4.39) can be used to find a solution with $u(0) = u_0$ and $u'(0) = v_0$, for any choice of u_0 and v_0 . What are c_1 and c_2 in terms of u_0 and v_0 ?

4.2.6 The Existence and Uniqueness of Solutions

The general solutions produced in the various cases are aptly named, for every solution to $mu'' + cu' + ku = 0$ can be expressed using these general solutions. To see why, first note the following theorem:

Theorem 4.2.1 If $m \neq 0$ then the ODE

$$mu''(t) + cu'(t) + ku(t) = 0$$

with initial conditions $u(0) = u_0$ and $u'(0) = v_0$ has a unique solution, and the solution exists for all t .

This theorem also holds if the initial conditions are given at an arbitrary time $t = t_0$, and even under more general conditions in which m, c , or k is not constant. For a proof see [101].

Why These Are Called General Solutions

Any solution to $mu''(t) + cu'(t) + ku(t) = 0$ has some initial data, namely $u_0 = u(0)$ and $v_0 = u'(0)$, and by virtue of Theorem 4.2.1, this $u(t)$ is the unique solution with this initial data. Moreover, in every case considered—underdamped, overdamped, and critically damped—we produced a general solution as in Definition 4.2.1 from which any initial conditions $u(0) = u_0$ and $u'(0) = v_0$ can be obtained. It follows that every solution to $mu''(t) + cu'(t) + ku(t) = 0$ is represented by the corresponding general solution we constructed. We have not missed anything.

4.2.7 Summary and a Physical Perspective

For the harmonic oscillator equation $mu'' + cu' + ku = 0$ with $m > 0$, $k > 0$, and $c \geq 0$, we've encountered four distinct cases. To summarize:

- **The Overdamped Case:** This occurs when $c^2 - 4mk > 0$ (the damping coefficient c is sufficiently large, $c > 2\sqrt{mk}$). In this case both roots r_1 and r_2 of the characteristic equation are real, distinct, and negative. A general solution is given by (4.25). Any solution decays exponentially in time and its graph crosses the horizontal axis at most once.
- **The Underdamped Case:** This occurs when $c^2 - 4mk < 0$. In this case the roots r_1 and r_2 of the characteristic equation are a complex-conjugate pair. If $c > 0$ these roots have negative real part. A general solution is given by (4.25), but a real-valued general solution can also be written in the form (4.33) with (4.29). The solution contains sines and cosines, and if $c > 0$ the amplitudes decays exponentially in time t .
- **The Undamped Case:** This occurs when $c = 0$ and might be considered a special case of an underdamped system. Here both roots to the characteristic equation are purely imaginary, $\pm i\sqrt{4mk}/(2m)$, which simplifies to $\pm i\sqrt{k/m}$. Solutions to the ODE are of the form (4.33) with $\alpha = 0$, so the solutions oscillate forever without decay. The period of the solution is $2\pi/\omega$ with $\omega = \sqrt{k/m}$, so the period P can also be expressed as

$$P = 2\pi\sqrt{m/k}.$$

Although most physical systems don't really have zero damping, when damping is close to zero it can be useful to posit an undamped model to gain insight into the system's behavior, and then move to the more realistic damped model.

- **The Critically Damped Case:** This is the razor's edge between overdamped and underdamped, and occurs when $c^2 = 4mk$. Note that since we assume $m > 0$ and $k > 0$ this also requires $c > 0$. In this case the characteristic equation has a double root at $-c/(2m)$. Any solution is of the form $u(t) = c_1 e^{-\alpha t} + c_2 t e^{-\alpha t}$ with $\alpha = c/(2m)$. Solutions decay to zero and do not oscillate, similar to the overdamped case.

4.2.8 Exercises

Exercise 4.2.1 For each set of parameters m , c , and k in (a)-(j), find the appropriate ODE that governs the corresponding spring-mass-damper system and write out the characteristic equation. Find the roots of the characteristic equation and determine whether the system is undamped, underdamped, critically damped, or overdamped.

- $m = 3, c = 24, k = 60$
- $m = 1, c = 0, k = 20$
- $m = 2, c = 12, k = 10$
- $m = 2, c = 16, k = 64$
- $m = 2, c = 4, k = 10$
- $m = 3, c = 21, k = 36$

- (g) $m = 2, c = 12, k = 18$
- (h) $m = 3, c = 18, k = 75$
- (i) $m = 2, c = 8, k = 6$
- (j) $m = 5, c = 10, k = 5$

Exercise 4.2.2 For each set of parameters m, c , and k in (a)-(h), find the appropriate ODE that governs the corresponding spring-mass-damper system, with $u(t)$ as the dependent variable. In each case the system is overdamped. Then form the characteristic equation, find its roots, and write out a general solution to the ODE. Finally, use the general solution to obtain the specific solution with initial conditions $u(0) = 2$ and $u'(0) = 3$. Graph the specific solution on the interval $0 \leq t \leq 5$.

- (a) $m = 1, c = 6, k = 8$
- (b) $m = 3, c = 9, k = 6$
- (c) $m = 2, c = 10, k = 12$
- (d) $m = 3, c = 21, k = 36$
- (e) $m = 2, c = 10, k = 8$
- (f) $m = 1, c = 7, k = 12$
- (g) $m = 3, c = 18, k = 24$
- (h) $m = 1, c = 4, k = 3$

Exercise 4.2.3 For each set of parameters m, c , and k in (a)-(h), find the appropriate ODE that governs the corresponding spring-mass-damper system, with $u(t)$ as the dependent variable. In each case the system is underdamped. Then form the characteristic equation and find its roots. Use this information to find a complex-valued general solution of the form $u(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t}$ and a real-valued general solution of the form (4.33). Use each general solution to find a solution with initial conditions $u(0) = 2$ and $u'(0) = 4$, and verify that the solutions are identical. Then graph the solution.

- (a) $m = 1, c = 4, k = 5$
- (b) $m = 2, c = 4, k = 20$
- (c) $m = 2, c = 16, k = 64$
- (d) $m = 1, c = 6, k = 18$
- (e) $m = 2, c = 8, k = 10$
- (f) $m = 3, c = 12, k = 60$
- (g) $m = 2, c = 16, k = 50$
- (h) $m = 3, c = 12, k = 39$

Exercise 4.2.4 For each set of parameters m, c , and k in (a)-(d), find the appropriate ODE that governs the corresponding spring-mass-damper system, with $u(t)$ as the dependent variable. In each case the system is critically damped. Then form the characteristic equation and find its root. Use this information to find a general solution of the form (4.39). Use the general solution to find a solution with initial conditions $u(0) = 2$ and $u'(0) = 4$, and graph the solution.

- (a) $m = 1, c = 4, k = 4$
- (b) $m = 3, c = 6, k = 3$
- (c) $m = 2, c = 8, k = 8$

- (d) $m = 5, c = 40, k = 80$

Exercise 4.2.5 Consider a building as modeled in Section 4.1.2 (see also Examples 4.5 and 4.7), but with roof mass $m = 20000$ kg and spring constant $k = 60000$ newtons per meter.

- Suppose the damping constant is $c = 80000$ newton-seconds per meter. A gust of wind imparts an initial velocity $u'(0) = 0.1$ meters per second to the roof; assume $u(0) = 0$. Write out and solve the ODE that governs $u(t)$, the displacement of the building from equilibrium. Plot $u(t)$ on the range $0 \leq t \leq 10$. Is this system overdamped, underdamped, undamped, or critically damped?
- Repeat part (a) but with $c = 40000$. Write the solution in a real-valued form as in Example 4.7.
- Repeat part (a) with $c = 0$. Write the solution in a real-valued form as in Example 4.7.
- What value for c would result in a critically damped system? Write out the solution in this case and plot on the range $0 \leq t \leq 10$.

Exercise 4.2.6 Consider a vibration table governed by (4.9) but with $d(t) = 0$ for all t (the ground is not in motion).

- Show that $y(t)$ satisfies the second-order linear nonhomogeneous ODE

$$my''(t) + cy'(t) + ky(t) = -mg.$$

- What is the equilibrium position of the table top? That is, if $y(t) = y_{eq}$, what is y_{eq} here, in terms of m, g , and k ?
- Define $w(t) = y(t) - y_{eq}$ (or $y(t) = y_{eq} + w(t)$), so $w(t)$ is the displacement of the table top from equilibrium. Show that $u(t)$ satisfies a homogeneous equation

$$mw''(t) + cw'(t) + wu(t) = 0.$$

- Take $g = 9.8$ meters per second squared, and suppose that $m = 100$ kg, $k = 10^4$ newtons per meter, and $c = 2000$ newton-seconds per meter. Verify that the ODE for $w(t)$ is critically damped.
- Suppose someone bumps the table top and imparts an initial velocity of $w'(0) = 0.01$ meters per second to it. Assume $w(0) = 0$ (the table was at equilibrium). Find $w(t)$, plot this function for $0 \leq t \leq 1$, and compute how long it will take for the table top to return to within 0.0001 meters of its equilibrium position.

Exercise 4.2.7 Consider a pendulum of length L that swings back and forth without friction. Let $\theta(t)$ be the angle that the pendulum makes with a vertical line; see Figure 4.32 and Sections 4.6.5 or 4.6.6, where we derive the ODE

$$\theta''(t) + \frac{g}{L}\theta(t) = 0$$

that the function $\theta(t)$ approximately satisfies, at least if the angle $\theta(t)$ remains relatively close to zero (say, $|\theta(t)| \leq \pi/6$, about 30 degrees).

- Which of the spring-mass models does this correspond to—overdamped, critically damped,

underdamped, or undamped?

- (b) Find a general solution to $\theta''(t) + \frac{g}{L}\theta(t) = 0$.
- (c) Find a formula for P , the period of the pendulum (one back and forth swing) in terms of g and L . Do a quick check on the reasonableness of your formula—what does it predict if L is larger or smaller? What if g were larger or smaller?

Exercise 4.2.8 An RLC circuit has inductance $L = 10^{-4}$ henries, resistance $R = 0.1$ ohms, and capacitance $C = 10^{-4}$ farads, with no voltage source, so $v(t) = 0$ volts. At time $t = 0$ the capacitor has charge $q(0) = 5 \times 10^{-4}$ coulombs and no current flows in the circuit. Set up and solve the appropriate ODE. Is this system underdamped, critically damped, or overdamped? Plot the solution on the time range $0 \leq t \leq 0.01$.

Exercise 4.2.9 The solution to an undamped spring-mass system is of the form

$$u(t) = A \cos(\omega t) + B \sin(\omega t) \quad (4.40)$$

for constants A and B . However, it is always possible and sometimes useful to exhibit the solution in the form

$$u(t) = C \sin(\omega t + \phi). \quad (4.41)$$

Here C is the **amplitude** of u and ϕ is the **phase shift**.

- (a) Apply the trigonometric identity $\sin(x+y) = \sin(x)\cos(y) + \cos(x)\sin(y)$ to (4.41) with $x = \omega t$ and $y = \phi$, then compare the result to the right side of (4.40) to show that these expressions will be identical as functions of t if

$$C \sin(\phi) = A \quad \text{and} \quad C \cos(\phi) = B. \quad (4.42)$$

Thus if we are given $u(t)$ in the form (4.41), we can express $u(t)$ in the form (4.40).

- (b) Use (4.42) to show that given A and B we can solve for C as

$$C = \sqrt{A^2 + B^2}.$$

Note $C \geq 0$.

- (c) Use (4.42) to show that given A and B then $\tan(\phi) = A/B$, and so we can solve for ϕ as

$$\phi = \arctan(A/B),$$

if we adjust properly for the cases when $B < 0$ or $B = 0$. Hint: it's just polar coordinates.

Exercise 4.2.10 An unforced spring-mass-damper system governed by $mu'' + cu' + ku = 0$ with mass $m = 3.3$ kg is displaced to initial position $u(0) = 1$ meter and released with no initial velocity. Measurements of the mass displacement are made at times $t = 2, 4, 6, 8, 10$ and yield the data in Table 4.1.

Use this data to estimate c and k . A possible outline: the data suggests an overdamped

system, so consider a solution to the ODE of the form

$$u(t) = c_1 e^{-r_1 t} + c_2 e^{-r_2 t},$$

where we may as well assume $0 < r_1 \leq r_2$. Use $u(0) = 1$ and $u'(0) = 0$ to show that $c_1 = -r_2/(r_1 - r_2)$ and $c_2 = r_1/(r_1 - r_2)$, then use these values for c_1 and c_2 in $u(t)$ and adjust r_1 and r_2 to obtain a good fit to the data (perhaps using least-squares). Finally, use the fact that $-r_1$ and $-r_2$ are roots to the characteristic equation to infer c and k .

Time (seconds)	2	4	6	8	10
Displacement (meters)	0.559	0.258	0.118	0.054	0.025

Table 4.1: Data for spring-mass-damper system in Exercise 4.2.10.

Exercise 4.2.11 Suppose an overdamped spring-mass-damper system has position $u(t)$ given by

$$u(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t}$$

with $r_1 \neq r_2$; note r_1 and r_2 are real and negative. Suppose also that $u(t^*) = 0$ and $u(t^{**}) = 0$ for two times $t = t^*$ and $t = t^{**}$ with $t^* \neq t^{**}$. Show that this forces $c_1 = c_2 = 0$, so $u(t)$ is the zero function. As a result, an overdamped system can return to its equilibrium position at most once, but not twice (unless it remains at equilibrium). Hint: treat $u(t^*) = 0$ and $u(t^{**}) = 0$ as two equations in two unknowns, c_1 and c_2 . Why is $c_1 = c_2 = 0$ the only solution?

Exercise 4.2.12 Suppose a critically damped spring-mass-damper system has position $u(t)$ given by

$$u(t) = c_1 e^{-\alpha t} + c_2 t e^{-\alpha t}.$$

Suppose also that $u(t^*) = 0$ and $u(t^{**}) = 0$ for two times $t = t^*$ and $t = t^{**}$ with $t^* \neq t^{**}$. Show that this forces $c_1 = c_2 = 0$, so $u(t)$ is the zero function. As a result, a critically damped system can return to its equilibrium position at most once, but not twice (unless it remains at equilibrium). Hint: treat $u(t^*) = 0$ and $u(t^{**}) = 0$ as two equations in two unknowns, c_1 and c_2 . Why is $c_1 = c_2 = 0$ the only solution?

Exercise 4.2.13 Suppose a spring-mass-damper ODE is underdamped, so $c^2 - 4mk < 0$. In this problem we show that with real-valued initial conditions $u(0) = u_0$ and $u'(0) = v_0$, the solution we obtain from the complex-valued general solution (4.25) is real-valued, even though many intermediate computations may involve complex numbers. Recall that if $z = a + bi$ is complex then the complex conjugate of z is the complex number $\bar{z} = a - bi$.

- (a) Show that the roots r_1 and r_2 of the characteristic equation are complex, distinct, and complex conjugates.
- (b) Use Euler's identity to show that the terms $e^{r_1 t}$ and $e^{r_2 t}$ in the general solution $u(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t}$ are complex conjugates.
- (c) Suppose we have initial conditions $u(0) = u_0$ and $u'(0) = v_0$ with u_0 and v_0 real numbers. Then $c_1 + c_2 = u_0$ and $r_1 c_1 + r_2 c_2 = v_0$ in the general solution. Show that c_1 and c_2 are

complex conjugates. Hint: use (4.30) to express r_1 and r_2 , then show that

$$c_1 = \frac{u_0}{2} - i(\alpha u_0 + v_0)/(2m) \text{ and } c_2 = \frac{u_0}{2} + i(\alpha u_0 + v_0)/(2m).$$

- (d) Show that with c_1 and c_2 as (c) the solution $u(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t}$ is real-valued. Hint: if z and w are complex numbers then $\overline{w+z} = \overline{w} + \overline{z}$ and $\overline{wz} = (\overline{w})(\overline{z})$. Also, $z + \overline{z}$ is real for any complex number z .

Exercise 4.2.14 According to the ideal gas law, the pressure P , volume V , temperature T (degrees Kelvin), and number of moles n of an ideal gas (one mole is the number 6.02×10^{23}) in a closed container satisfies $PV = nRT$, where $R \approx 8.3145$ joules per degree per mole is the universal gas constant. Suppose a cylinder with cross-sectional area A as in Figure 4.11 contains n moles of an ideal gas, above which lies a piston of mass m .

Assume that the temperature is constant and that the only forces acting on the piston are gravity and the gas pressure from inside the cylinder. In particular, we suppose that there is no atmosphere outside the apparatus pushing down on the piston. Let t denote time and y denote the vertical distance of the bottom of the piston from the bottom of the cylinder, as indicated; positive y is upward. Let g denote gravitational acceleration, with $g > 0$.

- (a) Use Newton's second law $F = ma$ to show that in the absence of any friction the position of the piston satisfies $my''(t) = \frac{nRT}{y(t)} - mg$.
- (b) Show that the equilibrium position of the piston (where gas pressure and gravity are balanced) is given by

$$y_{eq} = \frac{nRT}{mg}.$$

Hint: the upward force on the piston due to the gas pressure in the cylinder is just PA , since pressure is force per area. Also, $V = yA$ for the cylinder.

- (c) Let $u(t) = y(t) - y_{eq}$. Use Newton's second law to show that $u(t)$ obeys the second-order differential equation

$$mu''(t) = -mg + \frac{nRT}{u(t) + y_{eq}}. \quad (4.43)$$

Note that (4.43) is not linear.

- (d) Take $n = 0.01$ moles, $m = 1.0$ Kg, $g = 9.8$ meters per second squared, $T = 300$ degrees Kelvin, and $R = 8.3145$ (units are joules per degree per mole). Solve the ODE from part (c) using initial conditions $u(0) = 0.2$ meters and $u'(0) = 0$ meters per second, and then plot for $t = 0$ to $t = 20$. You'll have to do it numerically. What is the approximate frequency of oscillation of the piston?
- (e) The function $f(u) = 1/(u+a)$ has a tangent line approximation (or linearization) given by

$$f(u) \approx 1/a - u/a^2 + O(u^2)$$

at $u = 0$. Use this to show that (4.43) can be approximated by the linear second-order ODE

$$u''(t) + \frac{mg^2}{nRT}u(t) = 0. \quad (4.44)$$

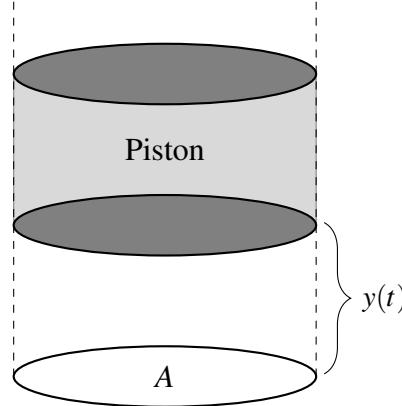


Figure 4.11: Piston in a gas-filled cylinder with cross-sectional area A .

This should be valid as long as $u \approx 0$ (the piston doesn't stray too far from equilibrium). Find the general solution to this linearized ODE exactly, and use it to estimate the frequency of the piston near equilibrium. Plot the solution with $u(0) = 0.2$ and $u'(0) = 0$ for $0 \leq t \leq 20$ and compare the plot to your answer in part (c). Can you see any difference in the solution to (4.43) and the linearized approximation (4.44)?

- (f) Repeat the solution process for $u(t)$ in parts (d) and (e) but with initial conditions $u(0) = 2$ and $u'(0) = 0$. Plot the solution $u(t)$ and compare this plot to the plot you obtained in part (e).

4.3 The Forced Harmonic Oscillator

The unforced harmonic oscillator (4.4) stems from the assumption that the only forces acting on the mass m are the spring force F_{spring} of (4.1) and frictional forces of the form $F_{damping}$, as quantified by (4.2). However, it is often the case that additional forces act on the mass. For example, in our simplified earthquake model, the shaking of the ground is the very thing that sets the mass in motion.

We now return to the nonhomogeneous forced harmonic oscillator ODE (4.3), reproduced here for convenience,

$$mu''(t) + cu'(t) + ku(t) = f(t). \quad (4.45)$$

As in the homogeneous case (4.4) two initial conditions are needed to specify a unique solution. The two-fold focus in this section is the general structure of solutions to (4.45) and how to find solutions with specified initial conditions. But first, let's look at an example.

Example 4.12 Consider a structure as in Figure 4.1 modeled as a spring-mass-damper system with $m = 5000$ kg, $k = 5 \times 10^5$ newtons per meter, and $c = 10^4$ newton-seconds per meter. With no additional forces on the mass the relevant ODE is $5000u'' + 10^4u' + (5 \times 10^5)u = 0$. The roots to the characteristic equation are approximately $-1 \pm 9.95i$ and in view of (4.33) a general solution is

$$u(t) = c_1 e^{-t} \cos(9.95t) + c_2 e^{-t} \sin(9.95t).$$

With initial conditions $u(0) = 0.01$ and $u'(0) = 0$ the solution is

$$u(t) = 0.01e^{-t} \cos(9.95t) + 0.001e^{-t} \sin(9.95t).$$

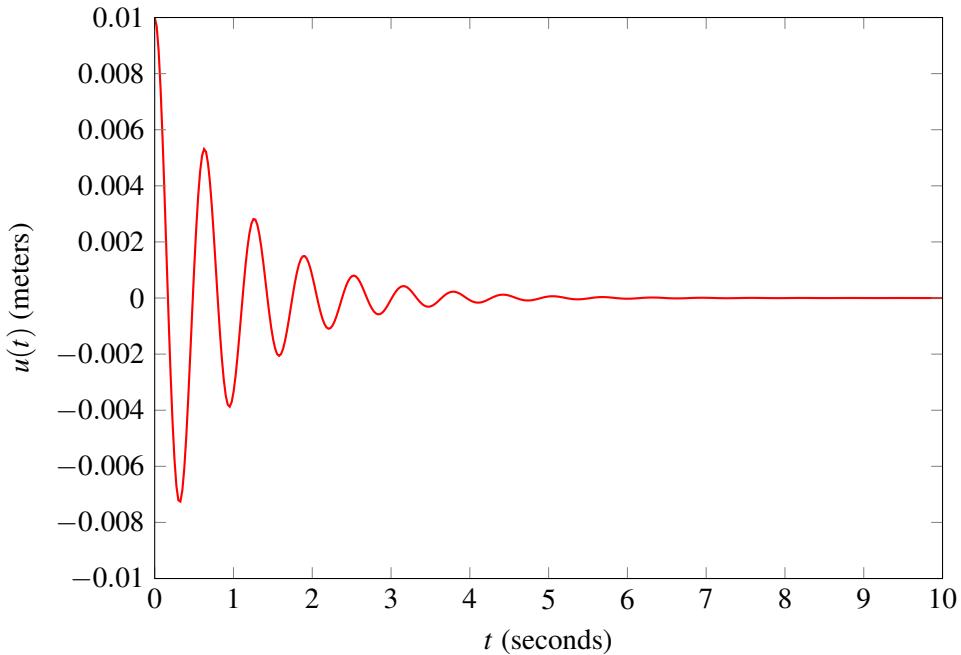


Figure 4.12: Response $u(t)$ of the unforced oscillator $5000u'' + 10^4u' + (5 \times 10^5)u = 0$ with $u(0) = 0.01$ and $u'(0) = 0$.

This function is plotted in Figure 4.12. As expected, since there are no external forces acting on the system after time $t > 0$, the oscillations damp out after a few cycles (though there may have been external forces for $t \leq 0$, in order to initiate the initial nonzero displacement). The sinusoidal portion of the damped oscillations has a frequency of $\frac{9.95}{2\pi} \approx 1.58$ hertz, or a period of about 0.63 seconds.

How would such a structure behave in an earthquake? Here we will model an earthquake as a driving force on the system (but see the corresponding Project "Earthquake Modeling" in Section 4.6 for a slightly more realistic model.) The response to the driven system may be quite different from the undriven system and depend very much on the nature of the driving force. To illustrate, suppose the mass m is at initial position $u(0) = 0$ with $u'(0) = 0$ when an earthquake strikes. We model the situation using (4.3). It's not clear what $f(t)$ is appropriate, but earthquake shaking often contains strong periodic components, with frequencies in the range of 0.2 to 2 hertz or higher; see Chapter 4 of [16]. Let's consider the choice $f(t) = 10^4 \sin(5t)$ in (4.3), which leads to the ODE

$$5000u''(t) + 10^4u'(t) + (5 \times 10^5)u(t) = 10^4 \sin(5t).$$

Here $f(t)$ is periodic with frequency $5/(2\pi) \approx 0.8$ hertz and amplitude 10^4 newtons, which is about 2040 pounds. In this case the solution to the ODE with $u(0) = u'(0) = 0$ is, to three significant figures,

$$u(t) = \underbrace{-0.0128e^{-t} \sin(9.95t)}_{\text{transient}} + \underbrace{0.0035e^{-t} \cos(9.95t)}_{\text{transient}} + \underbrace{0.0262 \sin(5t) - 0.0035 \cos(5t)}_{\text{periodic}}. \quad (4.46)$$

A graph of the response $u(t)$ is shown in Figure 4.13. The solution (4.46) has a **transient** portion stemming from the terms that contain e^{-t} ; this part of the solution quickly decays to zero. However, the two terms on the right in (4.46) involving $\sin(5t)$ and $\cos(5t)$ are the **periodic** portion of the solution and do not decay in time. The periodic portion persists, and quickly dominates the solution as t increases. The long-term response of the building is to shake sinusoidally at the

same frequency as the driving force $f(t)$, with an amplitude of about 0.026 meters. This response will continue as long as $f(t)$ remains active. In this section we'll focus on how this solution was obtained and its general structure. ■

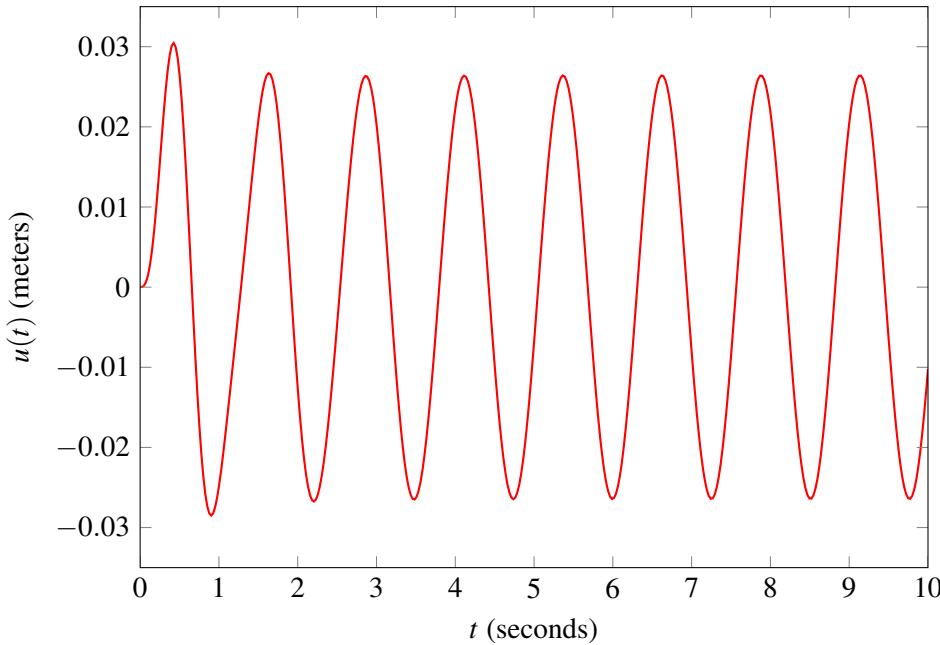


Figure 4.13: Response $u(t)$ of the forced oscillator $5000u'' + 10^4u' + (5 \times 10^5)u = f(t)$ with $f(t) = 10^4 \sin(5t)$.

4.3.1 Solving the Forced Harmonic Oscillator Equation

We will now develop a procedure for solving (4.3) with initial conditions $u(0) = u_0$ and $u'(0) = v_0$, for a variety of common choices for $f(t)$. This will lead to an understanding of the nature and structure of the solution. Linearity and the principle of superposition will be essential.

A General Solution to the Forced Equation

Finding a general solution to the nonhomogeneous ODE (4.3) involves the following steps:

1. Find a general solution $u(t) = u_h(t)$ to the homogeneous ODE $mu''(t) + cu'(t) + ku(t) = 0$.
2. Find any particular solution $u(t) = u_p(t)$ to $mu''(t) + cu'(t) + ku(t) = f(t)$, without worrying about initial conditions.
3. A general solution to $mu''(t) + cu'(t) + ku(t) = f(t)$ is then $u(t) = u_p(t) + u_h(t)$. The arbitrary constants in $u_h(t)$ can be adjusted to find the solution with the desired initial conditions.

We already know how to do Step 1. We'll consider how to accomplish Step 2 momentarily. But first, let's see why Step 3 works.

If $u_h(t)$ denotes a general solution to the homogeneous ODE, then

$$mu_h''(t) + cu_h'(t) + ku_h(t) = 0.$$

This general solution will be of the form $u_h(t) = c_1u_1(t) + c_2u_2(t)$ as in Definition 4.2.1, for appropriate basis functions $u_1(t)$ and $u_2(t)$, with c_1 and c_2 as arbitrary constants. Let $u_p(t)$ be any particular solution to the nonhomogeneous equation, so

$$mu_p''(t) + cu_p'(t) + ku_p(t) = f(t).$$

Consider the function $u(t) = u_h(t) + u_p(t)$. The function $u(t)$ satisfies the nonhomogeneous ODE, since

$$\begin{aligned} mu''(t) + cu'(t) + ku(t) &= m(u_h''(t) + u_p''(t)) + c(u_h'(t) + u_p'(t)) + k(u_h(t) + u_p(t)) \\ &= \underbrace{(mu_h''(t) + cu_h'(t) + ku_h(t))}_0 + \underbrace{(mu_p''(t) + cu_p'(t) + ku_p(t))}_{f(t)} \quad (4.47) \\ &= f(t), \end{aligned}$$

where we regrouped terms in the transition from the first line to the second. Notice that the linearity of differentiation and the ODE are both essential in carrying out the computation in (4.47). It follows that the function

$$u(t) = u_h(t) + u_p(t) = c_1 u_1(t) + c_2 u_2(t) + u_p(t)$$

satisfies $mu''(t) + cu'(t) + ku(t) = f(t)$ for any choice of c_1 and c_2 . In this case $u(t)$ is a general solution to the nonhomogeneous ODE, since c_1 and c_2 can be adjusted to obtain any desired initial conditions. Let us state this as a theorem.

Theorem 4.3.1 A general solution $u(t)$ to $mu''(t) + cu'(t) + ku(t) = f(t)$ can be obtained as

$$u(t) = u_h(t) + u_p(t), \quad (4.48)$$

where $u_h(t) = c_1 u_1(t) + c_2 u_2(t)$ is any general solution to the homogeneous equation $mu'' + cu' + ku = 0$ and $u_p(t)$ is any particular solution to the nonhomogeneous equation $mu'' + cu' + ku = f$.

■ **Example 4.13** Consider the ODE

$$u''(t) + 4u'(t) + 3u(t) = e^{-2t}$$

with initial conditions $u(0) = 1$ and $u'(0) = 4$. To apply Theorem 4.3.1 first use the procedure developed in Section 4.2 to compute a general solution to the homogeneous ODE $u'' + 4u' + 3u = 0$, namely $u(t) = c_1 e^{-t} + c_2 e^{-3t}$. A particular solution $u_p(t)$ to the nonhomogeneous ODE is $u_p(t) = -e^{-2t}$, as can be easily verified; again, we'll soon see how to construct $u_p(t)$. From Theorem 4.3.1 it follows that a general solution to $u''(t) + 4u'(t) + 3u(t) = e^{-2t}$ is

$$u(t) = u_h(t) + u_p(t) = c_1 e^{-t} + c_2 e^{-3t} - e^{-2t}. \quad (4.49)$$

The initial condition $u(0) = 1$ yields $c_1 + c_2 - 1 = 1$ and since $u'(t) = -c_1 e^{-t} - 3c_2 e^{-3t} + 2e^{-2t}$, $u'(0) = 4$ yields $-c_1 - 3c_2 + 2 = 4$. The solution for the constants is $c_1 = 4$ and $c_2 = -2$. The solution to the ODE with the desired initial conditions is then

$$u(t) = 4e^{-t} - 2e^{-3t} - e^{-2t}.$$

■

Reading Exercise 4.3.1 Verify that $u(t)$ as given in (4.49) satisfies $u''(t) + 4u'(t) + 3u(t) = e^{-2t}$.

Reading Exercise 4.3.2 According to Theorem 4.3.1, any particular solution $u_p(t)$ in (4.48) will yield a valid general solution. The function $u_p(t) = -e^{-2t} - 5e^{-t}$ is also a particular solution to the ODE in Example 4.13. (There are infinitely many other particular solutions.) Write out the general solution obtained from (4.48) with this choice of $u_p(t)$. Then adjust c_1 and c_2 to obtain $u(0) = 1$ and $u'(0) = 4$. Verify that we obtain exactly the same result as in Example 4.13.

The moral of Reading Exercise 4.3.2 is that any particular solution to the nonhomogeneous equation can be used to construct a general solution. Let's now consider a structured technique for producing a particular solution.

4.3.2 Finding a Particular Solution: Undetermined Coefficients

The central step in solving (4.3) is finding a particular solution $u = u_p(t)$. We will use the **method of undetermined coefficients**. This might also be aptly called the “method of educated guessing.” It is based on an informal observation that you may have made at some point in your mathematical experience: most elementary functions look like their own derivatives. The derivative of a polynomial is a polynomial. The derivative of an exponential function is an exponential function. The derivative of a sine or cosine is a cosine or sine. The same observation even applies to sums and products of these types of functions. This observation, when applied to the forcing function $f(t)$ in (4.45), can often be used to generate a particular solution. The essential idea is to seek a particular solution $u_p(t)$ that is of the same general form as the forcing function $f(t)$.

The best way to master this technique is to see it in action, so let’s consider a few examples.

■ **Example 4.14** Let us find a particular solution $u_p(t)$ of the ODE

$$u''(t) + 4u'(t) + 3u(t) = 6.$$

In this case the forcing function $f(t) = 6$ is a constant, which one might also think of as a zeroth-degree polynomial. Our guess for a particular solution $u_p(t)$ will be of this same form, a constant. But rather than trying a very specific guess like $u_p(t) = 5$ or $u_p(t) = \pi/2$ or such, let us try $u_p(t) = A$, where A is a constant to be determined. This gives some flexibility to adjust A in order to make this guess actually work. With $u_p(t) = A$ it follows that $u'_p = u''_p = 0$. As a result, when $u_p(t)$ is substituted into the ODE of interest the result is

$$3A = 6.$$

Then $A = 2$, so $u_p(t) = 2$ is a particular solution to $u'' + 4u' + 3u = 6$. ■

■ **Example 4.15** Let us find a particular solution $u_p(t)$ to

$$u''(t) + 4u'(t) + 3u(t) = 12 + 11t + 3t^2.$$

In this case $f(t) = 12 + 11t + 3t^2$ is a quadratic polynomial. The guess for $u_p(t)$ will be of the same form, $u_p(t) = A + Bt + Ct^2$, where the coefficients A, B , and C of the powers of t are constants to adjust to make this ansatz work. Then $u'_p(t) = B + 2Ct$ and $u''_p(t) = 2C$. Substituting $u_p(t)$ and these derivatives into the ODE yields

$$2C + 4(B + 2Ct) + 3(A + Bt + Ct^2) = 12 + 11t + 3t^2.$$

However, things are made much easier if terms with like powers of t on both sides are collected together, to obtain

$$(2C + 4B + 3A) + (8C + 3B)t + 3Ct^2 = 12 + 11t + 3t^2.$$

In order for the left and right sides above to be identical as functions of t the like powers of t on both sides must have the same coefficients. This yields

$$2C + 4B + 3A = 12, \quad 8C + 3B = 11, \quad 3C = 3.$$

The result is three equations in three unknowns, A, B , and C , with solution $C = 1, B = 1$, and $A = 2$. You should verify this. Then

$$u_p(t) = 2 + t + t^2$$

is a particular solution to $u''(t) + 4u'(t) + 3u(t) = 12 + 11t + 3t^2$. ■

This technique works for exponential forcing functions, too.

- **Example 4.16** Let us find a particular solution $u_p(t)$ to

$$u''(t) + 4u'(t) + 3u(t) = e^{-5t}.$$

Our guess will be an exponential function of the form $u_p(t) = Ae^{-5t}$; the undetermined coefficient A provides the flexibility needed to make the guess work in the ODE. For this choice of $u_p(t)$ we find $u'_p(t) = -5Ae^{-5t}$ and $u''_p(t) = 25Ae^{-5t}$. Substituting these into the ODE yields

$$(25 - 20 + 3)Ae^{-5t} = e^{-5t}$$

or $8A = 1$ after dividing both sides by e^{-5t} . This yields $A = 1/8$, so a particular solution is

$$u_p(t) = \frac{e^{-5t}}{8}.$$

Let's do one last example involving trigonometric functions.

- **Example 4.17** Let us find a particular solution $u_p(t)$ to

$$u''(t) + 4u'(t) + 3u(t) = \sin(2t).$$

Based on the above examples, it's tempting to try $u_p(t) = A \sin(2t)$. Then $u'_p(t) = 2A \cos(2t)$ and $u''_p(t) = -4A \sin(2t)$. Substitute these into the ODE and collect like $\sin(2t)$ and $\cos(2t)$ terms on the left to obtain

$$-A \sin(2t) + 4A \cos(2t) = \sin(2t).$$

Matching the sine terms on each side is easy, we merely need $-A = 1$, so $A = -1$. But the $4A \cos(2t)$ term on the left then becomes $-4 \cos(2t)$ with no match on the right side. You can convince yourself that no constant choice for A will work here. This ansatz has failed.

The problem is that the guess $u_p(t) = A \sin(2t)$ generates $\cos(2t)$ terms when differentiated, and there aren't any corresponding terms in the forcing function. Instead, let us try a guess

$$u_p(t) = A \sin(2t) + B \cos(2t)$$

with two adjustable constants. Then $u'_p(t) = 2A \cos(2t) - 2B \sin(2t)$ and $u''_p(t) = -4A \sin(2t) - 4B \cos(2t)$. In the ODE this becomes

$$-4A \sin(2t) - 4B \cos(2t) + 4(2A \cos(2t) - 2B \sin(2t)) + 3(A \sin(2t) + B \cos(2t)) = \sin(2t).$$

Again, group the $\sin(2t)$ and $\cos(2t)$ terms on the left and write the equation as

$$(-A - 8B) \sin(2t) + (8A - B) \cos(2t) = \sin(2t). \quad (4.50)$$

Thinking of the right side in (4.50) as $1 \sin(2t) + 0 \cos(2t)$ makes it clear that a particular solution will be obtained if A and B on the left side of (4.50) satisfy

$$-A - 8B = 1 \quad \text{and} \quad 8A - B = 0.$$

The solution to these two equations is $A = -1/65$ and $B = -8/65$. A particular solution is then

$$u_p(t) = -\frac{1}{65} \sin(2t) - \frac{8}{65} \cos(2t).$$

It should be pretty clear why this is called the “method of undetermined coefficients.” In each case we try an ansatz of the same general form as the forcing function $f(t)$, but with undetermined coefficients. We then substitute the ansatz into the ODE and adjust the coefficients as needed to obtain a solution.

Good Ansatzes

Not every forcing function $f(t)$ is amenable to using undetermined coefficients for finding a particular solution to $mu'' + cu' + ku = f$. Table 4.2 lists some common cases for $f(t)$ and corresponding guesses $u_p(t)$ that are usually, but not always, successful. In the table the functions p_n, P_n, q_n , and Q_n denote n th-degree polynomials in t . It's also worth noting that if u_1 is a particular solution to

$f(t)$	$u_p(t)$
a_0 (constant)	A (constant)
$p_n(t)$	$P_n(t)$
ae^{rt}	Ae^{rt}
$a\cos(\omega t) + b\sin(\omega t)$	$A\cos(\omega t) + B\sin(\omega t)$
$ae^{rt}\cos(\omega t) + be^{rt}\sin(\omega t)$	$Ae^{rt}\cos(\omega t) + Be^{rt}\sin(\omega t)$
$e^{rt}p_n(t)$	$e^{rt}P_n(t)$
$\cos(\omega t)p_n(t) + \sin(\omega t)q_n(t)$	$\cos(\omega t)P_n(t) + \sin(\omega t)Q_n(t)$

Table 4.2: Forcing functions $f(t)$ and reasonable ansatzes $u_p(t)$ for undetermined coefficients.

$mu'' + cu' + ku = f_1$ and u_2 is a particular solution to $mu'' + cu' + ku = f_2$, then by linearity $u_1 + u_2$ is a particular solution to $mu'' + cu' + ku = f_1 + f_2$.

Example: A Periodically-Driven Damped Harmonic Oscillator

Damped harmonic oscillators driven by a sinusoidal forcing function are extremely common. Let us consider an example.

■ **Example 4.18** Let's solve the linear second-order nonhomogeneous ODE

$$2u''(t) + 4u'(t) + 10u(t) = \cos(t)$$

with initial conditions $u(0) = 1$ and $u'(0) = -1$. First, the characteristic equation for the homogeneous ODE is $2r^2 + 4r + 10 = 0$ and has roots $-1 \pm 2i$. From this and (4.33) a general solution to the homogeneous ODE $2u''(t) + 4u'(t) + 10u(t) = 0$ is

$$u_h(t) = c_1 e^{-t} \cos(2t) + c_2 e^{-t} \sin(2t).$$

The next step is to find a particular solution $u_p(t)$ using the method of undetermined coefficients. Based on Table 4.2 an appropriate ansatz is $u_p(t) = A\cos(t) + B\sin(t)$. Substituting $u'_p(t) = -A\sin(t) + B\cos(t)$ and $u''_p(t) = -A\cos(t) - B\sin(t)$ into the nonhomogeneous ODE yields

$$(8A + 4B)\cos(t) + (-4A + 8B)\sin(t) = \cos(t),$$

after grouping the sine and cosine coefficients on the left. This last equation is satisfied if

$$8A + 4B = 1 \quad \text{and} \quad -4A + 8B = 0.$$

The solution is $A = 1/10$ and $B = 1/20$. Therefore a particular solution to the nonhomogeneous ODE is $u_p(t) = \frac{1}{10}\cos(t) + \frac{1}{20}\sin(t)$ and, from Theorem 4.3.1, a general solution to the nonhomogeneous ODE is

$$u(t) = u_p(t) + u_h(t) = \frac{1}{10}\cos(t) + \frac{1}{20}\sin(t) + c_1 e^{-t} \cos(2t) + c_2 e^{-t} \sin(2t).$$

We can construct the solution with the desired initial data by using this general solution to find $u(0) = 1/10 + c_1 = 1$ and $u'(0) = -c_1 + 2c_2 + 1/20 = -1$. These two equations have simultaneous solution $c_1 = 9/10$ and $c_2 = -3/40$. The full solution is then

$$u(t) = \frac{1}{10}\cos(t) + \frac{1}{20}\sin(t) + \frac{9}{10}e^{-t}\cos(2t) - \frac{3}{40}e^{-t}\sin(2t).$$

Compare $u(t)$ here to the solution (4.46) obtained in Example 4.12. We see the same solution structure, a periodic portion at the same frequency as the driving force $f(t) = \cos(t)$ and a transient portion that decays to zero over time. ■

Periodic and Transient Solutions

Example 4.18 illustrates a common and important physical situation in which a damped oscillator is driven by a sinusoidal forcing function of the form $f(t) = C\cos(\omega t) + D\sin(\omega t)$ for constants C, D , and a frequency ω . The ODE that governs this situation is $mu''(t) + cu'(t) + ku(t) = f(t)$. A particular solution $u_p(t)$ can be found using the method of undetermined coefficients, and will take the form $u_p(t) = A\cos(\omega t) + B\sin(\omega t)$ for suitable constants A and B . This solution $u_p(t)$ is periodic and at the same frequency as the driving force. A general solution for the homogeneous ODE $mu''(t) + cu'(t) + ku(t) = 0$ is $u_h(t) = c_1e^{r_1t} + c_2e^{r_2t}$, where r_1 and r_2 are the roots to the characteristic equation $mr^2 + cr + k = 0$. From Theorem 4.3.1 it follows that a general solution for the motion of a damped oscillator driven sinusoidally at a frequency ω is of the form

$$u(t) = \underbrace{c_1e^{r_1t} + c_2e^{r_2t}}_{\text{transient}} + \underbrace{A\cos(\omega t) + B\sin(\omega t)}_{\text{periodic}}. \quad (4.51)$$

In this situation the solution will always consist of two distinct pieces as labeled in (4.51):

1. A **transient** portion that decays in time, since r_1 and r_2 are negative real numbers, or if complex, have negative real part. Think of the transient portion of the solution as stemming primarily from the system's inherent physical properties and the parameters m, c , and k .
2. A **periodic** portion that oscillates and never decays. Think of this part of the solution as stemming from the driving force $f(t)$. The periodic portion remains as long as $f(t)$ is present. The periodic portion of the solution is also referred to as the **steady-state** portion of the solution.

More Undetermined Coefficients Examples

Here are some more complete examples that illustrate how to find a solution to a nonhomogeneous ODE $mu'' + cu' + ku = f$ for a given function f , construct a general solution to the ODE, and determine the solution that has the specific initial conditions.

■ **Example 4.19** A building is modeled by the driven spring-mass-damper ODE (4.45) with $m = 10^4$, $c = 2 \times 10^5$, and $k = 10^7$, with $u(t)$ as the displacement of the mass or roof. At time $t = 0$ we have $u(0) = 0$ and $u'(0) = 0$ when a force $f(t) = 10^5 e^{-5t} \sin(30t)$ is applied to the mass (think of it as a brief earthquake). Let us find the response $u(t)$ of the mass.

The roots to the characteristic equation $10^4 r^2 + (2 \times 10^5)r + 10^7 = 0$ are $-10 \pm 30i$, so from (4.33) a general solution to the homogeneous version of the ODE is

$$u_h(t) = c_1e^{-10t} \cos(30t) + c_2e^{-10t} \sin(30t).$$

Based on the fact that $f(t) = 10^5 e^{-5t} \sin(30t)$ and the information in Table 4.2, we now seek a particular solution of the form

$$u_p(t) = Ae^{-5t} \cos(30t) + Be^{-5t} \sin(30t).$$

Inserting $u_p(t)$ into the ODE and collecting like terms produces

$$250000e^{-5t}(A + 12B)\cos(30t) + 250000e^{-5t}(-12A + B) = 10^5 e^{-5t} \sin(30t).$$

Divide both sides by $10^5 e^{-5t}$ to obtain

$$\frac{5}{2}(A + 12B)\cos(30t) + \frac{5}{2}(-12A + B)\sin(30t) = \sin(30t).$$

Both sides must be identical as functions of t (think of the right side above as $1 \sin(30t) + 0 \cos(30t)$), so

$$\frac{5}{2}(A + 12B) = 0 \quad \text{and} \quad \frac{5}{2}(-12A + B) = 1$$

which yields, after a bit of algebra, $A = -24/725$ and $B = 2/725$. Thus a particular solution to $mu'' + cu' + ku = f$ here is given by

$$u_p(t) = -\frac{24}{725}e^{-5t} \cos(30t) + \frac{2}{725}e^{-5t} \sin(30t).$$

From Theorem 4.3.1 a general solution is $u(t) = u_p(t) + u_h(t)$, or

$$u(t) = -\frac{24}{725}e^{-5t} \cos(30t) + \frac{2}{725}e^{-5t} \sin(30t) + c_1 e^{-10t} \cos(30t) + c_2 e^{-10t} \sin(30t). \quad (4.52)$$

The conditions $u(0) = 0$ and $u'(0) = 0$ applied to (4.52) lead to equations

$$\begin{aligned} -24/725 + c_1 &= 0 \\ -10c_1 + 30c_2 + 36/145 &= 1. \end{aligned}$$

This yields $c_1 = 24/725$ and $c_2 = 2/725$. The solution to the ODE that models this situation is thus

$$u(t) = -\frac{24}{725}e^{-5t} \cos(30t) + \frac{2}{725}e^{-5t} \sin(30t) + \frac{24}{725}e^{-10t} \cos(30t) + \frac{2}{725}e^{-10t} \sin(30t).$$

Figure 4.14 shows the response $u(t)$ on the time interval $0 \leq t \leq 2$. ■

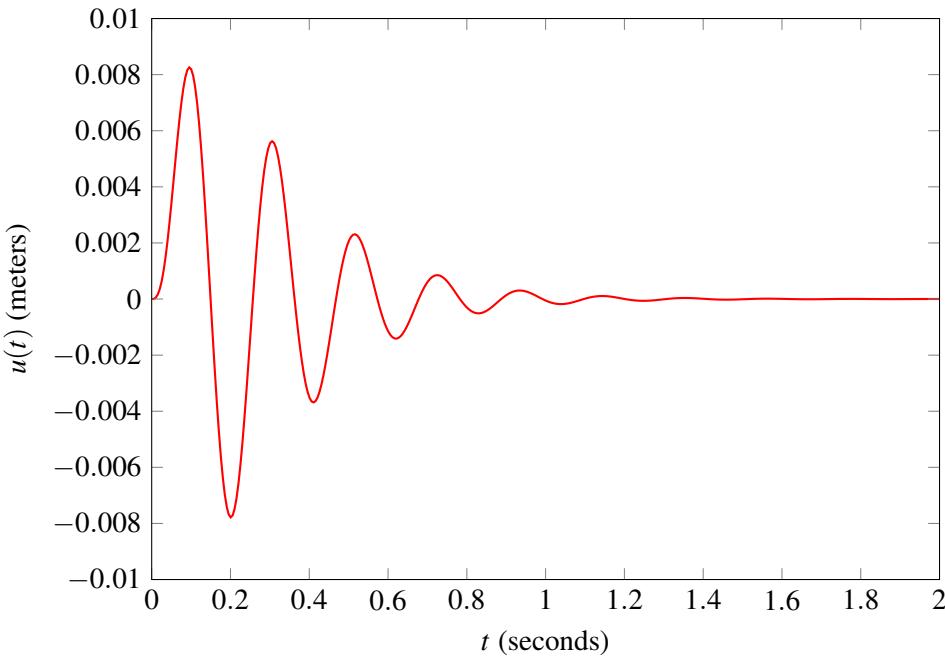


Figure 4.14: Solution $u(t)$ to $10^4u''(t) + (2 \times 10^5)u'(t) + 10^7u(t) = 10^5 \sin(30t)$ with $u(0) = u'(0) = 0$.

■ **Example 4.20** Recall Example 4.1 in which we modeled the front shock absorber on a mountain bike. When the front wheel is in contact with the ground (and so not moving vertically, while

supporting the bike and the rider's mass), the relevant ODE for the displacement $u(t)$ of the shock-damper is given by (4.6), reproduced here for convenience:

$$46u''(t) + 1700u'(t) + 15000u(t) = -450.8. \quad (4.53)$$

In that example it was assumed that the mass of the rider and bicycle combined is 92 kg, half of which is supported by the front shock, which accounts for the 46 coefficient in front of $u''(t)$ in (4.53). The spring constant is assumed to be 15000 newtons per meter and the viscous damping coefficient is 1700 newton-seconds per meter. The constant forcing function on the right side of (4.53) is $-mg = -450.8$ newtons, where gravitational acceleration is $g = 9.8$ meters per second squared and $m = 46$ kg.

Let us work out a general solution to (4.53). First, the characteristic equation for the homogeneous ODE $46u''(t) + 1700u'(t) + 15000u(t) = 0$ is

$$46r^2 + 1700r + 15000 = 0,$$

and it has roots $r_1 \approx -14.56$ and $r_2 \approx -22.40$. The roots are real and distinct, so this system is overdamped. A general solution to the homogeneous ODE is therefore

$$u_h(t) = c_1 e^{-14.56t} + c_2 e^{-22.40t}.$$

Since the forcing function is constant, a particular solution of the form $u_p(t) = u^*$ (constant) is appropriate. The ODE (4.53) then becomes $15000u^* = -450.8$, from which it follows that $u_p(t) = u^* = -0.03$ meters. From Theorem 4.3.1 a general solution to (4.53) is

$$u(t) = -0.03 + c_1 e^{-14.56t} + c_2 e^{-22.40t}. \quad (4.54)$$

Let's use this result to do some practical analysis. Suppose the rider of this mountain bike rides off a ledge or jump that's 1.5 meters in height. In the air there is no force on the shock and we expect that the shock displacement rapidly returns to the condition $u(t) = 0$, since here the homogeneous ODE holds and the solution decays very rapidly. Assume that at the moment of impact, $t = 0$, exactly half the weight of the bike and rider is absorbed by the shock. For $t > 0$ the front wheel is vertically motionless and in contact with the ground, hence the ODE (4.53) governs the shock's behavior. At the instant that the bike impacts the ground we have $u(0) = 0$, since the shock was not compressed in the air. The behavior of the shock for $t > 0$ can be determined from knowledge of $u'(0)$.

A standard physics result shows that an object that falls from a distance h under gravitational acceleration hits the ground with speed $\sqrt{2gh}$, if air resistance is negligible. Therefore, we estimate that the bike hits the ground at a speed of $v_0 \approx -5.42$ meter per second, negative because the bike is falling. Thus we take $u'(0) = v_0$. The initial data $u(0) = 0$ and $u'(0) = -5.42$, along with (4.54), leads to equations

$$-0.03 + c_1 + c_2 = 0 \quad \text{and} \quad -14.56c_1 - 22.40c_2 = -5.42.$$

The solution is $c_1 \approx -0.606$, $c_2 \approx 0.636$, and so the displacement $u(t)$ of the shock is

$$u(t) \approx -0.03 - 0.606e^{-14.56t} + 0.636e^{-22.40t}.$$

A plot of this function is shown in Figure 4.15. ■

Reading Exercise 4.3.3 Compute the maximum compression of the shock that occurs in Example 4.20, to three significant figures. Suppose the bike's front shock has a range of motion of 140 mm before bottoming out (the shock has reached the end of its travel and can no longer compress). Would that be a problem in this scenario?

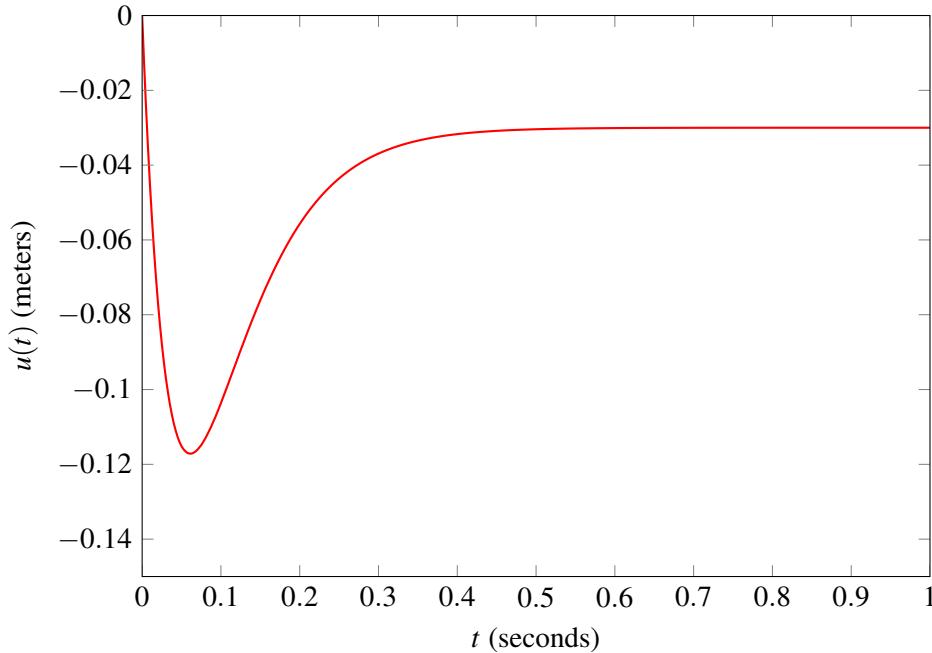


Figure 4.15: Displacement $u(t)$ of bicycle front shock after 1.5 meter jump.

4.3.3 When the Guess Fails

Under some circumstances the guesses in Table 4.2 fail, but it is usually possible to modify a failed guess to make it work. Consider the following example.

■ **Example 4.21** Let's look for a particular solution to

$$u''(t) + 4u'(t) + 3u(t) = e^{-t}. \quad (4.55)$$

Based on Table 4.2 (and more generally, the intuition that the ansatz should look like the forcing function) we try a particular solution of the form $u_p(t) = Ae^{-t}$. Inserting this into (4.55) leads to $0 = e^{-t}$, which is never true. The problem is that if $u_p(t) = Ae^{-t}$, then $u_p(t)$ is a solution to the homogeneous problem $u'' + 4u' + 3u = 0$, so no choice of A can ever work. ■

What can be done in the situation of Example 4.21? Let's take a cue from the double root case for the homogeneous equation. Specifically, instead of trying $u_p(t) = Ae^{-t}$, let's allow A to be a function of t , and so try a guess of the form

$$u_p(t) = A(t)e^{-t}.$$

Compute the derivatives

$$\begin{aligned} u'_p(t) &= -A(t)e^{-t} + A'(t)e^{-t} \\ u''_p(t) &= A(t)e^{-t} - 2A'(t)e^{-t} + A''(t)e^{-t}, \end{aligned}$$

then substitute this information into (4.55). Collect terms to find $e^{-t}(A''(t) + 2A'(t)) = e^{-t}$, or

$$A''(t) + 2A'(t) = 1. \quad (4.56)$$

Any choice for $A(t)$ that satisfies this equation will work, and there are many. One easy choice is to take $A(t)$ as a linear function of t , so $A''(t) = 0$. Equation (4.56) then becomes $2A'(t) = 1$ with solution $A(t) = t/2$. This means that a particular solution to (4.55) is

$$u_p(t) = A(t)e^{-t} = \frac{t}{2}e^{-t}.$$

This particular solution can then be used to form a general solution to (4.55) and obtain any desired initial conditions.

Notice that the particular solution $u_p(t) = te^{-t}/2$ that actually works is a minor modification of the guess Ae^{-t} suggested by Table 4.2: we could have started by multiplying the expected guess Ae^{-t} by t and instead tried $u_p(t) = Ate^{-t}$ in the ODE. Multiplying the standard guess by t frequently yields an ansatz that works.

General Advice for a Failed Guess

The standard guesses in the method of undetermined coefficients from Table 4.2 are guaranteed to fail whenever the forcing function $f(t)$ is itself a solution to the homogeneous version of the ODE $mu'' + cu' + ku = 0$. More generally, if the characteristic equation $mr^2 + cr + k = 0$ has roots r_1 and r_2 and the forcing function contains terms of the form $t^n e^{r_1 t}$ or $t^n e^{r_2 t}$ for $n \geq 0$, the standard guess from Table 4.2 will probably fail to yield a particular solution. Also note that since $\cos(\omega t)$ and $\sin(\omega t)$ are linear combinations of $e^{i\omega t}$ and $e^{-i\omega t}$, this observation also applies to functions $f(t)$ that involve $\sin(\omega t)$ or $\cos(\omega t)$.

Nevertheless, for a given forcing function $f(t)$ with ansatz $\phi(t)$ from Table 4.2, a modified guess can usually be made to work by allowing any undetermined coefficients in $\phi(t)$ be undetermined functions of t . Substituting such a guess into the ODE often leads to choices for these undetermined functions that works. The interested reader should consult [24] for a more thorough treatment of undetermined coefficients. We provide one last example to illustrate how simple the technique of undetermined coefficients is in practice, with a bit of experimentation and perseverance.

■ **Example 4.22** Let us find a particular solution to

$$2u''(t) + 4u'(t) + 10u(t) = e^{-t} \cos(2t), \quad (4.57)$$

then use this to construct a general solution to this ODE, and find a solution with initial data $u(0) = 2$ and $u'(0) = 4$.

The characteristic equation for the homogeneous version of (4.57) is $2r^2 + 4r + 10 = 0$ with roots $r = -1 \pm 2i$. A real-valued general solution is therefore

$$u_h(t) = c_1 e^{-t} \cos(2t) + c_2 e^{-t} \sin(2t).$$

Next we seek a particular solution $u_p(t)$. Based on Table 4.2 we might expect that $u_p(t) = Ae^{-t} \cos(2t) + Be^{-t} \sin(2t)$ will work, but this choice for $u_p(t)$ is a solution to the homogeneous equation and can never work—it leads to $0 = e^{-t} \cos(2t)$. Instead we'll replace this guess with

$$u_p(t) = A(t)e^{-t} \cos(2t) + B(t)e^{-t} \sin(2t). \quad (4.58)$$

Inserting $u = u_p$ into (4.57) and collecting like terms yields

$$(2A''(t) + 8B'(t))e^{-t} \cos(2t) + (-8A'(t) + 2B''(t))e^{-t} \sin(2t) = e^{-t} \cos(2t).$$

We will obtain a specific solution if $A(t)$ and $B(t)$ can be chosen so that

$$2A''(t) + 8B'(t) = 1 \quad \text{and} \quad -8A'(t) + 2B''(t) = 0.$$

This is a system of two coupled ODEs for two unknown functions, a topic for a later chapter, but all we need is one solution, any solution. This is where a little creative experimentation helps, and one possibility is to take $A(t) = 0$ so the equations become $8B'(t) = 1$ and $2B''(t) = 0$. We can then take $B(t)$ as a linear function of t to make $B''(t) = 0$, and then $8B'(t) = 1$ has a solution $B(t) = t/8$. Using these choices in (4.58) shows that

$$u_p(t) = \frac{te^{-t} \sin(2t)}{8}$$

is a solution to (4.57).

A general solution is then

$$u(t) = u_p(t) + u_h(t) = \frac{te^{-t} \sin(2t)}{8} + c_1 e^{-t} \cos(2t) + c_2 e^{-t} \sin(2t).$$

The initial condition $u(0) = 2$ yields $c_1 = 2$, while $u'(0) = 4$ forces $-c_1 + 2c_2 = 4$, and ultimately $c_2 = 3$. The solution to (4.57) with $u(0) = 2$ and $u'(0) = 4$ is

$$u(t) = \frac{te^{-t} \sin(2t)}{8} + 2e^{-t} \cos(2t) + 3e^{-t} \sin(2t).$$

■

A Few Remarks on Higher-Order or Variable-Coefficient ODEs

We've now seen how to solve a variety of first-order equations, as well as linear, constant-coefficient second-order equations. The observant reader may wonder about higher-order equations, or linear equations that do not have constant coefficients, or even nonlinear second or higher-order equations.

Higher-order linear, constant-coefficient ODEs can be handled in much the same manner as second-order ODEs. In the homogeneous case an ansatz $u(t) = e^{rt}$ provides a path forward. To illustrate, consider the third-order ODE $u'''(t) + u''(t) - 2u(t) = 0$. An ansatz of the form $u(t) = e^{rt}$ leads to the characteristic equation $r^3 + r^2 - 2 = 0$ with solutions $r = 1, r = -1 + i, r = -1 - i$. Linearity allows us to construct a general solution of the form

$$u(t) = c_1 e^{-t} + c_2 e^{(-1+i)t} + c_3 e^{(-1-i)t}.$$

With three initial conditions of the form $u(0) = u_0, u'(0) = u'_0$, and $u''(0) = u''_0$, we can solve for c_1, c_2 , and c_3 . A real-valued general solution can also be constructed in a manner similar to that which led to (4.33). In general an n th-order ODE leads to an n th-degree characteristic equation whose roots we must find. Roots of multiplicity higher than 1 add some complication and give rise to solution terms like $t^m e^{rt}$ for $m > 0$.

The nonhomogeneous case can be handled using undetermined coefficients in much the same way as was done for second-order ODEs. Alternatively, second and higher ODEs can be converted into systems of first-order ODEs and analyzed using techniques we will develop in Chapters 6 and 7.

Linear ODEs with variable coefficients can often be analyzed using series methods, in which we posit that the solution $u(t)$ has a series expansion $u(t) = a_0 + a_1 t + a_2 t^2 + \dots$, substitute this ansatz into the ODE, and then deduce information about the coefficients a_k ; see [24] for more on this topic. Second and higher-order nonlinear ODEs form a less unified subject, with techniques specific to small classes of problems. However, analysis may be aided by the ideas in Chapter 7.

4.3.4 Exercises

Exercise 4.3.1 For each ODE in parts (a)-(w):

- Find a general solution $u_h(t)$ to the homogeneous version of the ODE.
- Use the method of undetermined coefficients to find a particular solution $u_p(t)$ to the nonhomogeneous ODE. This answer is not unique.
- Write out a general solution to the nonhomogeneous ODE and use it to obtain initial conditions $u(0) = 2$ and $u'(0) = 3$.

- (a) $u''(t) + 9u'(t) + 20u(t) = 2e^{-3t}$
 (b) $4u''(t) + 16u'(t) + 12u(t) = -32e^t$

- (c) $u''(t) + 8u'(t) + 32u(t) = 32$
 (d) $4u''(t) + 12u'(t) + 8u(t) = 8$
 (e) $u''(t) + 4u'(t) + 3u(t) = 9t$
 (f) $u''(t) + 4u'(t) + 8u(t) = 10\sin(2t)$
 (g) $3u''(t) + 15u'(t) + 12u(t) = 10\sin(3t)$
 (h) $3u''(t) + 12u'(t) + 39u(t) = 21e^{-2t}\cos(4t)$
 (i) $4u''(t) + 12u'(t) + 9u(t) = t + t^2$
 (j) $3u''(t) + 15u'(t) + 12u(t) = 12te^{-2t}$
 (k) $4u''(t) + 28u'(t) + 40u(t) = 16t^2e^{-3t}$
 (l) $u''(t) + u(t) = t \sin(t)$
 (m) $u''(t) + 2u'(t) + 10u(t) = 10e^{-2t}$
 (n) $3u''(t) + 6u'(t) + 6u(t) = 6e^{-t}\cos(t)$
 (o) $3u''(t) + 12u'(t) + 39u(t) = 27te^{-2t}$
 (p) $3u''(t) + 24u'(t) + 96u(t) = 96$
 (q) $u''(t) + 5u'(t) + 4u(t) = 10\sin(2t)$
 (r) $4u''(t) + 24u'(t) + 20u(t) = 8e^{-2t}$
 (s) $u''(t) + 7u'(t) + 10u(t) = 15 + 25t$
 (t) $2u''(t) + 4u'(t) + 10u(t) = 6e^{-t}\cos(t)$
 (u) $u''(t) + 2u'(t) + 2u(t) = 25t \cos(t)$
 (v) $u''(t) + 4u'(t) + 5u(t) = 40\sin(5t)$
 (w) $u''(t) + u(t) = t$

Exercise 4.3.2 For each ODE in parts (a)-(i):

- Find a general solution $u_h(t)$ to the homogeneous version of the ODE.
- Use the method of undetermined coefficients to find a particular solution $u_p(t)$ to the nonhomogeneous ODE. However, in each of these the standard guess fails. Modify it appropriately, noting that the answer is not unique.
- Write out a general solution to the nonhomogeneous ODE and use it to obtain the specific solution with initial conditions $u(0) = 2$ and $u'(0) = 3$.

- (a) $u''(t) + 9u'(t) + 20u(t) = 2e^{-4t}$
 (b) $4u''(t) + 24u'(t) + 20u(t) = 8e^{-t}$
 (c) $4u''(t) + 16u'(t) + 12u(t) = 8e^{-3t}$
 (d) $u''(t) + u(t) = \cos(t)$
 (e) $u''(t) + 2u'(t) + 2u(t) = 2e^{-t}\sin(t)$
 (f) $u''(t) + 2u'(t) + 10u(t) = e^{-t}\sin(3t)$
 (g) $u''(t) + 4u'(t) + 8u(t) = 16e^{-2t}\cos(2t)$
 (h) $u''(t) + 4u'(t) + 4u(t) = te^{-2t}$
 (i) $u''(t) + u(t) = \sin(t)$

Exercise 4.3.3 Consider the ODE $mu''(t) + cu'(t) + ku(t) = e^{at}$ and suppose that a is not a root of this ODE's characteristic equation. Show that a guess of the form $u_p(t) = Ae^{at}$ for finding a particular solution will always work. Hint: just substitute $u_p(t)$ into the ODE and show you can always find A .

Exercise 4.3.4 Consider the ODE $mu''(t) + cu'(t) + ku(t) = f(t)$, and suppose that $k \neq 0$.

- Suppose $f(t) = a_0$ is constant. Show that an ansatz of the form $u_p(t) = A_0$ will yield a particular solution.
- Suppose $f(t) = a_0 + a_1 t$. Show that an ansatz of the form $u_p(t) = A_0 + A_1 t$ will yield a particular solution.
- Suppose $f(t)$ is an n th-degree polynomial. Will taking $u_p(t)$ as an n th-degree polynomial with undetermined coefficients always yield a particular solution? Why?

Exercise 4.3.5

- Redo the solution for the ODE (4.53) in the bike shock absorber Example 4.20, but change the damping constant from $c = 1700$ to $c = 10^4$, so the system is heavily overdamped. Plot the solution and redo Reading Exercise 4.3.3 with this parameters. What disadvantage might such a value of c have for the rider?
- Redo the solution for the ODE (4.53) in the bike shock absorber Example 4.20 but change the damping constant from $c = 1700$ to $c = 1200$. Show that the system is now underdamped. Plot the solution and redo Reading Exercise 4.3.3. What disadvantage might an underdamped system have for the rider?

Exercise 4.3.6 Recall the Hill-Keller ODE $v'(t) = P - kv(t)$ with initial condition $v(t_0) = 0$. We solved this ODE for $v(t)$ and then computed the sprinter's position $x(t)$ from $x'(t) = v(t)$ with initial condition $x(t_0) = 0$. However, if we pose the problem in terms of $x(t)$ directly, the Hill-Keller ODE becomes

$$x''(t) = P - kx'(t), \quad (4.59)$$

with initial conditions $x(t_0) = 0$ and $x'(t_0) = 0$.

- Equation (4.59) is a second-order, linear, constant-coefficient nonhomogeneous ODE. Write out its characteristic equation for the relevant homogeneous equation and find its roots. (Be careful—the constant k here appears in front of x' , not x .) One of the roots depends on k .
- Write out a general solution $x_h(t)$ to the homogeneous ODE.
- Find a particular solution $x_p(t)$ to (4.59).
- Find the solution with initial data $x(t_0) = 0$ and $x'(t_0) = 0$, and verify that we obtain the same result as in (3.39).

Exercise 4.3.7 Consider a vibration isolation table as modeled in Example 4.2 and the ODE (4.9). Suppose that the mass of the tabletop is $m = 100$ kg, at a nominal height of $L_0 = 1$ meter, and the isolation leg has spring constant $k = 10^4$ newtons per meter and damping constant $c = 2000$ newtons per meter second. Use $g = 9.8$ meters per second squared. Let $y(t)$ denote the height of the table as a function of time.

- Find the position $y(t) = y_{eq}$ of the table if only gravity acts on the tabletop (and $d(t) = 0$).
- Suppose the ground begins to vibrate according to $d(t) = 10^{-4} \cos(40\pi t)$ (20 hertz) at time $t = 0$. If the table has initial data $y(0) = y_{eq}$ and $y'(0) = 0$, find the motion $y(t)$ of the table. Plot $y(t)$ for $0 \leq t \leq 2$. What is the amplitude of the vibration of the table top

for $t > 1$? How does this compare to the amplitude of $d(t)$? Does the table effectively dampen this motion?

- (c) Consider an alternate scenario in which $d(t) = 0$ and the table is at its equilibrium position $y(t) = y_{eq}$ for $t < 0$. At time $t = 0$ a clumsy researcher drops something on the table and imparts an initial velocity $y'(0) = -0.1$ meter per second to the tabletop. Find the motion of the tabletop and plot this motion for $0 \leq t \leq 2$.

Exercise 4.3.8 An RLC circuit in the single-loop configuration of Figure 4.4 has an inductor with inductance $L = 0.1$ henries, resistance $R = 20$ ohms, and capacitance $C = 10^{-4}$ farads. The voltage source is $V(t) = 5$ volts. At $t = 0$ the capacitor is uncharged and no current flows in the circuit.

- Write out the appropriate nonhomogeneous ODE to model this circuit, with $q(t)$, the charge on the capacitor, as the dependent variable. What are the initial conditions for $q(t)$?
- Find a general solution $q_h(t)$ to the homogeneous version of the ODE. Based on your solution, is this system under, over, or critically damped?
- Use the method of undetermined coefficients to find a particular solution $q_p(t)$ to the nonhomogeneous ODE.
- Use your work from parts (b) and (c) to find a general solution to the nonhomogeneous ODE, and then find the solution with the required initial conditions.
- Plot the solution $q(t)$ on the range $0 \leq t \leq 0.1$ seconds. Plot the current flowing through the circuit on the same time range.

4.4 Resonance

Resonance is a phenomenon in which the response of a periodically forced spring-mass or similar system depends greatly on the frequency of the driving force. Sometimes resonance is undesirable, as in a building in an earthquake. Sometimes resonance is an essential part of how the system functions, such as in a classic RLC tuner circuit in a radio; see the project “Stay Tuned—RLC Circuits and Radio Tuning” in Section 4.6.

4.4.1 An Example of Resonance

In Example 4.12 of the last section we considered a single-story building in the presence of an earthquake. The building was modeled as a spring-mass-dashpot system $mu''(t) + cu'(t) + ku(t) = f(t)$ with $m = 5000$ kg, $k = 5 \times 10^5$ newtons per meter, $c = 10^4$ newtons per meter per second, and forcing function $f(t) = 10^4 \sin(5t)$ newtons. This forcing function represents a frequency of $5/(2\pi) \approx 0.8$ hertz at an amplitude of 10^4 newtons. With initial conditions $u(0) = 0$ and $u'(0) = 0$ the solution was

$$u(t) = \underbrace{-0.0128e^{-t} \sin(9.95t) + 0.0035e^{-t} \cos(9.95t)}_{\text{transient}} + \underbrace{0.0262 \sin(5t) - 0.0035 \cos(5t)}_{\text{periodic}}. \quad (4.60)$$

This function was graphed in Figure 4.13. The transient portion of the solution dies out rather rapidly, diminishing to less than one percent of its initial value within 5 seconds. The periodic portion on the right in (4.60) remains so long as the forcing $f(t)$ is active. This periodic response is sinusoidal, at exactly the same frequency as the driving function $f(t)$, and has an amplitude of $\sqrt{(-0.0035)^2 + 0.0262^2} \approx 0.0262$ meters (see Exercise 4.2.9).

Consider now how the building's response will change if the driving force is $f(t) = 10^4 \sin(10t)$. This driving force has the same amplitude as the previous case, 10^4 newtons, but at a different frequency, about 1.59 hertz. The solution with $u(0) = u'(0) = 0$ can be found using the techniques of Section 4.3 and is

$$u(t) = \underbrace{0.01e^{-t} \sin(9.95t) + 0.1e^{-t} \cos(9.95t)}_{\text{transient}} - \underbrace{0.1 \cos(10t)}_{\text{periodic}}. \quad (4.61)$$

The response of the building again has a transient portion, but as that portion decays exponentially, the solution consists primarily of the periodic term on the right in (4.61). This periodic response is at the same frequency as $f(t)$, but has a magnitude of 0.1 meters for the building's displacement. This is about a four-fold larger displacement than that induced by $f(t) = 10^4 \sin(5t)$, caused simply by changing the driving frequency from 0.8 to 1.6 hertz.

It's also instructive to examine the acceleration induced by these driving forces. The acceleration for each forcing function is graphed in Figure 4.16. For $f(t) = 10^4 \sin(5t)$ the initial acceleration peaks at about 1 meter per second squared, roughly 0.1 g's, but settles to a periodically varying value of around 0.7 meters per second squared, around 0.07 g. But with $f(t) = 10^4 \sin(10t)$ the acceleration assumes a longer-term oscillation with values around 10 meters per second squared, over 1 g. This is a far more destructive situation.

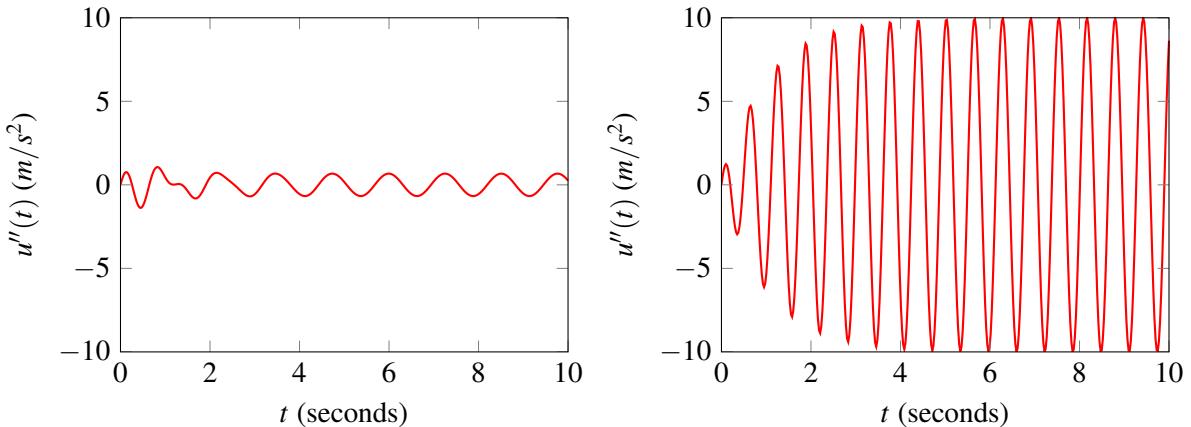


Figure 4.16: Left panel: acceleration $u''(t)$ for the forced oscillator $5000u'' + (10^4u' + (5 \times 10^5)u = f(t)$ with $f(t) = 10^4 \sin(5t)$. Right panel: same, but with $f(t) = 10^4 \sin(10t)$.

Despite the fact that in each case the sinusoidal driving function has amplitude 10^4 newtons, driving this system at 1.59 hertz elicits a much more vigorous response from the building, as compared to 0.8 hertz. By no small coincidence, the undriven version of this system has a natural frequency of about $10/(2\pi) \approx 1.59$ hertz.

4.4.2 Periodic Forcing

In Section 4.3.2 we considered the response of a damped harmonic oscillator to sinusoidal forcing. In particular, equation (4.51) shows that the response of the oscillator consists of a transient portion that decays in time and a periodic or steady-state portion that oscillates at the same frequency as the driving force. Let's extend that analysis to obtain a more detailed understanding of how the periodic portion of the solution depends on the driving force, as a prelude to understanding the phenomenon called "resonance."

We return to the driven harmonic oscillator (4.3)

$$mu''(t) + cu'(t) + ku(t) = f(t), \quad (4.62)$$

reproduced here for convenience. Suppose the forcing function $f(t)$ is sinusoidal and of the form

$$f(t) = C \sin(\omega t + \phi), \quad (4.63)$$

where C , ω , and ϕ are specified constants. We assume $\omega \geq 0$. The function $f(t)$ in (4.63) is sinusoidal with C as the amplitude, ω the frequency, and ϕ acting as a phase shift. As shown in Exercise 4.2.9, such a function $f(t)$ can also be written in the form

$$f(t) = A_1 \sin(\omega t) + A_2 \cos(\omega t) \quad (4.64)$$

by taking $A_1 = C \cos(\phi)$ and $A_2 = C \sin(\phi)$. We will work with whichever form, (4.63) or (4.64), is most convenient at the time.

As was discussed in the analysis leading up to (4.51), the general solution to (4.62) with $f(t)$ as in (4.63) or (4.64) is of the form

$$\begin{aligned} u(t) &= u_h(t) + u_p(t) \\ &= \underbrace{c_1 u_1(t) + c_2 u_2(t)}_{u_h(t), \text{ transient}} + \underbrace{A \cos(\omega t) + B \sin(\omega t)}_{u_p(t), \text{ periodic}} \end{aligned}$$

where $u_h(t) = c_1 u_1(t) + c_2 u_2(t)$ is the general solution to the homogeneous ODE $mu'' + cu' + ku = 0$ and $u_p(t)$ is a particular solution to the nonhomogeneous ODE, of the given form. If $c > 0$ then both $u_1(t)$ and $u_2(t)$ involve decaying exponentials and quickly approach zero as t increases, so only the periodic portion $u_p(t)$ of the solution remains. The periodic response satisfies

$$mu_p''(t) + cu_p'(t) + ku_p(t) = C \sin(\omega t + \phi) \quad (4.65)$$

and is of the form

$$u_p(t) = A \cos(\omega t) + B \sin(\omega t), \quad (4.66)$$

for a suitable choice of A and B . The function $u_p(t)$ is sinusoidal at frequency ω and has amplitude $\sqrt{A^2 + B^2}$. Of particular interest is how this amplitude depends on the parameters C , ω , and ϕ .

Reading Exercise 4.4.1 Consider the periodically forced ODE $u''(t) + 2u'(t) + 10u(t) = \sin(\omega t)$, where $\omega > 0$ is some driving frequency. Verify that

$$u(t) = \underbrace{c_1 e^{-t} \cos(3t) + c_2 e^{-t} \sin(3t)}_{\text{transient}} + \underbrace{A \cos(\omega t) + B \sin(\omega t)}_{\text{periodic}}$$

provides a general solution with

$$A = -\frac{2\omega}{\omega^4 - 16\omega^2 + 100} \quad \text{and} \quad B = \frac{10 - \omega^2}{\omega^4 - 16\omega^2 + 100}.$$

Then show that the amplitude $\sqrt{A^2 + B^2}$ of the periodic portion of the solution is $1/\sqrt{\omega^4 - 16\omega^2 + 100}$. Plot this amplitude as a function of ω for $0 \leq \omega \leq 10$. Explain what this plot tells you about the response of the system, in particular, the amplitude of the periodic portion, for various driving frequencies.

Computing the Periodic Response

Inserting $u_p(t)$ in the form (4.66) into the ODE (4.65) leads to

$$\begin{aligned} &(-Am\omega^2 + Bc\omega + Ak) \cos(\omega t) + (-Bm\omega^2 - Ac\omega + Bk) \sin(\omega t) \\ &= C \sin(\phi) \cos(\omega t) + C \cos(\phi) \sin(\omega t), \end{aligned} \quad (4.67)$$

after performing all the differentiations and collecting the $\cos(\omega t)$ and $\sin(\omega t)$ terms separately on the left in (4.67), as well as expanding the right hand side of (4.65) as $C \sin(\omega t + \phi) = C \sin(\phi) \cos(\omega t) + C \cos(\phi) \sin(\omega t)$. In order for (4.67) to be satisfied identically in t , the constants A and B must be chosen so that the coefficients of the $\cos(\omega t)$ and $\sin(\omega t)$ terms on both sides of (4.67) match, which yields

$$(k - m\omega^2)A + c\omega B = C \sin(\phi) \\ -Ac\omega + (k - m\omega^2)B = C \cos(\phi).$$

This is a system of two linear equations in two unknowns A and B . Some mildly tedious algebra leads to the solution

$$A = C \left(\frac{\sin(\phi)(k - m^2\omega^2) - \cos(\phi)c\omega}{(m\omega^2 - k)^2 + c^2\omega^2} \right), \\ B = C \left(\frac{\cos(\phi)(k - m^2\omega^2) + \sin(\phi)c\omega}{(m\omega^2 - k)^2 + c^2\omega^2} \right). \quad (4.68)$$

In (4.66) this yields the periodic solution $u_p(t)$ to (4.65).

The Amplitude of the Periodic Response

Our real interest is in the amplitude of the response $u_p(t)$, which is given by the quantity $\sqrt{A^2 + B^2}$. Despite the complexity of A and B in (4.68), the quantity $\sqrt{A^2 + B^2}$ simplifies considerably to

$$\sqrt{A^2 + B^2} = \underbrace{\frac{C}{\sqrt{(m\omega^2 - k)^2 + c^2\omega^2}}}_{\psi(\omega)}, \quad (4.69)$$

where we define $\psi(\omega)$ as this amplitude, considered primarily as a function of the driving frequency ω . We also assume $C > 0$. Some observations concerning the response amplitude $\psi(\omega)$ are

- The response amplitude $\psi(\omega)$ is proportional to C (see (4.63), the amplitude of the driving force).
- The response amplitude $\psi(\omega)$ does not depend on the input phase ϕ .
- If $c \neq 0$ the denominator of $\psi(\omega)$ is always positive for $\omega \geq 0$, and so ψ is defined for all $\omega \geq 0$.
- If $c = 0$ then $\psi(\omega)$ is undefined at $\omega = \sqrt{k/m}$, and both one-sided limits satisfy

$$\lim_{\omega \rightarrow (\sqrt{k/m})^-} \psi(\omega) = \lim_{\omega \rightarrow (\sqrt{k/m})^+} \psi(\omega) = \infty.$$

Periodic Forcing: Examples

Let's consider several examples of periodically forced systems: overdamped, underdamped, and undamped.

■ **Example 4.23** Consider the overdamped spring-mass system governed by

$$u''(t) + 4u'(t) + 3u(t) = \sin(\omega t + \phi).$$

Here we are assuming that the driving amplitude $C = 1$, while ω and ϕ are unspecified in the forcing function $f(t)$ of (4.63). According to the analysis above, the periodic response $u_p(t)$ of the system will be of the form (4.66) and have amplitude

$$\psi(\omega) = \frac{1}{\sqrt{(\omega^2 - 3)^2 + 16\omega^2}}$$

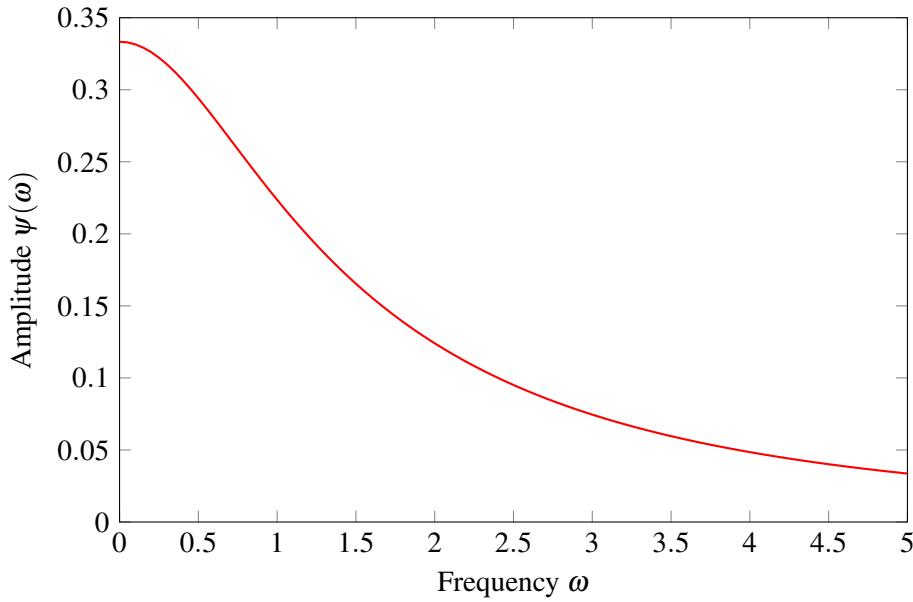


Figure 4.17: Periodic response amplitude for $u''(t) + 4u'(t) + 3u(t) = \sin(\omega t)$ as a function of the driving frequency ω .

which does not depend on ϕ . A plot of $\psi(\omega)$ is shown in Figure 4.17. In this case the system responds most vigorously (with the greatest amplitude) when ω is close to zero. The response amplitude drops off rapidly as the driving frequency increases.

To further illustrate, Figure 4.18 shows the actual response $u(t)$ of the system to forcing at $\omega = 1$ in the left panel and $\omega = 5$ in the right panel, with initial conditions $u(0) = u'(0) = 0$ in each case. The vertical scaling is the same, though note the horizontal scaling differs. After transients die out, the periodic response is assumed at an amplitude that can be read off of the graph in Figure 4.17. ■

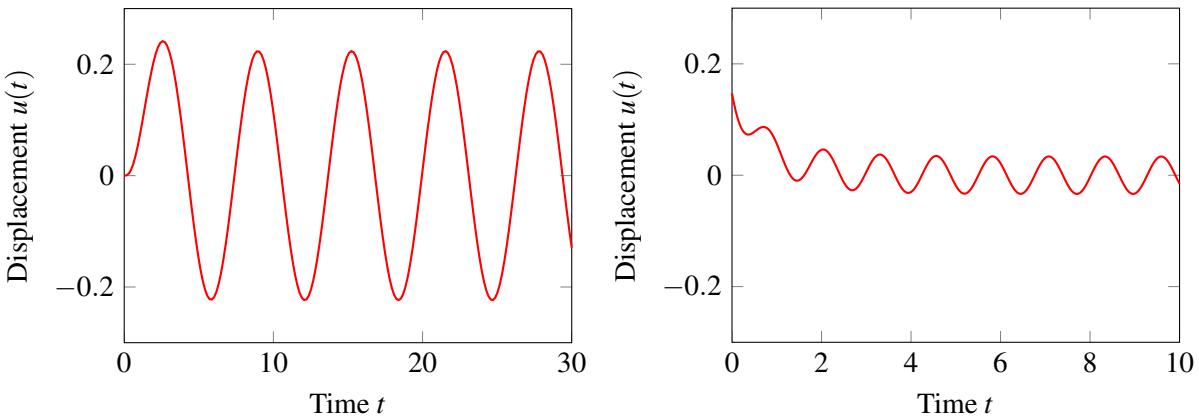


Figure 4.18: Left panel: solution to $u''(t) + 4u'(t) + 3u(t) = \sin(\omega t)$ with $u(0) = u'(0) = 0$ and $\omega = 1$. Right panel: same, with $\omega = 5$.

■ **Example 4.24** Consider the underdamped spring-mass system governed by

$$u''(t) + u'(t) + 10u(t) = \sin(\omega t + \phi).$$

Again the amplitude of the driving force is $C = 1$ with ω and ϕ unspecified in the forcing function

$f(t)$ of (4.63). According to the analysis above, the periodic response $u_p(t)$ of the system will be of the form (4.66) and have amplitude

$$\psi(\omega) = \frac{1}{\sqrt{(\omega^2 - 10)^2 + \omega^2}}.$$

A plot of $\psi(\omega)$ is shown in Figure 4.19. The system has a much more vigorous response to driving frequencies near $\omega \approx 3$.

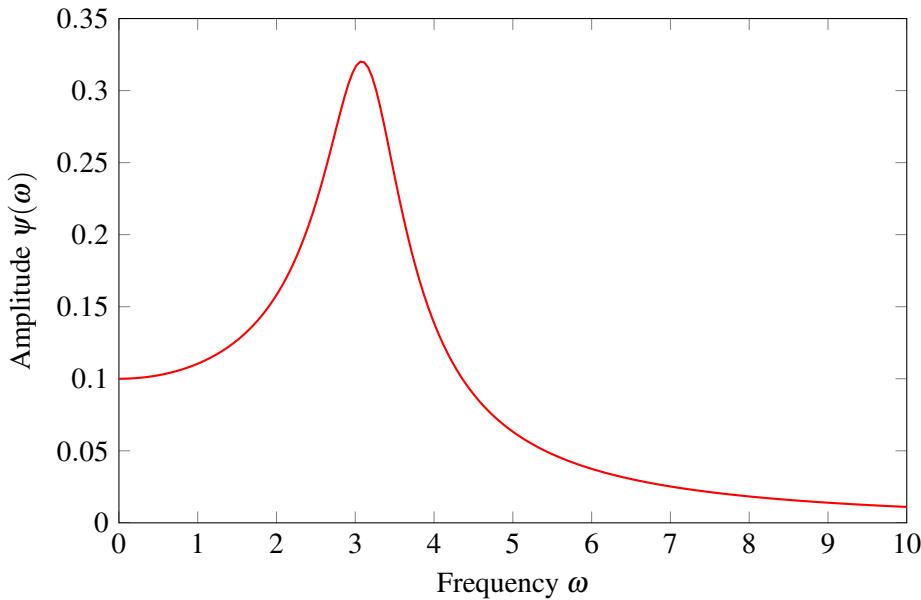


Figure 4.19: Periodic response amplitude for $u''(t) + u'(t) + 10u(t) = \sin(\omega t)$ as a function of the driving frequency ω .

Figure 4.20 shows the response $u(t)$ of the system to forcing at $\omega = 3$ in the left panel and $\omega = 5$ in the right panel, with initial conditions $u(0) = u'(0) = 0$ in each case. After transients die out, the periodic response is assumed at an amplitude that can be read off of the graph in Figure 4.19. The response at $\omega = 3$ has much greater amplitude than that at $\omega = 5$, despite the fact that the driving force in each case has amplitude $C = 1$.

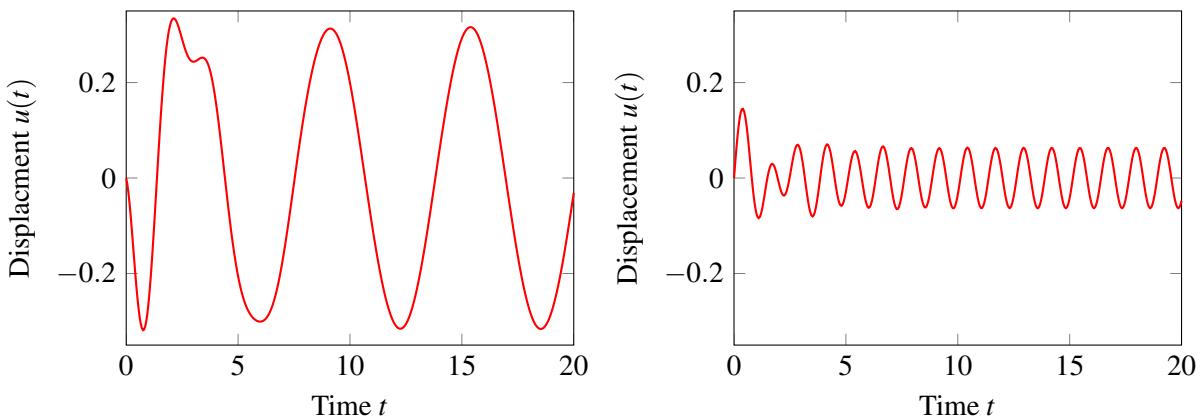


Figure 4.20: Left panel: solution to $u''(t) + u'(t) + 10u(t) = \sin(\omega t)$ with $u(0) = u'(0) = 0$ and $\omega = 3$. Right panel: same, with $\omega = 5$.

Reading Exercise 4.4.2 Show that the maximum value of $\psi(\omega) = 1/\sqrt{(\omega^2 - 10)^2 + \omega^2}$ graphed in Figure 4.19 occurs at $\omega = \sqrt{38}/2 \approx 3.08$.

An important note: The characteristic equation for the unforced system $u''(t) + u'(t) + 10u(t) = 0$ is $r^2 + r + 10 = 0$ with roots $r = -1/2 \pm i\sqrt{39}/2 \approx -0.5 \pm 3.12i$. Thus the solution to the unforced system is a superposition of $e^{-t/2} \cos(3.12t)$ and $e^{-t/2} \sin(3.12t)$, and so the unforced system vibrates at a natural frequency of $\omega = 3.12$ radians per second, quite close to the frequency where it responds most vigorously when forced. ■

■ **Example 4.25** Consider the undamped spring-mass system governed by

$$2u''(t) + 18u(t) = 5 \sin(\omega t + \phi).$$

Here we've used amplitude $C = 5$ in the forcing function $f(t)$ of (4.63), and left ω and ϕ undefined. According to the analysis above, the periodic response $u_p(t)$ of the system will be of the form (4.66) and have amplitude

$$\psi(\omega) = \frac{5}{\sqrt{(2\omega^2 - 18)^2}} = \frac{5}{2|\omega^2 - 9|}.$$

A plot of $\psi(\omega)$ is shown in Figure 4.21. The periodic response to forcing at $\omega = 3$ is unbounded; the graph is clipped vertically at $\psi = 10$. And indeed, for this undamped system the natural frequency of vibration for the unforced system is at precisely $\omega = 3$ radians per unit time.

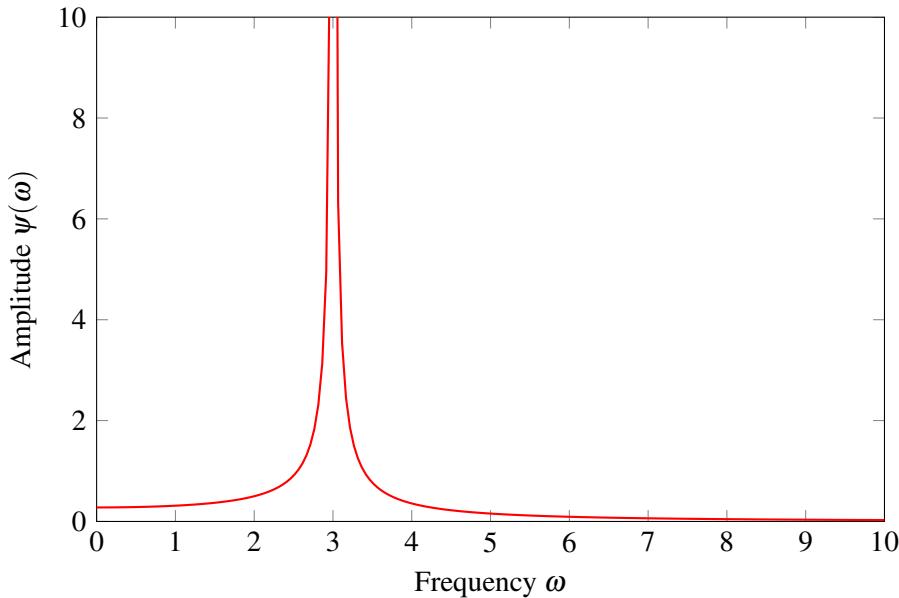


Figure 4.21: Periodic response amplitude for $2u''(t) + 18u(t) = 5 \sin(\omega t)$ as a function of the driving frequency ω .

What does the system do when actually driven with a forcing function of the form $f(t) = C \sin(3t)$? In Figure 4.22 we show the solution to $2u''(t) + 18u(t) = 5 \sin(3t)$ with $u(0) = 1$ and $u'(0) = 0$. This solution can be found using the techniques of Section 4.3. The expected particular solution $u_p(t) = A \sin(3t) + B \cos(3t)$ fails in this case, but an approach similar to that of Example 4.22 succeeds and yields the particular solution $u_p(t) = -\frac{5}{12}t \cos(3t)$. The full solution with the desired initial conditions turns out to be $u(t) = \frac{5}{36} \sin(3t) + \cos(3t) - \frac{5}{12}t \cos(3t)$. In this case, with no friction to dissipate energy, the system soaks up the energy provided at $\omega = 3$ radians per second and never attains a periodic response—the amplitude increases without bound. ■

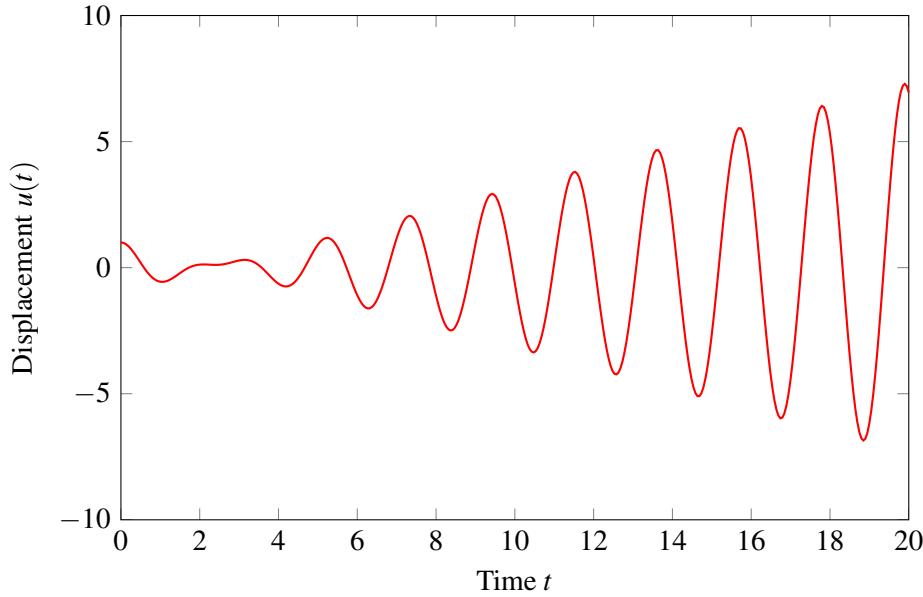


Figure 4.22: Solution to $2u''(t) + 18u(t) = 5 \sin(3t)$ with $u(0) = 1$ and $u'(0) = 0$.

Analysis of Resonance

The phenomenon in which a lightly damped (or undamped) system responds vigorously at one or more frequencies is illustrated in Examples 4.24 and 4.25 and is called **resonance**. Let's consider the general spring-mass-damper system and look at exactly when and how resonance can occur.

Consider the amplitude of the periodic response of a system governed by $mu'' + cu' + ku = C \sin(\omega t + \phi)$. This amplitude is $\psi(\omega)$ in (4.69), and depends on the driving force amplitude C as well as the driving frequency ω (but not ϕ). Let's take C out of consideration so we can focus on the effect of ω , by defining

$$G(\omega) = \frac{1}{\sqrt{(m\omega^2 - k)^2 + c^2\omega^2}}. \quad (4.70)$$

The function G is often called the **gain function** for the system. The amplitude of the system response to forcing $f(t) = C \sin(\omega t + \phi)$ is then $\psi(\omega) = G(\omega)C$. The equation $\psi(\omega) = G(\omega)C$ is the motivation for the terminology “gain function”, for the forcing amplitude C is amplified by a factor $G(\omega)$ to produce the amplitude of the system response. Note, however, that $f(t)$ and $u(t)$ have different physical dimensions.

Consider the graph of $G(\omega)$. It's easy to see from (4.70) that $G(\omega) \geq 0$ in all cases and $\lim_{\omega \rightarrow \infty} G(\omega) = 0$. For heavily damped systems this graph looks like that of Figure 4.17, and decreases monotonically with increasing ω . But for less heavily damped systems, the graph of $G(\omega)$ has a unique peak for some positive value of ω . The location of this peak can be found by setting $G'(\omega) = 0$, which yields

$$G'(\omega) = -\frac{4m(m\omega^2 - k)\omega + 2c^2\omega}{2((m\omega^2 - k)^2 + c^2\omega^2)^{3/2}} = 0.$$

The solution in ω is obtained by setting the numerator equal to zero, that is, $4m(m\omega^2 - k)\omega + 2c^2\omega = 0$. One solution is $\omega = 0$; look at Figures 4.17, 4.19, and 4.21. Dividing by $\omega \neq 0$ leaves us with $4m(m\omega^2 - k) + 2c^2 = 0$. The only other nonnegative solution is $\omega = \omega_{res}$ where

$$\omega_{res} = \frac{\sqrt{4km - 2c^2}}{2m} = \sqrt{\frac{k}{m} - \frac{1}{2} \left(\frac{c}{m}\right)^2}. \quad (4.71)$$

If $4km - 2c^2 > 0$, then ω_{res} is a positive real number. The inequality $4km - 2c^2 > 0$ is equivalent to $c^2 < 2km$ or $c < \sqrt{2km}$. Thus, the graph of $G(\omega)$ has a (unique) peak for $\omega > 0$ when the system damping is sufficiently light, $c < \sqrt{2km}$. You can check that ω_{res} has dimension T^{-1} , the correct dimension for a frequency. We summary these observations in the following definition:

Definition 4.4.1 A system governed by $mu'' + cu' + ku = f(t)$ has a **resonant frequency** ω_{res} given by (4.71) if $c^2 < 2km$.

Let's make two remarks concerning resonance:

- When $c = 0$ the resonant frequency is $\omega_{res} = \sqrt{k/m}$, which is exactly the natural frequency of the undriven system. This is called **pure resonance**, and is illustrated by Example 4.25.
- A Taylor series approximation to $\sqrt{4km - 2c^2}$ with respect to c at $c = 0$ shows that

$$\sqrt{4km - 2c^2} = 2\sqrt{km} - \frac{c^2}{2\sqrt{km}} + O(c^4).$$

Using the right hand side above in (4.71) shows that

$$\omega_{res} = \sqrt{\frac{k}{m}} - \frac{c^2}{4\sqrt{km^3}} + O(c^4). \quad (4.72)$$

Thus for lightly damped systems ($c \approx 0$), the resonant frequency is very close to $\sqrt{k/m}$. This approximation is used quite often in practice.

Reading Exercise 4.4.3 Show that $G(0) = 1/k$. Also show that

$$\lim_{\omega \rightarrow \infty} m\omega^2 G(\omega) = 1.$$

Conclude that if ω is large then we are justified in writing $G(\omega) \approx \frac{1}{m\omega^2}$.

Reading Exercise 4.4.4 Suppose a spring-mass-damper system with mass m , damping c , and spring constant k is underdamped (that is, $c^2 < 4mk$). Will resonance occur? If resonance does occur, show that this resonant frequency is less than the natural frequency of the unforced system.

Beats

The undamped spring-mass system of Example 4.25 oscillates with ever-increasing amplitude when it is driven at its natural frequency. What happens in an undamped system when driven at other frequencies, especially those close to the natural frequency?

Consider an undamped spring-mass system

$$mu''(t) + ku(t) = C \sin(\omega t + \phi). \quad (4.73)$$

The general solution to the homogeneous (or undriven) system is

$$u_h(t) = c_1 \cos(\omega_0 t) + c_2 \sin(\omega_0 t),$$

where $\omega_0 = \sqrt{k/m}$ is the natural frequency of this system. If the driving frequency ω does not equal ω_0 in (4.73), then one can use the method of undetermined coefficients to find that a particular solution is

$$u_p(t) = \frac{C \sin(\omega t + \phi)}{\omega^2 - \omega_0^2}.$$

A general solution to (4.73) is then

$$\begin{aligned} u(t) &= u_h(t) + u_p(t) = c_1 \cos(\omega_0 t) + c_2 \sin(\omega_0 t) + \frac{C \sin(\omega t + \phi)}{\omega^2 - \omega_0^2} \\ &= c_1 \cos(\omega_0 t) + c_2 \sin(\omega_0 t) + \frac{C \sin(\phi)}{\omega^2 - \omega_0^2} \cos(\omega t) + \frac{C \cos(\phi)}{\omega^2 - \omega_0^2} \sin(\omega t) \end{aligned} \quad (4.74)$$

where the second line follows from the identity $\sin(a+b) = \sin(a)\cos(b) + \cos(a)\sin(b)$. The constants c_1 and c_2 in (4.74) are determined by the initial conditions. In this case there is no transient response; with no damping, the contribution of $u_h(t)$ never decays. A driven undamped system can exhibit some peculiar behavior, as the next example demonstrates.

■ Example 4.26 Consider the undamped system of Example 4.25, governed by $2u''(t) + 18u(t) = 5 \sin(\omega t)$ with $\omega = 4$ and initial conditions $u(0) = 0$ and $u'(0) = 0$. The solution in this case is

$$u(t) = \frac{10}{21} \sin(3t) - \frac{5}{14} \sin(4t).$$

The $\sin(3t)$ term stems from the system's natural unforced response and the initial conditions, but since there is no damping, this portion of the solution never decays. The $\sin(4t)$ term is due to the driving force. The function $u(t)$ is plotted in Figure 4.23. It consists of a superposition of a sine wave with period $2\pi/3$ and a sine wave of period $2\pi/4$, and is therefore itself periodic. ■

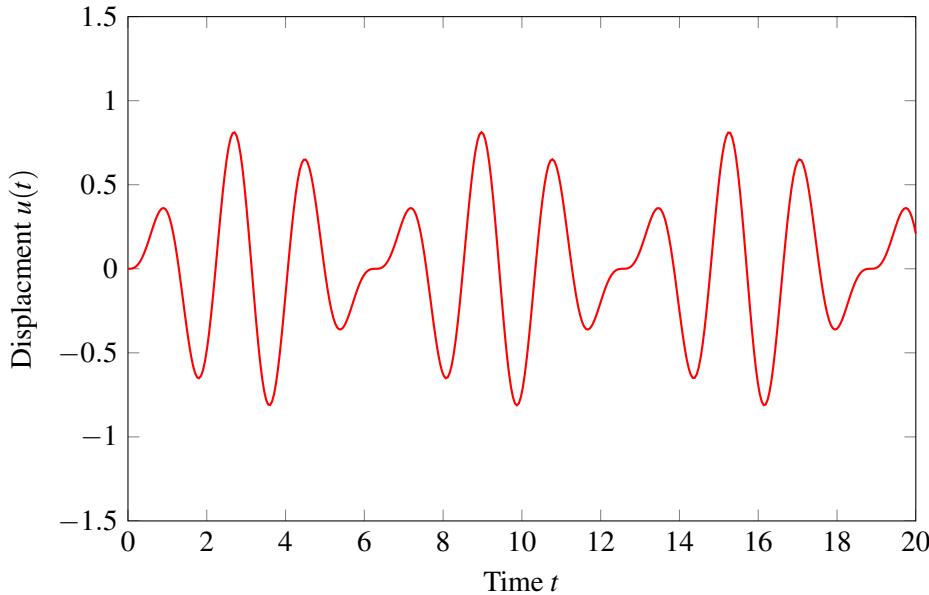


Figure 4.23: Solution to $2u''(t) + 18u(t) = 5 \sin(4t)$ with $u(0) = 0$ and $u'(0) = 0$.

Reading Exercise 4.4.5 What is the period of the solution in Example 4.26?

■ Example 4.27 Consider now what happens if the driving frequency is close to but not equal to the natural frequency of the undamped system. Let us examine $2u''(t) + 18u(t) = 5 \sin(3.2t)$ with $u(0) = u'(0) = 0$. The solution is

$$u(t) = \frac{200}{93} \sin(3t) - \frac{125}{62} \sin(3.2t).$$

This solution is plotted on the range $0 \leq t \leq 15$ in the left panel of Figure 4.24. The plot looks very much like the pure resonance of Example 4.22, but plotting out to $t = 100$ shows that the amplitude

does not continue to increase. Instead we see here a peculiar **beat** phenomenon, in which a high frequency sinusoidal function exhibits a slowly varying amplitude. The appearance of such beats is typical of an undamped system when driven near its natural frequency. The solution looks very much like a sine wave of the form $C \sin(3t + \phi)$ for some phase shift ϕ , but with an amplitude $C = C(t)$ that varies slowly and periodically.

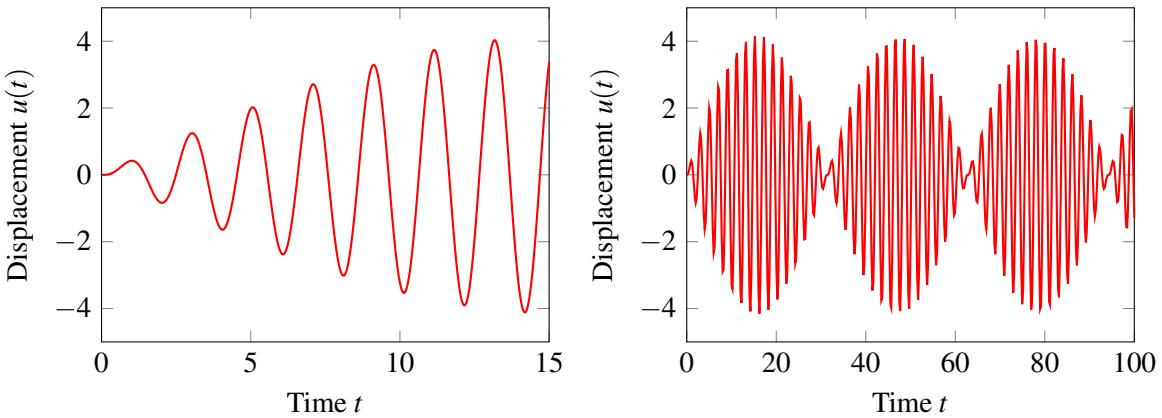


Figure 4.24: Left panel: solution to $2u''(t) + 18u(t) = 5 \sin(3.2t)$ with $u(0) = 0$ and $u'(0) = 0$ for $0 \leq t \leq 15$. Right panel Same solution $u(t)$ plotted for $0 \leq t \leq 100$.

This phenomenon occurs in lightly damped systems too, until the transient portion of the solution dies out. A physical example is the common auditory phenomenon in which two objects (e.g., tuning forks) emit closely spaced frequencies. The superposition is perceived by the ear as a kind of periodic beat in the intensity of the composite sound. ■

For an audio demonstration of the beat phenomenon, see [6].

Analysis of the Beat Phenomenon

To understand the beat phenomenon more quantitatively, consider an undamped oscillator with natural frequency ω_0 driven sinusoidally with a function of the form $f(t) = C \sin(\omega t + \phi)$. As detailed in (4.74), a general solution is

$$u(t) = c_1 \cos(\omega_0 t) + c_2 \sin(\omega_0 t) + A \cos(\omega t) + B \sin(\omega t),$$

where $A = \frac{C \sin(\phi)}{\omega^2 - \omega_0^2}$ and $B = \frac{C \cos(\phi)}{\omega^2 - \omega_0^2}$. The analysis that follows can be done quite generally, for any choice of c_1, c_2, A , and B , but for clarity let's focus on a specific choice, the case in which $c_1 = -A$ and $c_2 = B = 0$. In this case $u(t) = -A \cos(\omega_0 t) + A \cos(\omega t)$. Let us further specify that $A = 1$ so that

$$u(t) = \cos(\omega t) - \cos(\omega_0 t). \quad (4.75)$$

This function is graphed in Figure 4.25 for the case $\omega_0 = 2$ and $\omega = 1.8$ (the solid red curve).

Reading Exercise 4.4.6 What initial conditions does $u(t)$ in (4.75) satisfy?

To understand the behavior of $u(t)$ in (4.75) we make use of the trigonometric identity $\cos(x) - \cos(y) = 2 \sin((y-x)/2) \sin((x+y)/2)$ with $x = \omega t$ and $y = \omega_0 t$ to find

$$\begin{aligned} u(t) &= \cos(\omega t) - \cos(\omega_0 t) \\ &= 2 \sin((\omega_0 - \omega)t/2) \sin((\omega + \omega_0)t/2). \end{aligned} \quad (4.76)$$

Suppose the system is driven at a frequency close to resonance, so that ω is close to ω_0 , say $\omega = \omega_0 - \delta$ where $\delta = \omega_0 - \omega$ is close to zero. Then (4.76) can be written as

$$u(t) = \underbrace{2 \sin(\delta t/2)}_{\text{amplitude}} \sin((\omega_0 - \delta/2)t). \quad (4.77)$$

The right side of (4.77) can be interpreted as a sine wave, $\sin((\omega_0 - \delta/2)t)$, that oscillates with a frequency close to ω_0 , but with slowly varying amplitude $2 \sin(\delta t/2)$ (slow because δ is close to zero). It is this slowly varying amplitude that gives rise to the beat phenomenon.

In Figure 4.25 is shown the graph of $u(t)$ (solid red curve) for the case $\omega_0 = 2$, $\omega = 1.8$, so $\delta = 0.2$. Here

$$u(t) = \cos(1.8t) - \cos(2t) = 2 \sin(0.1t) \sin(1.9t).$$

Also shown is the graph of $\pm 2 \sin(\delta t/2)$ (the dashed black curve) that defines the slowly varying amplitude of the $\sin(1.9t)$ piece. The quantity $\pm 2 \sin(\delta t/2)$ that defines the amplitude of the rapidly varying sinusoidal piece is called the **envelope** of the solution. Note that this amplitude varies sinusoidally at a frequency of $|\delta/2|$ radians per unit time ($\delta < 0$ is possible, hence the absolute values), or a period of $4\pi/|\delta|$. However, the beats occur at twice this frequency, with a period of

$$P_{beat} = \frac{2\pi}{|\delta|} = \frac{2\pi}{|\omega_0 - \omega|}. \quad (4.78)$$

time units.

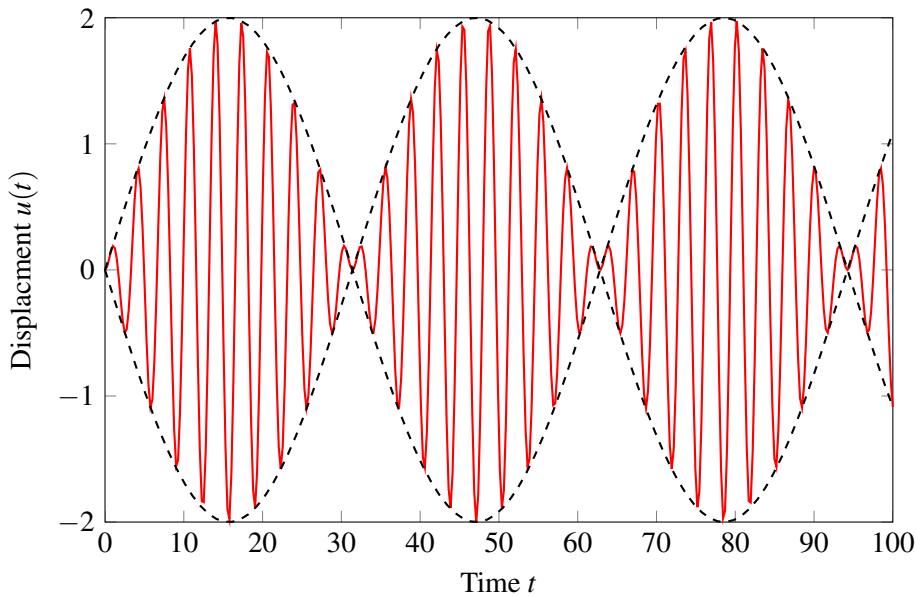


Figure 4.25: Plot of $u(t) = \cos(1.8t) - \cos(2t)$ (solid red curve) and amplitude envelope $\pm 2 \sin(0.1t)$ (dashed black curve).

4.4.3 Exercises

Exercise 4.4.1 For each ODE in parts (a)-(h)

- Plot the gain function $G(\omega)$ as defined by (4.70) on the given range for ω .
- Use the method of undetermined coefficients to compute the periodic response of the system for the given driving function $f(t)$, compute the amplitude of the response, and verify the result agrees with the graph of G .
- If the system exhibits resonance, find the frequency at which the system resonates.

- $2u''(t) + u'(t) + 8u(t) = f(t)$, $f(t) = 3 \sin(4t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 10$.
- $2u''(t) + 6u'(t) + 8u(t) = f(t)$, $f(t) = 3 \sin(4t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 10$.
- $2u''(t) + 4u'(t) + 20u(t) = f(t)$, $f(t) = 5 \cos(2t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 10$.
- $20u''(t) + 2u'(t) + 100u(t) = f(t)$, $f(t) = \cos(2.25t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 10$.
- $20u''(t) + 2u'(t) + 100u(t) = f(t)$, $f(t) = \sin(10t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 10$. Compare the amplitude of the periodic response to that of (d).
- $u''(t) + u'(t) + 10000u(t) = f(t)$, $f(t) = 5 \cos(100t) + \sin(100t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 200$.
- $u''(t) + 10u'(t) + u(t) = f(t)$, $f(t) = 2 \cos(2t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 5$.
- $u''(t) + u(t) = f(t)$ (undamped), $f(t) = \cos(2t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 3$.

Exercise 4.4.2 An RLC series circuit has inductor $L = 10^{-4}$ henries, resistor $R = 2$ ohms, and capacitor $C = 10^{-6}$ farads, with a voltage source $V(t) = \sin(\omega t)$ in series with these components. Write out the ODE that governs the charge on the capacitor. Compute the gain function for this circuit, plot the gain function over the range $0 \leq \omega \leq 10^6$, and find the resonant frequency for this system.

Exercise 4.4.3 Find the gain function for the ODE $Lq''(t) + Rq'(t) + q(t)/C = \sin(\omega t)$ that governs an RLC circuit (in terms of R, L, C , and ω , treating q as the output) and show that this circuit has resonant frequency $\omega_{res} = \frac{\sqrt{4L/C - 2R^2}}{2L}$.

Exercise 4.4.4 Consider the underdamped system $25u''(t) + 10u'(t) + 26u(t) = f(t)$ with $f(t) = \cos(t) + \cos(5t)$ and initial data $u(0) = 0$ and $u'(0) = 0$.

- (a) Plot the gain function $G(\omega)$ as defined by (4.70) for $0 \leq \omega \leq 5$.
- (b) Compute the resonant frequency of this system.
- (c) Find the function $u(t)$ and plot it for $0 \leq t \leq 50$. Identify (approximately) on the graph that part of the solution that is transient.
- (d) For t sufficiently large the system settles into a periodic response. What is the period and radial frequency of this response? Why do we see little evidence of the $\cos(5t)$ term in $f(t)$ in the system periodic response? Hint: compute $G(1)$ and $G(\omega)$.

Exercise 4.4.5 Suppose a spring-mass-damper system with mass m , damping c , and spring constant k exhibits resonance. Show that if resonance occurs, then the maximum value for the gain function is $1/(c\omega_{nat})$, where $\omega_{nat} = \frac{\sqrt{4km - c^2}}{2m}$ is the system's natural frequency (recall

(4.29)).

Exercise 4.4.6 In certain situations in which a system $mu''(t) + cu'(t) + ku(t) = f(t)$ is driven by a periodic forcing function, say $f(t) = C \sin(\omega t + \phi)$, we want to know the amplitude of $u'_p(t)$ or even $u''_p(t)$ for the periodic response, rather than $u_p(t)$. That is, we are concerned with the amplitude of the velocity or acceleration of the mass, or the current q' in an RLC circuit. (For an example see Exercise 4.4.8.)

As shown previously in this section, the periodic response is $u_p(t) = A \cos(\omega t) + B \sin(\omega t)$ (this was (4.66)) where A and B are given by (4.68).

- Show that the magnitude of $u'_p(t)$ is given by $C\omega G(\omega)$ where $G(\omega)$ is defined by (4.70).
- Show that the value of $\omega > 0$ that maximizes the amplitude of $u'_p(t)$ is $\omega'_{res} = \sqrt{k/m}$ (which does not depend on c). Compare ω'_{res} to the usual resonant frequency $\omega_{res} = \sqrt{k/m - c^2/2m^2}$ at which the amplitude of $u_p(t)$ itself is maximized, in particular when c is close to zero.
- Repeat this analysis to find the value of ω''_{res} that maximizes the amplitude of $u''_p(t)$. Compare ω''_{res} to ω_{res} when c is close to zero.

Exercise 4.4.7 Engineers often quantify the sharpness of the peak in the gain function $G(\omega)$ for a system that exhibits significant resonance by using the **Q-factor**. There are various, slightly different definitions for the Q-factor, but one common one is

$$Q = \frac{\omega_{res}}{\omega_+ - \omega_-}. \quad (4.79)$$

In equation (4.79), ω_{res} is the resonant frequency of the system. The frequencies ω_- and ω_+ are defined by the requirement that

$$G(\omega_-) = G(\omega_+) = \frac{G(\omega_{res})}{\sqrt{2}} \quad (4.80)$$

with $\omega_- < \omega_{res}$ and $\omega_+ > \omega_{res}$, where $G(\omega)$ is the gain function defined in (4.70). See Figure 4.26 for an illustration. The larger the value of Q , the sharper the resonance peak.

For each system below, compute $G(\omega)$ and plot it on the given range. Then solve (4.80) for each of ω_- and ω_+ and compute Q for the system using (4.79).

- $m = 25, c = 10, k = 26, 0 \leq \omega \leq 5$.
- $m = 2, c = 1, k = 20, 0 \leq \omega \leq 10$.
- $m = 2, c = 0.1, k = 20, 0 \leq \omega \leq 10$.
- $m = 2, c = 0.01, k = 20, 0 \leq \omega \leq 10$.
- $m = 2, c = 0, k = 20, 0 \leq \omega \leq 10$. (Based on (b)-(d), how should Q be defined here?)

Exercise 4.4.8 A single-story building is modeled as a spring-mass-damper system with $m = 5000$ kg, $c = 10000$ newtons per meter per second, and $k = 5 \times 10^5$ newtons per meter. With $u(t)$ as the displacement of the roof mass, suppose initial data $u(0) = 0$ and $u'(0) = 0$ when an earthquake strikes at time $t = 0$ and exerts force $f(t) = 10^4 \cos(\omega t)$, where ω may lie anywhere in the range $0 \leq \omega \leq 6\pi$ (0 to 3 hertz.)

- What is the resonant frequency for this building? What is the maximum possible amplitude

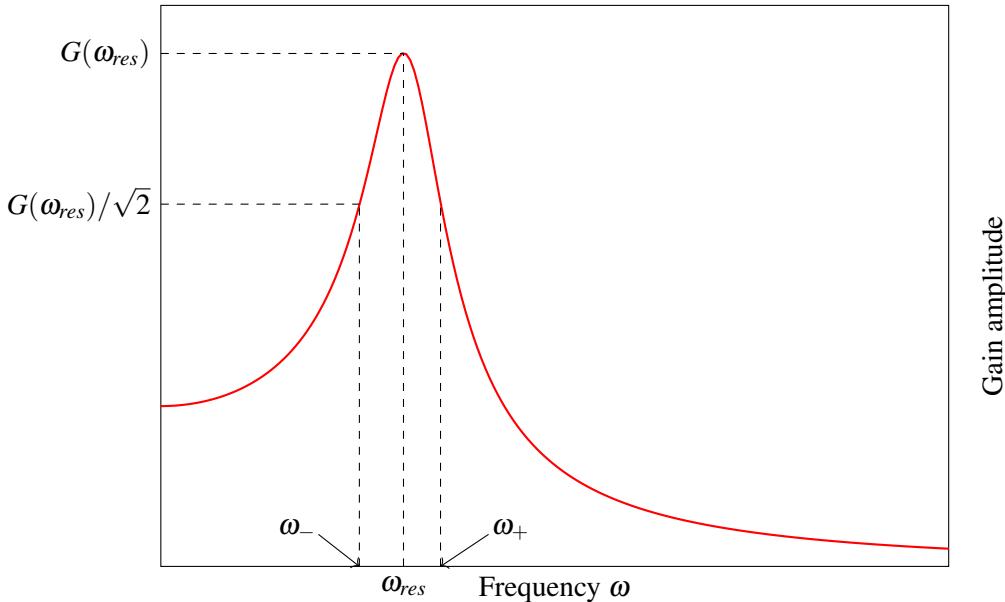


Figure 4.26: Illustration for definition of Q in equation (4.79).

- of the building's displacement? Hint: start by computing and plotting $G(\omega)$.
- What is the maximum possible amplitude of the building's acceleration for frequencies in the given range? Hint: $G(\omega)$ gives the amplitude $\sqrt{A^2 + B^2}$ of the periodic solution $u_p(t) = A \cos(\omega t) + B \sin(\omega t)$; the acceleration is $u''_p(t) = -\omega^2(A \cos(\omega t) + B \sin(\omega t))$. Find the amplitude of $u''_p(t)$ as a function of ω and maximize. See also Exercise 4.4.6.
 - Suppose the acceleration of the building's periodic motion must be kept below 6 meters per second squared for frequencies between 0 and 3 hertz. What is the smallest value of the damping coefficient c that accomplishes this?

Exercise 4.4.9 Each undamped ODE below is driven near its resonant frequency ω_0 by a driving function $f(t) = \cos(\omega t)$. Find the solution with the given initial conditions, then plot the solution on the given time interval. Find the natural frequency of the undriven system and use (4.78) to explain the graph, in particular the period of the beats.

- $u''(t) + u(t) = \cos(0.9t)$, $u(0) = u'(0) = 0$, on the range $0 \leq t \leq 250$.
- $u''(t) + u(t) = \cos(1.2t)$, $u(0) = u'(0) = 0$, on the range $0 \leq t \leq 100$.
- $u''(t) + 4u(t) = \cos(1.9t)$, $u(0) = 0$, $u'(0) = 0$, on the range $0 \leq t \leq 200$.
- $u''(t) + 4u(t) = \cos(1.99t)$, $u(0) = 0$, $u'(0) = 0$, on the range $0 \leq t \leq 1000$.

Exercise 4.4.10 Consider an undamped system $mu''(t) + ku(t) = C \sin(\omega t)$ with $u(0) = u'(0) = 0$, where $C > 0$. Let $\omega_0 = \sqrt{k/m}$ denote the natural frequency of the undamped system.

- Verify that the solution to this ODE is

$$u(t) = -\frac{C\omega}{m\omega_0(\omega_0^2 - \omega^2)} \sin(\omega_0 t) + \frac{C}{m(\omega_0^2 - \omega^2)} \sin(\omega t).$$

- (b) Use the triangle inequality $|x + y| \leq |x| + |y|$ for any real numbers x and y to argue that

$$|A \sin(\omega_0 t) + B \sin(\omega t)| \leq |A| + |B|$$

for all t . Use this to show that the maximum value of $|u(t)|$ in part (a) is bounded by

$$|u(t)| \leq \frac{C}{m\omega_0|\omega_0 - \omega|}.$$

The amplitude of the beat is therefore inversely proportional to $1/|\omega_0 - \omega|$.

- (c) Verify this result by checking it against the graph of $u(t)$ from Example 4.27.

Exercise 4.4.11 Lightly driven systems can exhibit the beat phenomenon, at least until the transient portion of the solution dies out. Solve the ODE $10u''(t) + 0.25u'(t) + 10u(t) = \sin(1.1t)$ with $u(0) = 0$ and $u'(0) = 0$. Plot the solution on the interval $0 \leq t \leq 400$. What is the natural frequency of the undriven version of this spring-mass system? What is the period of the beat, and how does it compare to that of the undamped system predicted by (4.78)?

4.5 Scaling and Nondimensionalization for ODEs

4.5.1 Motivation: Nonlinear Springs

The spring-mass models we've considered so far have been built on Hooke's law, which posits a linear relationship $F = kx$ for the force F necessary to stretch a spring a distance x from equilibrium. This type of linear relationship between forces and displacements (or stresses and strains) is used much more generally in mechanics and materials science. For all materials this model has limits, however, beyond which the relationship becomes nonlinear. Consider, for example, a so-called **hard spring** in which the force-displacement relationship is

$$F = k_1x + k_2x^3, \quad (4.81)$$

where k_1 and k_2 are nonnegative constants. The case $k_2 = 0$ in (4.81) is the usual linear Hooke's law (4.1). If $k_2 > 0$ and $|x|$ is sufficiently small then the k_2x^3 term is much smaller in magnitude than the k_1x term and F in (4.81) is close to the force predicted by the usual Hooke's law with spring constant k_1 . But if $|x|$ is large, then the k_2x^3 dominates. Note that both terms on the right in (4.81) are always of the same sign, so when $|x|$ is large the force from (4.81) substantially exceeds that predicted by the linear Hooke's law. This is the basis for the term "hard spring."

Reading Exercise 4.5.1 Show that if $|x| \leq 0.1\sqrt{k_1/k_2}$ then $|k_2x^3| \leq 0.01|k_1x|$. That is, the magnitude of the nonlinear k_2x^3 term on the right in (4.81) is less than one percent of the magnitude of the linear k_1x term, and so we might consider ignoring it. Of course, one percent can be altered to any prescribed tolerance.

Suppose that a spring obeys (4.81) in an undamped spring-mass-damper model. The same reasoning that led to (4.4) now yields a nonlinear second-order ODE

$$mu''(t) + k_1u(t) + k_2u^3(t) = 0 \quad (4.82)$$

with $u(t)$ as the displacement of a mass m . In some settings the model (4.82) might be more accurate than the ODE (4.4) based on Hooke's law, but with one big drawback: The ODE (4.82) has no analytical solution. We would therefore be forced to use numerical or qualitative techniques to analyze (4.82). But if the nonlinear term $k_2u^3(t)$ is sufficiently small, as in Reading Exercise

4.5.1, perhaps the linear ODE $mu''(t) + k_1 u(t) = 0$ is a good enough model. This would provide the luxury of an analytical solution.

The trick is to figure out when the nonlinear term can be ignored. There are three parameters in (4.82): m, k_1 , and k_2 . Any specific solution to (4.82) also requires two initial conditions, say $u(0) = u_0$ and $u'(0) = v_0$. For some combinations of these five parameters, the nonlinear term doesn't matter; for other combinations it does. Moreover, the time interval over which we seek a solution may come into play.

■ **Example 4.28** Consider (4.82) in the case that $m = 10, k_1 = 5$, and $k_2 = 0.25$, with initial conditions $u(0) = 1$ and $u'(0) = 0$. The left panel of Figure 4.27 shows the graph of the solution $u(t)$ as a solid red curve, computed numerically. The dashed black curve is the solution to the linear ODE $mu'' + k_1 u = 0$ (cubic term omitted) with the same initial data, $u(0) = 1$ and $u'(0) = 0$. Both solutions are graphed on the time interval $0 \leq t \leq 40$. The agreement on this time interval might be considered good enough for some purposes, and so the analytical solution to the linear ODE could be used, at least for these parameter values.

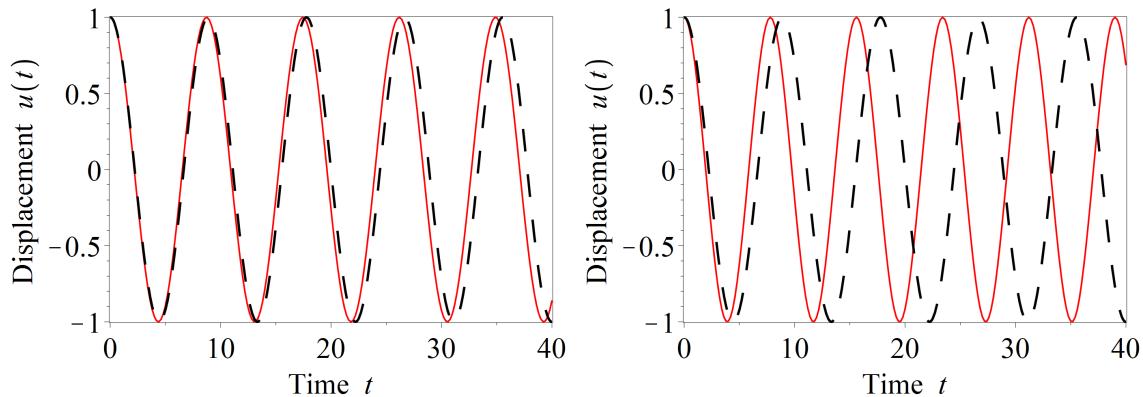


Figure 4.27: Left panel: Solution to (4.82) with $m = 10, k_1 = 5, k_2 = 0.25$ with initial data $u(0) = 1$ and $u'(0) = 0$ (solid red curve) and solution to (4.82) with cubic term $k_2 u^3$ omitted (dashed black curve). Right panel: Same, but with $k_2 = 2$.

The right panel of Figure 4.27 shows the solutions to these nonlinear and linear ODEs but with $k_2 = 2$ and with the remaining parameters the same as in the left panel. In this case the solutions are completely out of phase by time $t = 40$. A solution to the linear ODE is not likely to be an adequate approximation to the corresponding solution to the nonlinear ODE in this circumstance. ■

With the six parameters m, k_1, k_2, u_0 and v_0 and the time interval $0 \leq t \leq T$, using brute force to sort out what combinations give reasonable agreement between the linear and nonlinear versions of the ODE and which combinations do not is hopeless. What we need is a method to reduce the number of parameters to be considered and distill the ODE down to a more elemental form, so that what is important and what is not important is more readily apparent. This task is facilitated by tools from the subject of **scaling and nondimensionalization**. These techniques build on the ideas presented in Section 1.5, and are also useful tools for parameter estimation problems.

4.5.2 Characteristic Variable Scales

Differential equations that model physical situations always have one or more **characteristic variable scales** for the independent and dependent variables. To illustrate, consider an undamped spring-mass system governed by $mu''(t) + ku(t) = 0$, with initial data $u(0) = u_0$ and $u'(0) = 0$. The solution to this ODE is $u(t) = u_0 \cos(t\sqrt{k/m})$. This function is oscillatory with a period $P = 2\pi\sqrt{m/k}$. In this case we can say that a characteristic time scale for the motion of the mass is

$\sqrt{m/k}$ (dimensionless constants like 2π are usually omitted). It's easy to check that since $[m] = M$ and $[k] = MT^{-2}$, the quantity $\sqrt{m/k}$ does indeed have the dimension of time. So, for example, if $m = 4 \text{ kg}$ and $k = 1 \text{ newton per meter}$, then a characteristic time scale for the motion is $\sqrt{m/k} = 2 \text{ seconds}$. Based on this characteristic time scale, if we measure the mass position $u(t) = \cos(t/2)$ (period $4\pi \approx 12.6 \text{ seconds}$) every two seconds we will see the smooth and gradual change in $u(t)$. On the other hand, measuring $u(t)$ every 1000 seconds will produce a meaningless cloud of data with no pattern. Measuring $u(t)$ every 10^{-6} seconds will make it seem that $u(t)$ isn't changing at all from one time to the next, and we will see nothing without taking a lot of data. Measuring $u(t)$ every two seconds is about right to see what's going on with this system.

Time scales varies widely from one physical scenario to another. Some oscillators, say the quartz crystal in a watch, have a time scale measured in microseconds. Others, like the vibration of a tall building, have time scales measured in seconds. Others may have time scales measured in millions of years. It all depends on m and k , and more generally the ODE and the physics. The dependent variable (solution to the ODE) also has its own scale. In the spring-mass example above, $u(t) = u_0 \cos(t\sqrt{k/m})$ has a characteristic length scale u_0 , since $u(t)$ varies between $-u_0$ and $+u_0$; this length scale might be nanometers, meters, or parsecs (a parsec is a unit of length, not time, Han.)

Understanding the various scales in an ODE provides a number of benefits. The first is simply for intuition and sanity checks. If we're modeling something on an atomic scale and the solution we find evolves on a time scale of years, we probably messed up the model or the solution. Another good reason to understand the time scale of an ODE is for efficient numerical solutions. If a nonlinear pendulum (see Section 4.6) swings with a time scale of seconds then the numerical solver can likely use time steps of 0.1 seconds or so; time steps of 10^{-6} seconds will be wasteful. When estimating the parameters in a system from data, an understanding of the time or length (or other) scales may help us to decide how much data to collect. The spring-mass system in the Modeling Project "Parameter Estimation with Second-Order ODEs" of Section 4.6.3 has a time scale on the order of one second, so taking 50 position measurements per second is clearly enough. Finally, if there are multiple time scales in an ODE and these scales are of widely varying magnitudes, this can mean trouble for numerical algorithms. This issue is common when we confront system of ODEs as in Chapters 6 and 7. There are specific techniques for combating the difficulties that arise in this situation and these are addressed in Section 7.5.

One of the most important benefits of understanding the various scales in an ODE is that it lets us determine conditions under which certain parameters or terms in an ODE are negligible and so may be omitted. Such omissions can lead to considerable simplification of the analysis, as illustrated below.

Finding Characteristic Variable Scales

How can the characteristic scales of an ODE be determined without solving the ODE? The answer lies in the dimensions of the various physical constants, parameters, and initial conditions that enter into the ODE. Let us proceed by looking at some examples.

■ **Example 4.29** Consider the undamped spring-mass ODE $mu''(t) + ku(t) = 0$ with $u(0) = u_0$ and $u'(0) = 0$. The dimension of the various constants are $[m] = M$, $[k] = MT^{-2}$, and $[u_0] = L$. Any characteristic time scale t_c for this ODE will be encoded in the values of these constants. We hypothesize

$$t_c = m^\alpha k^\beta u_0^\gamma \quad (4.83)$$

for some constants α, β , and γ . Since $[t_c] = T$, dimensional consistency in (4.83) demands that

$$[t_c] = [m]^\alpha [k]^\beta [u_0]^\gamma \text{ or (noting that } T = M^0 L^0 T^1)$$

$$\begin{aligned} M^0 L^0 T^1 &= M^\alpha (MT^{-2})^\beta L^\gamma \\ &= M^{\alpha+\beta} L^\gamma T^{-2\beta} \end{aligned} \quad (4.84)$$

after grouping and simplifying the M, L , and T exponents. Matching exponents on the left and right in (4.84) leads to $\alpha + \beta = 0$, $\gamma = 0$, and $-2\beta = 1$, three equations in unknowns α, β and γ . The unique solution is $\alpha = 1/2, \beta = -1/2, \gamma = 0$. We conclude from (4.83) that

$$t_c = m^{1/2} k^{-1/2} u_0^0 = \sqrt{m/k},$$

which is precisely what was obtained from the ODE solution $u(t) = u_0 \cos(t \sqrt{k/m})$. ■

Reading Exercise 4.5.2 Redo Example 4.29, but this time look for a characteristic spatial scale $u_c = m^\alpha k^\beta u_0^\gamma$ by adjusting α, β , and γ appropriately to obtain dimensional consistency.

An ODE may have multiple scales in time or space, as the next example illustrates.

■ **Example 4.30** Consider a spring-mass-damper system governed by $mu''(t) + cu'(t) + ku(t) = 0$; we won't incorporate the dimensions for the initial data, just the system parameters. Then $[m] = M$, $[c] = MT^{-1}$, and $[k] = MT^{-2}$. For a characteristic time scale $t_c = m^\alpha c^\beta k^\gamma$ the constants α, β , and γ must satisfy

$$\begin{aligned} M^0 L^0 T^1 &= (M^\alpha)(MT^{-1})^\beta(MT^{-2})^\gamma \\ &= M^{\alpha+\beta+\gamma} L^0 T^{-\beta-2\gamma} \end{aligned}$$

after collecting and simplifying the exponents. Matching exponents on the left and right above yields equations

$$\begin{aligned} \alpha + \beta + \gamma &= 0 \\ -\beta - 2\gamma &= 1. \end{aligned}$$

There are infinitely many solutions to these equations. Specifically, we may (for example) choose α arbitrarily and then solve for β and γ in terms of α as

$$\beta = 1 - 2\alpha \quad \text{and} \quad \gamma = \alpha - 1. \quad (4.85)$$

Alternatively we could choose β and solve for α and γ , or choose γ and solve for α and β , but all roads lead to the same conclusions. Based on (4.85), this problem has infinitely many time scales. With β and γ as in (4.85), any characteristic time scale t_c has the form

$$t_c = m^\alpha c^{1-2\alpha} k^{\alpha-1} = \frac{c}{k} \left(\frac{mk}{c^2} \right)^\alpha \quad (4.86)$$

for any choice of α . Note that on the right in (4.86) the quantity c/k has dimension time, while the parenthesized quantity mk/c^2 in (4.86) is dimensionless, and so is $(mk/c^2)^\alpha$ for any α . This makes it easy to see that the right side of (4.86) has the dimension of time for any α .

It might seem a bit disconcerting that there are infinitely many time scales we can construct with this process, but this is common. Different choices for α in (4.86) emphasize different aspects of how the solution to $mu''(t) + cu'(t) + ku(t) = 0$ evolves in time. ■

Reading Exercise 4.5.3 Show that the choice $\alpha = 1/2$ in (4.86) yields the time scale $t_c = \sqrt{m/k}$ seen previously that captures the period of the oscillatory portion of the mass motion.

Reading Exercise 4.5.4 Show that the choice $\alpha = 1$ in (4.86) yields the time scale $t_c = m/c$. What aspect of the mass's motion does this emphasize? Hint: a general solution to $mu'' + cu' + ku = 0$ is of the form

$$u(t) = c_1 e^{-\frac{c}{2m}t} \cos(\omega t) + c_2 e^{-\frac{c}{2m}t} \sin(\omega t).$$

4.5.3 Nondimensionalization: Logistic Equation Example

We now examine a technique known as **nondimensionalizing** (sometimes also known as **scaling**) the variables that appear in an ODE. This technique involves replacing variables with dimensions like time, length, etc., by equivalent **dimensionless variables**. As we shall see, nondimensionalizing an ODE can greatly simplify analysis and reduce the number of parameters involved. It can highlight commonality in problems of vastly different scales, such as a skyscraper swaying in the breeze and a quartz crystal in an electronic circuit. It may also give insight into which terms in an ODE are important and which terms might be ignored, potentially simplifying analysis of the equation. This technique makes use of the characteristic variable scales we can find using the techniques above. In this section and the next we consider two examples of this technique.

For our first example, recall the logistic equation (1.10) for population growth,

$$\frac{du}{dt} = ru(t)(1 - u(t)/K). \quad (4.87)$$

Here $u(t)$ is the population of some species as a function of time, r is the intrinsic growth rate of the species, and K is the maximum number of individuals that the environment can support, called the carrying capacity of the environment. Let us **nondimensionalize** this ODE.

Dimension of the Variables

The function $u(t)$ quantifies how many individuals of the species are present, and so might be considered dimensionless. However, as was noted in Exercise 1.5.6, it can be helpful to assign a dimension to the species count, say N . Thus we have $[u] = N$ and, in particular, $[K] = N$. This makes the term u/K in (4.87) dimensionless, so that $(1 - u/K)$ is well-defined and dimensionless (since 1 is dimensionless). Then $[du/dt] = NT^{-1}$ and dimensional consistency in (4.87) demands $[r] = T^{-1}$.

Characteristic Scales

There are two parameters, r and K in (4.87), with dimensions $[r] = T^{-1}$ and $[K] = N$. To form characteristic time and population scales t_c and u_c , respectively, consider the expressions

$$t_c = r^\alpha K^\beta \quad (4.88)$$

$$u_c = r^\gamma K^\delta. \quad (4.89)$$

For $[t_c] = T$ dimensional consistency in (4.88) leads to $T = (T^{-1})^\alpha N^\beta = T^{-\alpha} N^\beta$, which gives $\alpha = -1$ and $\beta = 0$. Then the characteristic time scale is $t_c = 1/r$. The condition $[u_c] = N$ in (4.89) gives $N = (T^{-1})^\gamma N^\delta = T^{-\gamma} N^\delta$, which leads to $\gamma = 0$ and $\delta = 1$. The characteristic population scale is $u_c = K$. In summary, the characteristic variable scales for the independent (time) and dependent (population) variables here are

$$t_c = 1/r \quad \text{and} \quad u_c = K. \quad (4.90)$$

This may seem rather clear after the fact.

By performing the analysis that led to (4.90), we've already deduced something important about this ODE: we expect that the solution to (1.10) changes on a time scale comparable to $1/r$, and the population exists on a scale comparable to K . There is no need to solve the ODE to figure this out.

Dimensionless Variables

The next step is to nondimensionalize both the independent (time) and dependent (population) variables. This process involves replacing the time variable t and population variable u with equivalent rescaled variables that are (like radians, for example) nondimensional. If done correctly, this simplifies the ODE under consideration and reduces the number of physical parameters.

To illustrate in the present example, we use $t_c = 1/r$ to define a new time-like independent variable τ as

$$\tau = t/t_c = rt \quad \text{or} \quad t = t_c\tau = \tau/r. \quad (4.91)$$

The variable τ is dimensionless since $[t] = T$ and $[t_c] = T$, so $[\tau] = [t/t_c] = TT^{-1} = 1$. Thus τ is a sort of rescaled or dimensionless time. For example, if the ODE has a characteristic time scale $t_c = 100$ days and we're interested in the solution $u(t)$ at time $t = 500$ days, this corresponds to $\tau = 5$; here τ would measure time in increments of 100 days, instead of single days like t .

We use $u_c = K$ to define a new dimensionless dependent variable, a dimensionless population \bar{u} , as

$$\bar{u} = u/u_c = u/K \quad \text{or} \quad u = u_c\bar{u} = K\bar{u}. \quad (4.92)$$

Again, this simply changes the scale for measuring population. If $u_c = 10^6$ then \bar{u} is the population measured in millions.

This process of nondimensionalizing variables as in (4.91) or (4.92) is sometimes also called **rescaling**, since it involves rescaling the independent and dependent variables in an ODE.

Nondimensionalizing the ODE

After defining nondimensional variables, the next step is to rewrite the ODE in terms of these variables. The function $u(t)$ will be replaced by its equivalent version with dimensionless variables, here $\bar{u}(\tau)$. The relation between $\bar{u}(\tau)$ and $u(t)$ is simple: $\bar{u}(\tau) = u(t)/u_c$, or

$$\bar{u}(\tau) = u(t)/K, \quad (4.93)$$

making use of $u_c = K$. That is, \bar{u} measures the population in increments of size K , and time is measured with respect to τ instead of t . Equivalently, multiply both sides of (4.93) by u_c and find

$$u(t) = K\bar{u}(\tau). \quad (4.94)$$

We will replace $u(t)$ in the logistic ODE (4.87) with the function $\bar{u}(\tau)$. To do this we need to determine how the first derivative of u relates to that of \bar{u} .

Start with (4.94) and differentiate both sides with respect to t to find

$$\begin{aligned} \frac{du}{dt}(t) &= \frac{d}{dt}(K\bar{u}(\tau)) \\ &= K \frac{d\bar{u}}{d\tau}(\tau) \frac{d\tau}{dt} \quad (\text{use the chain rule}) \\ &= \frac{K}{t_c} \frac{d\bar{u}}{d\tau}(\tau) \quad (\text{since } d\tau/dt = 1/t_c) \\ &= rK \frac{d\bar{u}}{d\tau}(\tau) \quad (\text{since } t_c = 1/r). \end{aligned} \quad (4.95)$$

To nondimensionalize the ODE (4.87) use (4.94) to replace each $u(t)$ with $K\bar{u}(\tau)$ and use (4.95) to replace $\frac{du}{dt}(t)$ with $rK\frac{d\bar{u}}{d\tau}(\tau)$. This yields

$$rK \frac{d\bar{u}}{d\tau}(\tau) = rK\bar{u}(\tau)(1 - K\bar{u}(\tau)/K).$$

A few fortuitous cancellations and some algebra show that

$$\frac{d\bar{u}}{d\tau}(\tau) = \bar{u}(\tau)(1 - \bar{u}(\tau)). \quad (4.96)$$

All of the physical parameters have disappeared.

The ODE (4.96) is a rescaled, dimensionless version of the original ODE (4.87). We can move back and forth between solutions $\bar{u}(\tau)$ to (4.96) and solutions $u(t)$ to (4.87) using the correspondences

$$u(t) = u_c \bar{u}(\tau) = K \bar{u}(rt)$$

or

$$\bar{u}(\tau) = u(t)/u_c = u(\tau/r)/K.$$

For example, a general solution $\bar{u}(\tau) = 1/(1 + Ce^{-\tau})$ to (4.96) is fairly easy to find with separation of variables. Since $u(t) = K \bar{u}(rt)$, it follows that

$$u(t) = K \bar{u}(rt) = \frac{K}{1 + Ce^{-rt}}.$$

In addition to clarifying the variable scales for the ODE, there is another advantage of nondimensionalization: aside from the rescaling in time and population, the solution to (4.87) will behave, qualitatively, like the solution to (4.96), and the latter equation might be easier to analyze. But the real power of rescaling is best illustrated by applying nondimensionalization to the harvested logistic equation, as we do in the next section.

Reading Exercise 4.5.5 Suppose $u(t) = t^2$, $r = 1/3$ (so $t_c = 3$), and $u_c = K = 7$. Use (4.93) to write out $\bar{u}(\tau)$ explicitly as a function of τ . Then verify (4.95).

4.5.4 Nondimensionalization: Harvested Logistic Equation Example

Let's reconsider the logistic ODE of Section 4.5.3, but this time with constant harvesting at a rate of h individuals per unit time, which can be modeled by subtracting an h from the right side of (4.87) to obtain

$$\frac{du}{dt} = ru(t)(1 - u(t)/K) - h. \quad (4.97)$$

Unlike equation (1.12), the harvesting here is not proportional to $u(t)$, but at a flat rate. The ODE (4.97) does have an analytical solution, and it is significantly more complicated than the solution to the standard logistic equation (4.87) in which $h = 0$. Under what conditions can the harvesting rate h be taken as zero and the solution to (4.87) used as a reasonable approximation to the solution to (4.97)? This is can be illuminated by nondimensionalizing (4.97).

Characteristic Scales

We begin by finding characteristic scales for the independent and dependent variables. As in Section 4.5.3, $[r] = T^{-1}$ and $[K] = N$, where N denotes the dimension population. The harvesting parameter h has dimension $[h] = NT^{-1}$. Form a time scale $t_c = r^\alpha K^\beta h^\gamma$, which leads to $T = T^{-\alpha-\gamma} N^{\beta+\gamma}$. Matching dimensions yields the equations $-\alpha - \gamma = 1$ and $\beta + \gamma = 0$, which have infinitely many solutions. Let us take γ as a free variable and then solve for $\alpha = -1 - \gamma$ and $\beta = -\gamma$. Thus anything of the form

$$t_c = r^{-1-\gamma} K^{-\gamma} h^\gamma = \frac{1}{r} \left(\frac{h}{rK} \right)^\gamma \quad (4.98)$$

can be used as a characteristic time scale, with any choice of γ . Note that the quantity $1/r$ on the right side in (4.98) has the dimension time, while $h/(rK)$ and any power thereof is dimensionless. We saw a similar structure in equation (4.86). Different choices for γ lead to different time scales that emphasize different aspects of system's evolution when the ODE is rescaled. Choosing a useful scale can be a bit hit and miss, and may require some experimentation. In this case let us make the simple choice $\gamma = 0$, which yields $t_c = 1/r$ as before.

Reading Exercise 4.5.6 By considering $u_c = r^\alpha K^\beta h^\gamma$, show that a characteristic population scale u_c is of the form

$$u_c = K \left(\frac{h}{rK} \right)^\gamma$$

for some γ .

Using $\gamma = 0$ in Reading Exercise 4.5.6 yields the same characteristic population scale $u_c = K$ that was found in Section 4.5.3. Let's go with these choices for t_c and u_c , and see where they lead.

Rescaling the ODE

With the choices $u_c = K$ and $t_c = 1/r$, equations (4.93), (4.94), and (4.95) remain valid. Replace each $u(t)$ with $K\bar{u}(\tau)$ and $\frac{du}{dt}(t)$ with $rK\frac{d\bar{u}}{d\tau}(\tau)$ in (4.97) to find

$$rK \frac{d\bar{u}}{d\tau}(\tau) = rK\bar{u}(\tau)(1 - K\bar{u}(\tau)/K) - h.$$

Divide through by rK to obtain

$$\frac{d\bar{u}}{d\tau}(\tau) = \bar{u}(\tau)(1 - \bar{u}(\tau)) - \varepsilon. \quad (4.99)$$

where $\varepsilon = \frac{h}{rK}$; note that ε is dimensionless, as it must be in order to be subtracted from \bar{u} . Compare (4.99) to (4.96). The solutions to (4.99) correspond to those of (4.97) via the relations $t = t_c\tau = \tau/r$ and $u(t) = u_c\bar{u}(\tau) = K\bar{u}(\tau)$.

Interpretation of the Nondimensional ODE and Conclusions

The nondimensionalized ODE (4.99) lets us make insightful conclusions about when harvesting is significant and when it is not by showing how much h , in conjunction with r and K , influences the solution. When $\varepsilon = \frac{h}{rK}$ is close to zero, the solution to (4.99) should be close to the solution in which $\varepsilon = 0$, which corresponds to no harvesting. To illustrate, Figure 4.28 shows the solution to (4.99) when $\varepsilon = 0.05$ (solid red curve) and $\varepsilon = 0$ (dashed black curve), both with $\bar{u}(0) = 0.2$.

Based on Figure 4.28, if the solution with $\varepsilon = 0$ is deemed to be a sufficiently good approximation to the solution when $\varepsilon \leq 0.05$ then perhaps the no-harvesting logistic model ($\varepsilon = 0$) can be used in place of the more complex harvested model. Back in the original ODE (4.97) the condition $\varepsilon \leq 0.05$ becomes the condition

$$\frac{h}{rK} \leq 0.05 \quad (4.100)$$

since $\varepsilon = \frac{h}{rK}$. This gives a quantitative criteria for how small h should be in order to drop the harvesting term: it is not the actual value of h that matters, it is the size of h in relation to the product rK . On the other hand, if the threshold 0.05 on the right in (4.100) does not provide sufficiently good agreement with solutions obtained using $\varepsilon = 0$, then a smaller threshold for ε can be used. In any case it is the dimensionless quantity $\frac{h}{rK}$ that is relevant when deciding whether to ignore the harvesting term.

Reading Exercise 4.5.7 Suppose $K = 10^6$ and $r = 0.03$ (units reciprocal years). What value of h will satisfy (4.100)? If r increases, how does the allowable value of h change? Why does that make sense? If K increases, how does the allowable value of h change? Why does that make sense?

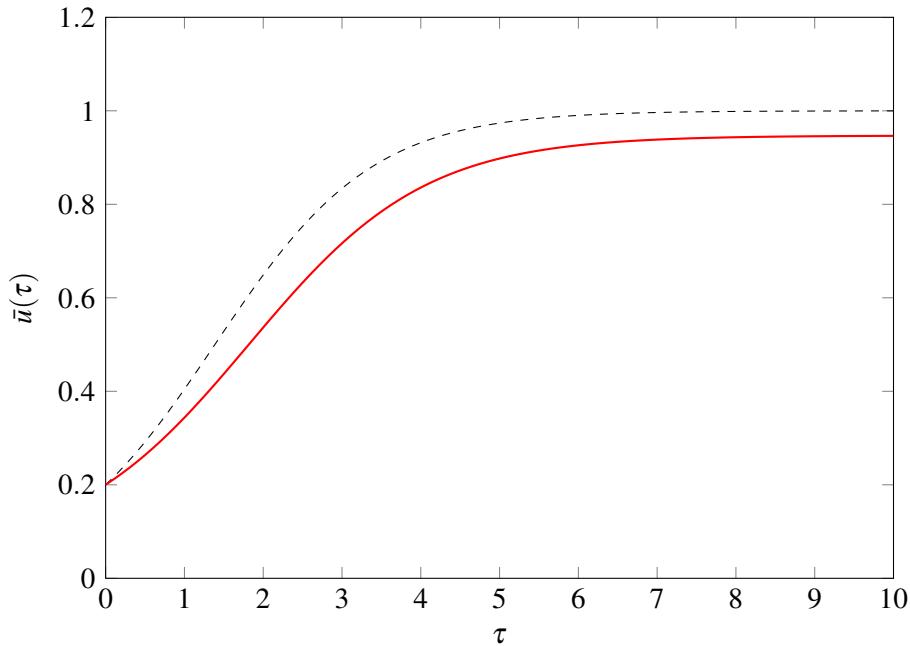


Figure 4.28: Solution to (4.99) with $\varepsilon = 0$ (dashed black curve) and with $\varepsilon = 0.05$ (solid red curve), both with $\bar{u} = 0.2$.

4.5.5 The General Outline for Nondimensional Rescaling

The techniques used in Sections 4.5.3 and 4.5.4 work in more general circumstances. Consider an ODE in which the various coefficients appear in unspecified or symbolic form, e.g., the stiffness of a spring is k (and not 3.2 newtons per meter). Suppose the dependent variable is $u(t)$. To nondimensionalize the ODE:

1. Identify the physical parameters that appear in the ODE, such as masses, growth rates, etc. The initial condition(s) may also come into play, if any are specified (we'll look at such an example shortly.) Find the dimension of each of these quantities.
2. Find the possibilities for a characteristic time scale t_c for the problem, with t_c as a product of powers of the ODE parameters. There may be a unique time scale, as in Section 4.5.3, or many, as in Section 4.5.4. Similarly construct a characteristic scale u_c for the dependent variable. If there is more than one possibility for either scale, you may have to experiment to find one that gives the insights you seek. See the exercises at the end of the section for some examples. In any case, understanding the scales present in the problem provides valuable insight.
3. Define a dimensionless time variable τ according to

$$\tau = t/t_c \quad \text{or} \quad t = t_c\tau. \quad (4.101)$$

Next define a dimensionless dependent variable \bar{u} as

$$\bar{u}(\tau) = u(t)/u_c \quad \text{or} \quad u(t) = u_c\bar{u}(\tau). \quad (4.102)$$

From $u(t) = u_c\bar{u}(\tau)$, $\tau = t/t_c$, and the chain rule it follows that

$$\frac{du}{dt} = \frac{u_c}{t_c} \frac{d\bar{u}}{d\tau}. \quad (4.103)$$

If the second derivative d^2u/dt^2 is needed in terms of \bar{u} then differentiating both sides of (4.103) with respect to t again and using the chain rule shows that

$$\frac{d^2u}{dt^2} = \frac{u_c}{t_c^2} \frac{d^2\bar{u}}{d\tau^2}. \quad (4.104)$$

Higher derivatives can also be computed if needed.

4. In the original unscaled ODE use (4.102)-(4.104) to replace $u(t)$ by $u_c\bar{u}(\tau)$, du/dt by $\frac{u_c}{t_c} \frac{d\bar{u}}{d\tau}$, and d^2u/dt^2 by the right side of (4.104). An initial condition $u(0) = u_0$ becomes $\bar{u}(0) = u_0/u_c$, and $\frac{du}{dt}(0) = v_0$ becomes $\frac{u_c}{t_c} \frac{d\bar{u}}{d\tau}(0) = v_0$ or $\frac{d\bar{u}}{d\tau} = t_c v_0/u_c$. Note that the initial data for \bar{u} is dimensionless.
5. Simplify the nondimensional ODE. Examine the remaining terms in the nondimensional ODE (like ϵ in (4.99)). This will help provide insight into what combinations of parameters influence the solution, which terms in the ODE are the most important, and perhaps even allow you to drop some terms.

This technique is applicable to ODEs that have independent variables other than t , and is also applicable to partial differential equations.

4.5.6 Back to the Hard Spring

As a final example, let's return to the hard spring model (4.82), nondimensionalize the ODE, and examine conditions under which the cubic term in the model may be omitted. As already noted, one brute force approach would be to numerically experiment, varying each of m, k_1, k_2, u_0 and possibly the solution interval $0 \leq t \leq T$, in all combinations, to figure out which combinations give good agreement between the linear and nonlinear equations, but this would require a huge amount of work. By rescaling and nondimensionalizing the ODE we can greatly reduce the amount of effort necessary while gaining insight into what factors make the cubic term important and what factors don't. For example, the mass m turns out to be irrelevant.

Characteristic Scales

There are three parameters, m, k_1 , and k_2 involved in (4.82). Their dimensions are

$$[m] = M, \quad [k_1] = MT^{-2}, \quad [k_2] = MT^{-2}L^{-2}.$$

The dimensions of m and k_1 have been derived many times before. To compute $[k_2]$ note that from (4.81), $[k_2]$ must have dimension force divided by length cubed. Based on Figure 4.27 we also expect that the magnitude of u_0 is important, for if $|u_0|$ is large then $|k_1 u_0| \ll |k_2 u_0^3|$ and the nonlinearity of the spring really comes into play immediately. The initial velocity v_0 could also be thrown into the mix, but for simplicity let's work under the assumption that $v_0 = 0$.

The first step is construct a characteristic time scale of the form

$$t_c = m^\alpha k_1^\beta k_2^\gamma u_0^\delta. \quad (4.105)$$

It's straightforward to verify (see Exercise 4.5.10) by matching dimensions on the left and right in (4.105) that any characteristic time scale is of the form

$$t_c = m^{1/2} k_1^{-1/2-\gamma} k_2^\gamma u_0^{2\gamma} = \sqrt{\frac{m}{k_1}} \left(\frac{k_2 u_0^2}{k_1} \right)^\gamma \quad (4.106)$$

for some choice of γ . Note that $\sqrt{m/k}$ has the dimension of time, while the parenthesized expression $k_2 u_0^2/k_1$ on the right in (4.106) is dimensionless. This same general structure for t_c was present in (4.86) and (4.98), in which there were infinitely many choices for t_c , each of the form $t^* q^\gamma$ where t^*

has the dimension of time, q is a dimensionless constant formed from the parameters in the ODE, and γ is an arbitrary free variable.

Given that there are infinitely many time scales, which one is appropriate? In Example 4.29 the choice $t_c = \sqrt{m/k}$ was used, corresponding to $\gamma = 0$ in (4.106). Moreover, $\sqrt{m/k}$ captures the period of the linear system, and, based on Figure 4.27, it seems that the main difference in the linear and nonlinear problems is the period of the solution. It then makes sense to use $\gamma = 0$ in (4.106), so

$$t_c = \sqrt{m/k_1}.$$

If this time scale provides no insight or simplifications, we'll amend it and try again.

For a characteristic length scale u_c consider

$$u_c = m^\alpha k_1^\beta k_2^\gamma u_0^\delta.$$

Analysis similar to that done for t_c leads to

$$u_c = u_0 \left(\frac{k_2 u_0^2}{k_1} \right)^\gamma \quad (4.107)$$

for any choice of γ , which again has the same dimensionless quantity $k_2 u_0^2 / k_1$ —there is a deeper reason. We can freely choose γ , and one obvious choice is $\gamma = 0$, which leads to

$$u_c = u_0.$$

As with our choice for t_c , if this choice for u_c in the rescaled ODE yields no insight or simplification, we can go back and try something else.

The Nondimensional ODE

In accordance with the general procedure of (4.101), set $\tau = t/t_c = t\sqrt{k_1/m}$ (or $t = \tau\sqrt{m/k_1}$). Use (4.102) to define a rescaled dependent variable as $\bar{u}(\tau) = u(t)/u_0$, or $u(t) = u_0 \bar{u}(\tau)$ (using $u_c = u_0$). Then from (4.104) it follows that

$$\frac{d^2u}{dt^2}(t) = \frac{u_0 k_1}{m} \frac{d^2\bar{u}}{d\tau^2}(\tau).$$

With these substitutions the ODE (4.82) for $\bar{u}(\tau)$ can be written as

$$\frac{d^2\bar{u}}{d\tau^2} + \bar{u} + \varepsilon \bar{u}^3 = 0, \quad (4.108)$$

where

$$\varepsilon = \frac{k_2 u_0^2}{k_1}. \quad (4.109)$$

The quantity ε in (4.109) is the same dimensionless quantity that appeared in (4.106) and (4.107). The initial condition $u(0) = u_0$ becomes $\bar{u}(0) = u(0)/u_c = u_0/u_c = 1$. The solutions $u(t)$ to (4.82) with $u(0) = u_0$ and the solution $\bar{u}(\tau)$ to (4.108) with $\bar{u}(0) = 1$ are in precise correspondence, with the relations $t = t_c \tau = \tau\sqrt{m/k_1}$ and $u(t) = u_c \bar{u}(\tau)$.

Let's look at the role of ε a bit more closely. Assume $u_0 \neq 0$, for otherwise the solution to (4.108) is $u(t) = 0$ for all t , which is uninteresting. With this assumption, $\varepsilon = 0$ corresponds to $k_2 = 0$, so there is no cubic term in (4.82) or (4.108). But if $\varepsilon > 0$ and is sufficiently close to zero then the solution to (4.108) should be close to the solution to $\frac{d^2\bar{u}}{d\tau^2} + \bar{u} = 0$ in some reasonable sense. Therefore the solution to (4.82) with the corresponding values of k_1, k_2 , and u_0 will be close to

the solution to $mu''(t) + ku(t) = 0$. For example, in the left panel of Figure 4.27 we showed the solution to (4.82) with $m = 10, k_1 = 5, k_2 = 0.25$, and initial position $u_0 = 1$ (solid red curve) which corresponds to $\varepsilon = 0.05$. We also showed the solution when $k_2 = 0$, which corresponds to $\varepsilon = 0$ (dashed black curve). The right panel showed the solution when $k_2 = 2$ (other parameters the same as the left panel), corresponding to $\varepsilon = 0.4$, along with the solution when $k_2 = 0$.

The analysis illustrates a quantitative and efficient way to determine when the cubic term in (4.82) matters when solving on an interval $0 \leq t \leq T$. It boils down to a condition of the form $\varepsilon \leq \varepsilon_0$ for the nondimensional ODE (4.108) or, equivalently, the condition

$$\frac{k_2 u_0^2}{k_1} \leq \varepsilon_0 \quad (4.110)$$

for the ODE (4.82), where ε_0 is some threshold that depends on the level of agreement that we desire between solutions to the nonlinear ODE and solutions to the linearized version. This threshold can be determined numerically, by solving (4.108) with $\varepsilon = 0$ and then with $\varepsilon = \varepsilon_0$ for some chosen $\varepsilon_0 > 0$. If the agreement is good enough, use this ε_0 as the condition in (4.110). If not, decrease ε_0 and try again. We should note that if the time interval is $0 \leq t \leq T$ in the original ODE then the interval for (4.108) is $0 \leq \tau \leq T/t_c$.

■ **Example 4.31** Here's an illustration. Consider (4.82) with $m = 200, k_1 = 10$, and $u_0 = 2$, on the time interval $0 \leq t \leq 200$. How small must k_2 be for the solution with the cubic term $k_2 u^3$ present in (4.82) to agree well with the solution that omits the cubic term? The time scale here is $t_c = \sqrt{m/k_1} = \sqrt{20}$, so for the nondimensional ODE we'll work on the interval $0 \leq \tau \leq 200/\sqrt{20} \approx 44.7$. When $\varepsilon = 0$ the solution to (4.108) is $\bar{u}(\tau) = \cos(\tau)$. A numerical solution to (4.108) with $\varepsilon = 0.02$ is shown in Figure 4.29 as the solid red curve, and the solution with the cubic term omitted ($\varepsilon = 0$) is shown as the dashed black curve. If this agreement is deemed good enough then $\varepsilon_0 = 0.02$

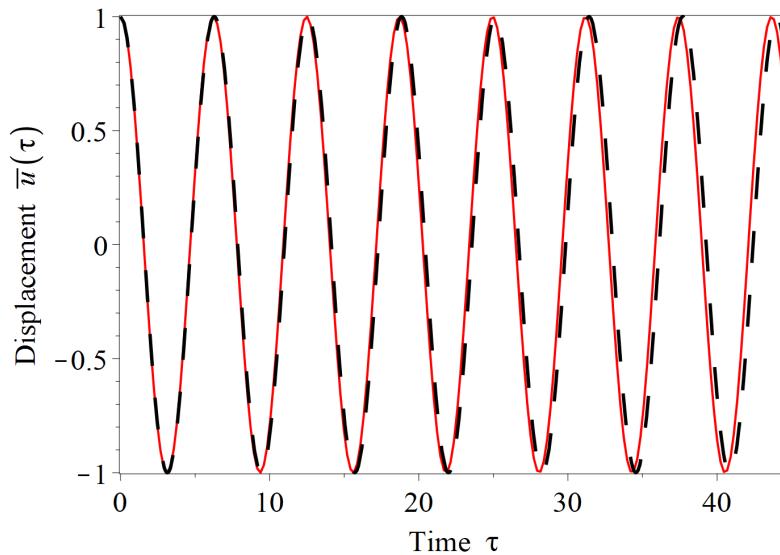


Figure 4.29: Solution to (4.108) with $\varepsilon = 0.02$ and initial data $\bar{u}(0) = 1, \bar{u}'(0) = 0$ (solid red curve) and solution to (4.108) with cubic term omitted (dashed black curve).

can be used in (4.110). With $m = 200, k_1 = 10$, and $u_0 = 2$ the condition for k_2 in (4.82) becomes $\frac{k_2 u_0^2}{k_1} \leq 0.02$, or $k_2 \leq 0.05$. Notice that in all the analysis above for the hard spring, the mass m was not relevant, an important insight. ■

Summary

Rescaling through the introduction of dimensionless variables can make the analysis and solution of ODEs (and other problems in applied mathematics) much easier. In particular

- Appropriate rescaling can illuminate the nature of the solution to an ODE by indicating the natural scale for the quantities of interest, in time, space, or other physical dimensions.
- Rescaling allows us to determine when and how equations can be simplified by dropping negligible terms.
- Rescaling can aid in the numerical solution of a problem. For example, when solving (4.82) numerically for a large selection of different values of m , k_1 , k_2 , and u_0 , the introduction of dimensionless variables shows that we really only need to solve (4.108) for the relevant values of ε , which may lead to much less work.
- The introduction of dimensionless variables is an essential part of many types of experimentation. For example, you can't easily put a full-sized jumbo jet into a wind tunnel; you'd use a scale model. But will the results for the scale model correspond to the behavior of a full-sized jet? Dimensional analysis of the type we've done can help assure that it will. See [20] for more on this topic and dimensional analysis in general.

4.5.7 Exercises

Exercise 4.5.1 Suppose $u'(t) = -ku(t)$ with initial condition $u(0) = u_0$ governs some physical process, where u is a mass, so $[u] = M$. Of course t is time, so $[t] = T$. Find the dimensions of k and show that any characteristic time scale of the form $t_c = k^\alpha u_0^\beta$ is given by $t_c = 1/k$. Show that any characteristic mass scale of the form $u_c = k^\alpha u_0^\beta$ is given by $u_c = u_0$. Then show that the nondimensionalized ODE is $d\bar{u}/dt = -\bar{u}$ with initial data $\bar{u}(0) = 1$.

Exercise 4.5.2 Suppose $u'(t) = -ku^2(t)$ with initial condition $u(0) = u_0$ governs some physical process, where u is a mass, so $[u] = M$. Of course t is time, so $[t] = T$. Find the dimensions of k and show that any characteristic time scale of the form $t_c = k^\alpha u_0^\beta$ is given by $t_c = 1/(ku_0)$. Show that any characteristic mass scale of the form $u_c = k^\alpha u_0^\beta$ is given by $u_c = u_0$. Then show that the nondimensionalized ODE is $d\bar{u}/dt = -\bar{u}^2$ with initial data $\bar{u}(0) = 1$.

Solve the ODE $u'(t) = -ku^2(t)$ with initial condition $u(0) = u_0$ and compute how long it takes for the solution $u(t)$ to decay from the initial amount u_0 to half of that amount, $u_0/2$. Is this in accord with the time scale t_c ?

Exercise 4.5.3 Newton's law of cooling is the ODE $u'(t) = -k(u(t) - A)$, where $u(t)$ is the temperature of some object in an environment with ambient temperature A and $k > 0$ is some constant; assume $A \neq 0$ for this problem. We'll use Θ for the dimension temperature, so $[A] = \Theta$. Deduce the dimension of k , find a characteristic time scale t_c and characteristic temperature scale u_c for this ODE in terms of the parameters k and A . Then nondimensionalize the ODE. What does an initial condition $u(0) = u_0$ become, in terms of the rescaled dependent variable \bar{u} ? To what feature of the solution to the Newton cooling ODE does the characteristic scale u_c correspond?

Exercise 4.5.4 The Hill-Keller ODE was $v'(t) = P - kv(t)$, where $v(t)$ is the velocity of a sprinter, $P > 0$ is the maximum acceleration the sprinter is capable of (from a standing start), and $k > 0$ is some constant. Deduce the dimension of k , find characteristic time scale t_c and

velocity scale v_c for this ODE in terms of P and k , and then nondimensionalize the ODE. What does the initial condition $v(t_0) = 0$ become? To what feature of the solution to the Hill-Keller ODE does the characteristic scale v_c correspond?

Exercise 4.5.5 Recall the salt tank ODE $u'(t) = rc_1 - \frac{r}{V}u(t)$ from Example 1.5 in Chapter 1. Here r is the rate (volume per time) of fluid entering and exiting a tank of volume V , with c_1 as the concentration of some substance (mass per volume) and $u(t)$ is the amount of the substance in the tank at time t , measured on a per mass basis. Find the dimensions of r, c_1 , and V , and then use them to show that the unique characteristic time scale of the form $t_c = V^\alpha b^\beta c_1^\gamma$ is $t_c = V/r$ and the unique characteristic mass scale is $u_c = c_1 V$. Use these to nondimensionalize the ODE.

Exercise 4.5.6 A pendulum of length R swings without friction in an environment with gravitational acceleration g . The mass of the bob at the end of the pendulum is m .

- (a) Let t_c denote the characteristic time scale for the problem and suppose that

$$t_c = R^\alpha g^\beta m^\gamma$$

for constants α, β , and γ . Find the dimension of each of R, g , and m , and show that the characteristic time scale for this problem is given by $t_c = \sqrt{R/g}$. Thus the time scale does not depend on the mass of the bob—already a valuable piece of information. Note that you don't actually need to use the ODE that governs a swinging pendulum to deduce this.

- (b) The ODE for the angle $\theta(t)$ that the pendulum makes with the vertical is $\frac{d^2\theta}{dt^2} + \frac{g}{R}\theta(t) = 0$. Since θ is already nondimensional (it's an angle) we will just use $\theta_c = 1$ (dimensionless) for the characteristic angle. Nondimensionalize this ODE using $t_c = \sqrt{R/g}$ and $\theta_c = 1$.

Exercise 4.5.7 The charge $q(t)$ on a capacitor in a series RC circuit with no voltage source obeys

$$R \frac{dq}{dt} + \frac{q(t)}{C} = 0.$$

Suppose the capacitor has initial charge $q(0) = q_0$. Find a characteristic time scale t_c and a characteristic charge scale q_c for this circuit in terms of R, C , and q_0 , and then show that the nondimensionalized ODE is

$$\frac{d\bar{q}}{d\tau} + \bar{q}(\tau) = 0$$

with $\bar{q}(0) = 1$. Hint: recall that $[R] = ML^2T^{-1}Q^{-2}$ and $[C] = M^{-1}L^{-2}T^2Q^2$, with Q as the dimension of electric charge.

Exercise 4.5.8 The charge $q(t)$ on a capacitor in a series RLC circuit with no voltage source obeys

$$L \frac{d^2q}{dt^2} + R \frac{dq}{dt} + \frac{q(t)}{C} = 0.$$

Suppose the capacitor has initial charge $q(0) = q_0$ and the current in the circuit is $q'(0) = 0$.

- (a) Show that any characteristic time scale t_c of the form $t_c = L^\alpha R^\beta C^\gamma q_0^\delta$ is in fact given by

$$t_c = \sqrt{LC} \left(\frac{CR^2}{L} \right)^{\beta/2}$$

for some β . Show that any charge scale q_c is of the form

$$q_c = q_0 \left(\frac{CR^2}{L} \right)^{\beta/2}.$$

Hint: see the hint from the last problem, and use $[L] = ML^2Q^{-2}$ (where L in this case is being used for both inductance and the dimension of length, so try not to confuse them).

- (b) With characteristic time scale $t_c = \sqrt{LC}$ ($\beta = 0$ above) and charge scale $q_c = q_0$ (also $\beta = 0$) show that the nondimensionalized ODE can be written as

$$\frac{d^2\bar{q}}{d\tau^2} + \gamma \frac{d\bar{q}}{d\tau} + \bar{q}(\tau) = 0$$

with initial conditions $\bar{q}(0) = 1$ and $\frac{d\bar{q}}{d\tau}(0) = 0$, where $\gamma = R\sqrt{C/L}$.

Exercise 4.5.9 A modification to the logistic equation that has been used for population modeling is

$$\frac{du}{dt} = -ru \left(1 - \frac{u}{K} \right) \left(1 - \frac{u}{P} \right) \quad (4.111)$$

where r is the population intrinsic growth rate and K is the carrying capacity, as we've previously studied, while P is a population size that satisfies $0 < P < K$. The idea is that P is the minimum sustainable population, below which the species becomes extinct.

- (a) Sketch a phase portrait for (4.111) and verify that if $u(t)$ satisfies $0 < u < P$, then solutions decay to zero, while if $u > P$ solutions approach K .
- (b) If we use N to denote the dimension of population then $[K] = [P] = N$. What is $[r]$ here?
- (c) Show that the only characteristic time scale of the form $t_c = r^\alpha K^\beta P^\gamma$ is $t_c = r^{-1}(K/P)^\beta$. We will take $\beta = 0$ so that $t_c = 1/r$. If we sample the population at periodic times (perhaps to collect data for parameter estimation), what implication does this have for how often we should sample?
- (d) Show any characteristic population scale of the form $u_c = r^\alpha K^\beta P^\gamma$ is $u_c = K^\beta P^\gamma$ where $\beta + \gamma = 1$ or equivalently $u_c = K(P/K)^\gamma$.
- (e) Nondimensionalize (4.111) using $t_c = 1/r$ and $u_c = K$ (corresponding to $\gamma = 0$ in $u_c = K(P/K)^\gamma$). What does an initial condition $u(0) = u_0$ become in the nondimensional problem?
- (f) Nondimensionalize (4.111) using $t_c = 1/r$ and $u_c = P$ (corresponding to $\gamma = 1$ in $u_c = K(P/K)^\gamma$). What does an initial condition $u(0) = u_0$ become in the nondimensional problem?
- (g) Consider a harvested version of equation (4.111) of the form

$$\frac{du}{dt} = -ru \left(1 - \frac{u}{K} \right) \left(1 - \frac{u}{P} \right) - hu \quad (4.112)$$

in which $h > 0$ is a harvesting rate. The population is harvested at a rate proportional to

the population. Using $t_c = 1/r$ and $u_c = K$, show that the nondimensionalized function $\bar{u}(\tau) = u(t)/K$ (where $\tau = t/t_c = rt$) satisfies the ODE

$$\frac{d\bar{u}}{d\tau} = -\bar{u}(1-\bar{u}) \left(1 - \frac{K}{P}\bar{u}\right) - \varepsilon\bar{u}, \quad (4.113)$$

where $\varepsilon = h/r$.

- (h) Suppose that in the harvested ODE (4.112) we have $P = K/10$. Write out the ODE (4.113) in this case. Show that if $\varepsilon > 2.025$, then all solutions to (4.113) that start with $u(0) > 0$ converge to zero; it might help to find the fixed points for (4.113) and sketch a phase portrait that shows the dependence on ε . What bound does this put on h/r in order to avoid extinction in (4.112)?

Exercise 4.5.10 In Example 4.31, show that any characteristic time scale t_c is of the form given in (4.106) and any characteristic length scale is of the form given in (4.107).

Exercise 4.5.11 Suppose an object of mass m falls at velocity $v(t)$ under the influence of gravitational acceleration g . Let's take $v > 0$ to indicate downward motion, which is all we're interested in. As it falls the mass experiences a drag force $F(v)$ that is a function of the object's velocity v . Suppose that $F(v) = -k_1v - k_2v^2$ for some positive constants k_1 and k_2 , so that the force is always upward when $v > 0$. From $F = ma$ we find

$$m \frac{dv}{dt} = mg - k_1v(t) - k_2v^2(t). \quad (4.114)$$

Note $g > 0$ here.

- (a) What are the dimensions of k_1 and k_2 ? Hint: $F(v)$ is a force, so k_1v and k_2v^2 must be forces as well.
 (b) Show that any characteristic time $t_c = m^\alpha g^\beta k_1^\gamma k_2^\delta$ that can be formed from m, g, k_1 , and k_2 is of the form

$$t_c = \frac{m}{k_1} \left(\frac{mgk_2}{k_1^2} \right)^{\delta_t}$$

for some choice of δ_t (note the quantity in parentheses is dimensionless).

- (c) Show that any characteristic velocity that can be formed from m, g, k_1 , and k_2 is of the form

$$v_c = \frac{mg}{k_1} \left(\frac{mgk_2}{k_1^2} \right)^{\delta_v}$$

for some choice of δ_v .

- (d) Nondimensionalize (4.114) using $t_c = \frac{m}{k_1}$ ($\delta_t = 0$ in (b)) and $v_c = \frac{mg}{k_1}$ ($\delta_v = 0$ in (c)). Show that this leads to an ODE for \bar{v} of the form

$$\frac{d\bar{v}}{d\tau} = 1 - \bar{v} - \varepsilon\bar{v}^2, \quad (4.115)$$

where $\varepsilon = mgk_2/k_1^2$ (dimensionless).

Find the analytical solution to (4.115) with initial data $\bar{v}(0) = 0$ in the case that $\varepsilon = 0$ (which corresponds to $k_2 = 0$). Call this solution $\bar{v}_0(\tau)$. Then compute (numerically or symbolically) the solution to (4.115) with $\bar{v}(0) = 0$ for each of $\varepsilon = 0.01, 0.1$, and 1.0 , and

then plot each along with $\bar{v}_0(\tau)$ for $0 \leq \tau \leq 5$. Experiment. How small must ε be before the solution to (4.115) agrees well with \bar{v}_0 on this nondimensional time interval?

Based on your observations, what is a reasonable quantitative criterion (involving m, g, k_1 , and k_2) for dropping the quadratic term in $F(v)$ and using the simpler linear ODE?

- (e) Show that the choice $\delta_t = -1/2$ for t_c and $\delta_v = -1/2$ for v_c leads to a rescaled ODE of the form

$$\frac{d\bar{v}}{d\tau} = 1 - \varepsilon\bar{v} - \bar{v}^2. \quad (4.116)$$

What is ε here? Mimic the computations in part (d) above to find a reasonable quantitative criterion (involving m, g, k_1 , and k_2) for dropping the linear term in $F(v)$ and using the ODE in which $F(v)$ is purely quadratic.

4.6 Modeling Projects

In this section we offer six modeling projects that incorporate ideas we've seen in this chapter. The last two concern different approaches to modeling a swinging pendulum, one based on conservation of energy, and one based on Newton's second law of motion. The latter incorporates friction into the model.

4.6.1 Project: Earthquake Modeling

Let us consider a more realistic model of a single-story building in an earthquake. Unlike previous models, the driver for the shaking of the building will be ground motion, instead of a force directly applied to the building's roof mass.

See Figure 4.30 in which a single-story building is modeled as in Section 4.1, as a point mass m suspended by walls that act as springs. In the present case, however, the building's foundation which rests upon the earth (the light gray rectangle) can move with respect to the horizontal as the ground moves in an earthquake; this horizontal axis provides a fixed inertial frame of reference, and we use x for the horizontal coordinate. Suppose the foundation moves according to $x = r(t)$ for some function $r(t)$. Let $u(t)$ denote the displacement of the roof mass m with respect to the foundation (not the x axis). This means that the position of m with respect to the x axis is $u(t) + r(t)$. Our goal is to derive the ODE that $u(t)$ obeys and then explore this model. The model will be based on Newton's second law of motion $F = ma$.

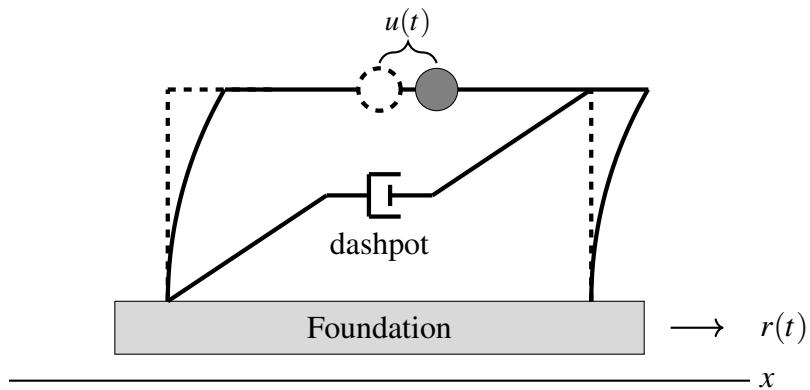


Figure 4.30: Simplified one-story building.

Modeling Exercise 6.1.1 Assume that the walls exert a force F_{walls} on the mass m according to the relative deflection of the walls with respect to the foundation, so this force is proportional to $u(t)$. Write down a reasonable expression for F_{walls} that depends on $u(t)$, using k for the constant of proportionality. Keep in mind Modeling Tip 1.1.1.

Modeling Exercise 6.1.2 Assume that the frictional force $F_{friction}$ on the roof mass is proportional to the relative velocity between the roof and the foundation (viscous damping), and hence this force is proportional to $u'(t)$. Why is this reasonable? Write down a reasonable expression for $F_{friction}$, using c for any constant of proportionality. Again, keep in mind Modeling Tip 1.1.1.

Modeling Exercise 6.1.3 The position of the roof mass m with respect to the inertial frame of reference, the x axis, is $u(t) + r(t)$. Use this along with Newton's second law of motion and the forces from Modeling Exercises 6.1.1 and 6.1.2 to justify the ODE

$$mu''(t) + cu'(t) + ku(t) = -mr''(t). \quad (4.117)$$

Modeling Exercise 6.1.4 As a quick sanity check, suppose $r(t) = 1$ (the foundation is at constant position 1 meter to the right of its zero position) and $u(0) = 0$ with $u'(0) = 0$. That is, the roof mass is at zero deflection with respect to the foundation. Why should $u(t) = 0$ for all t here? Does this choice satisfy (4.117)? Repeat this thought experiment if $r(t) = vt + b$ (the foundation is moving a constant speed v).

Modeling Exercise 6.1.5 Suppose that $m = 5000$ kg, $k = 5 \times 10^5$ newtons per meter, and $c = 5 \times 10^4$ newtons per meter per second. Also suppose that $r(t) = 0.02 \cos(\pi t)$, so the foundation moves back and forth with amplitude 0.02 meter and period 2 seconds. If $u(0) = u'(0) = 0$, solve (4.117) for $u(t)$. Plot the solution on the time interval $0 \leq t \leq 10$. How much do the building walls deflect from equilibrium? Does this seem reasonable?

Modeling Exercise 6.1.6 Show that this structure is underdamped. What is its natural frequency? Repeat Modeling Exercise 6.1.5 with $r(t) = 0.02 \cos(10t)$, which is close to the building's natural frequency. What displacement from equilibrium do the walls undergo? How does the amplitude of this displacement compare to that of Modeling Exercise 6.1.5?

Modeling Exercise 6.1.7 Compute the roof's periodic displacement response $u_p(t)$ with respect to the foundation when the foundation motion is $r(t) = 0.02 \cos(\omega t)$; leave ω unspecified. Then compute the amplitude of $u_p(t)$ as a function of ω . Plot this amplitude on the range $0 \leq \omega \leq 10\pi$ (0 to 5 hertz). What is the maximum amplitude of $u_p(t)$ in this frequency range?

Modeling Exercise 6.1.8 Suppose the building's roof is not designed to withstand a displacement of more than 0.05 meters relative to the foundation when being driven at any frequency in the range 0 to 5 hertz. What is the smallest damping coefficient (with m and k as already given) that will suffice?

Modeling Exercise 6.1.9 The displacement $u(t)$ of the building's roof may not be the only issue; the acceleration experienced by the building (and occupants) is also a concern. Recall that the position of the roof mass relative to the foundation is $u(t)$, and if the foundation itself is in motion with respect to the fixed x axis (a fixed inertial frame of reference) then the roof's position with respect to the x axis is $u(t) + r(t)$. This motion of the roof is what we will focus on, and in particular, the acceleration experienced by the roof mass, $u''_p(t) + r''(t)$. If $r(t) = 0.02 \cos(\omega t)$, compute the amplitude of the periodic acceleration response $u''_p(t) + r''(t)$ using $m = 5000$, $k = 5 \times 10^5$, and $c = 10^4$, as a function of ω . What is the maximum amplitude acceleration experienced by the roof in this frequency range?

Modeling Exercise 6.1.10 Suppose that in the setting of Modeling Exercise 6.1.9, the roof is not designed to withstand an acceleration in excess of 5 meters per second squared. With the same

values of m and k , find the smallest value of c that accomplishes this.

4.6.2 Project: Stay Tuned—RLC Circuits and Radios

A radio antenna may receive signals from many different stations, all operating at different frequencies. For example, in the United States the traditional commercial AM (amplitude modulation) radio frequency band spans 530 kilohertz to 1700 kilohertz. The radio waves from the various nearby transmitters induce tiny voltage differences across the span of the antenna, each at the frequency of the transmitting station. These signals are then processed and amplified to produce an audio signal. Given that signals from all nearby stations impinge on the antenna, why don't we hear all the stations at once? How does a radio receiver's circuitry select which station or frequency to amplify and play for the user?

A classic method for tuning to one received frequency is the use of an RLC circuit, similar to that of Figure 4.4. Think of $V(t)$ in that diagram as the entirety of the signal received from the antenna, all frequencies and stations mixed together (although this circuit is not precisely how the components would be arranged in an actual radio). The goal here is not to provide a circuit schematic, but merely to illustrate how an RLC circuit can be used to sort one frequency out of the cacophony of the airwaves.

Let's say the voltage source $V(t)$ in the circuit of Figure 4.4 inputs a signal to the system that may contain a superposition of many frequencies in the range of 530 kilohertz to 1700 kilohertz. We wish to tune in to 910 AM, that is, to a signal being carried at 9.1×10^5 hertz, corresponding to radial frequency $\omega_{res} = (2\pi)(9.1 \times 10^5) \approx 5.718 \times 10^6$. This can be accomplished by adjusting the values of the capacitor, resistor, and inductor so that the circuit resonates at this frequency. The output of the circuit will be the voltage across the resistor R , which will be routed to other circuitry in the radio for further processing, for example, amplification.

Modeling Exercise 6.2.1 Suppose the inductor has an inductance of $L = 3.5 \times 10^{-4}$ henries ($350 \mu\text{H}$) and the resistor has a value of $R = 10$ ohms. Find a capacitance so that the resonant frequency of the circuit in Figure 4.4 is 910 kilohertz (5.713×10^6 radians per second). The result of Exercise 4.4.3 may be helpful.

Modeling Exercise 6.2.2 In many applications practitioners use $1/\sqrt{LC}$ for the resonant frequency of an RLC circuit (ignoring R). Would that make much difference here? To find out, take $L = 3.5 \times 10^{-4}$ henries and $C = 1.0 \times 10^{-10}$ farad, with each of $R = 1, 10$, and 100 ohms, and compare the resonant frequency of the system as given by the formula in Exercise 4.4.3 to the quantity $1/\sqrt{LC}$. You may find it helpful to recall the computation that led up to (4.72).

Perform a quadratic Taylor expansion on $\frac{\sqrt{4L/C - 2R^2}}{2L}$ (the resonant frequency of an RLC circuit as given in Exercise 4.4.3) with respect to R at $R = 0$ to show that

$$\frac{\sqrt{4L/C - 2R^2}}{2L} \approx \frac{1}{\sqrt{LC}} - \frac{\sqrt{C}}{4L^{3/2}}R^2.$$

How does this justify the use of $1/\sqrt{LC}$ as an approximation to the resonant frequency when R is small?

Modeling Exercise 6.2.3 With the values of L and R specified in Modeling Exercise 6.2.1, and the value of C that you found in that Modeling Exercise, write out the ODE that governs $q(t)$, the charge on the capacitor, for this circuit. Use $V(t)$ (unspecified) for the voltage source.

Modeling Exercise 6.2.4 Find a general solution $q_h(t)$ to the unforced homogeneous ODE you found in Modeling Exercise 6.2.3, in a real-valued form as per (4.33); this captures the transient response in the forced system. Given the rule of thumb that the function $e^{-\alpha t}$ effectively decays to zero at time $t = 5/\alpha$, show that the transients in this system decay to zero in about 3.5×10^{-4} seconds.

Modeling Exercise 6.2.5 Compute the gain function $G(\omega)$ for this circuit using the values of R , L , and C from Modeling Exercise 6.2.3, where here the gain quantifies the amplitude of the periodic response $q_p(t)$. Plot $G(\omega)$ on the range $0 \leq \omega \leq 10^7$, and make sure the resonant frequency is at 910 kilohertz (5.713×10^6 radians per second).

Modeling Exercise 6.2.6 The periodic current in the circuit, $I_p(t)$, is $I_p(t) = q'_p(t)$. Compute the amplitude of $I_p(t)$ when $V(t) = \sin(\omega t)$; this amplitude should depend on ω . Use this to compute the periodic voltage $V_p(t) = RI_p(t)$ across the resistor (with $R = 10$ ohms). As mentioned above, this voltage is the output of the RLC circuit and routed to other circuitry for further processing. Let $H(\omega)$ denote the amplitude of this sinusoidal voltage. Compute $H(\omega)$ as a function of ω , and plot $H(\omega)$ on the range $0 \leq \omega \leq 10^7$.

Modeling Exercise 6.2.7 Compute $H(\omega_{res})$ where $\omega_{res} = 5.718 \times 10^6$; it should be close to 1 volt. That is, the voltage across the resistor is of the same amplitude as $V(t) = \sin(\omega_{res}t)$, as if the inductor and capacitor were not present in the circuit. Then compute $H(\omega_{res} \pm (2\pi)(1.0 \times 10^4))$ (the resonant frequency of the circuit plus or minus 10 kilohertz). What is the amplitude of the voltage across R for a signal at these frequencies? Do you see how this RLC circuit effectively tunes out frequencies that are very far from the resonant frequency?

Modeling Exercise 6.2.8 One way to sharpen the response of this RLC filter circuit (so it screens out unwanted frequencies even more effectively) is to decrease R . Redo Modeling Exercises 6.2.1 to 6.2.7 with $R = 1$ ohm, and compare the new value of $H(\omega_{res} \pm (2\pi)(1.0 \times 10^4))$ to that from Modeling Exercise 6.2.7. Comment: does the smaller value of R more effectively screen out signals at unwanted frequencies?

4.6.3 Project: Parameter Estimation with Second-Order ODEs

This modeling project is based on the SIMIODE Modeling Scenario “Models Motivating Second-Order,” [128].

One of the fundamental tasks of mathematical modeling is validation, where the predictions of a model are compared to the real world. The goal of validation is to determine whether the model is accurate enough to be used for whatever purpose we have in mind. This may go beyond obtaining agreement with the specific data at hand, and also involve validating the more general principles upon which the model is based. For example, in the spring-mass-dashpot model (4.3), is Hooke’s law a reasonable description for the force exerted by a stretched spring? Is viscous damping in the form (4.2) realistic?

One of this project’s authors collected data on May 9, 2013 using the spring-mass system depicted in Figure 4.31, in which a mass was suspended on a spring attached to a rod on a stand. Data on the vertical position of the mass was collected using a Vernier “Go! Motion” detector apparatus shown on the floor below the hanging mass. The distance from the detector to the base of the mass was recorded in a file on a PC. The mass, m , was measured to be 0.200 kg. Through the interface between the motion detector and a PC, data on the position of the mass was sampled for approximately 30 seconds at 50 data points per second for a total number of 1,500 observations. An initial portion of irregular data was excised (when the mass had not yet been released). The resulting data set contains 1461 data points, starting at time $t = 0$, in 0.02 second increments, up to final time $t = 29.18$ seconds. The time $t = 0$ corresponds approximately to a point at which the mass was at maximum positive displacement. The data is in the Excel spreadsheet `spring_mass_data_clean.xls` on the book website [8].

Our goal is to estimate the damping parameter c and spring constant k from this data, and determine whether the model (4.3) is reasonable. To facilitate computations, Maple, Mathematica, Matlab, and Sage worksheets are posted on the book website [8].

Modeling Exercise 6.3.1 As mentioned above, the data consists of the distance from the detector

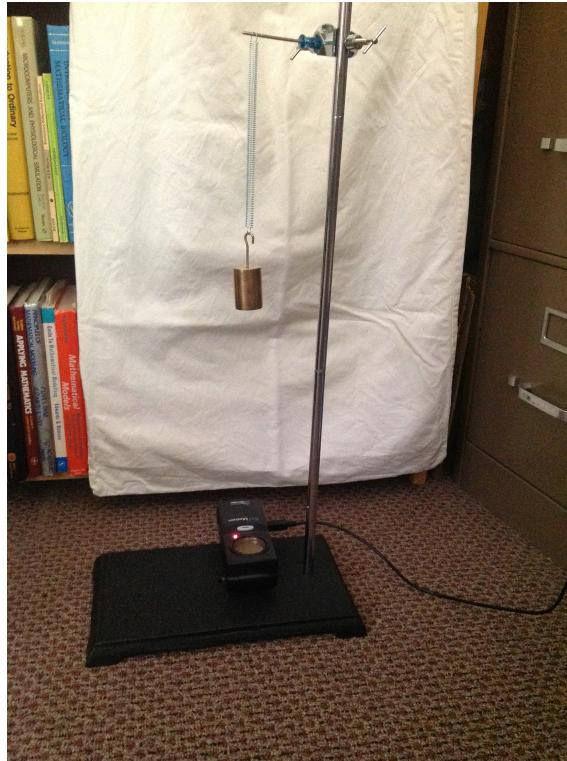


Figure 4.31: The apparatus for collecting data on the bouncing spring-mass system with the Vernier “Go! Motion” detector on the floor. The spring was pulled down and released and the detector collected data on the changing distance between the mass and the collection head of the detector.

head to the base of the mass hanging from the spring. The mean distance is not zero. You should begin by plotting the data.

The mean distance u^* can be estimated by taking the arithmetic mean of the data over the many oscillations the spring makes during the 29.18 seconds. Compute this mean value and then form a centered data set by subtracting u^* from each position measurement. Plot the centered data to make sure it looks reasonable.

The centered data embodies the excursions of the mass about its equilibrium position, and eliminates the effect of gravity. That is, our system may be considered as governed by the homogeneous spring-mass equation (4.4). We will use $y(t)$ to denote this displacement from equilibrium, so $y(t) = u(t) - u^*$.

The function $y(t)$ should satisfy the ODE $my''(t) + cy'(t) + ky(t) = 0$, with $y(0) = y_0$ for some y_0 . The data is edited so that at $t = 0$ the mass is at a point of maximum excursion from equilibrium, paused, and about to return toward equilibrium. As such, an initial velocity of $y'(0) = 0$ is appropriate.

One way to estimate c and k is from the period of the spring-mass oscillations and the decay rate of the oscillation amplitude. To approach this note that the solution to (4.4) is (4.33) for some choice of d_1 and d_2 . We reproduce that equation here for convenience,

$$y(t) = d_1 e^{-\alpha t} \cos(\omega t) + d_2 e^{-\alpha t} \sin(\omega t), \quad (4.118)$$

where we have from (4.29) that $\alpha = \frac{c}{2m}$ and $\omega = \frac{\sqrt{4mk-c^2}}{2m}$. The value of m is known.

Modeling Exercise 6.3.2 What is the initial position $y(0) = y_0$ of the mass in the centered data set? What choice for d_1 in (4.118) does this y_0 value dictate?

Given that the data is edited to start with $y'(0) \approx 0$, the initial data looks very much like a cosine function, so let us take $d_2 = 0$ (for now).

Modeling Exercise 6.3.3 Count the number of oscillations that the spring goes through over the course of the data set, and use it to estimate the period P^* of this damped spring-mass system. What is the corresponding value for ω ?

Modeling Exercise 6.3.4 The value of α in (4.118) dictates the rate of decay of the oscillations. Find a good choice for α . Hint: take a guess at α , then plot $y(t)$ in (4.118) with $d_2 = 0$ and the values you obtained for d_1 and ω ; then adjust α .

Modeling Exercise 6.3.5 Given that

$$\alpha = \frac{c}{2m} \quad \text{and} \quad \omega = \frac{\sqrt{4mk - c^2}}{2m},$$

and that you know $m = 0.2$ kg, come up with estimates for c and k .

The spring constant was determined experimentally to be $k = 17.306$, by plotting force versus displacement data for a variety of different masses and using Hooke's law to fit a linear relationship $F = k \cdot x$, where x is the displacement of the mass and F is the force in newtons necessary to obtain that displacement. How does your estimate of k compare?

Modeling Exercise 6.3.6 Suppose that m itself had not been measured. Would it be possible to estimate all three parameters, m, c , and k , from the data? If so, how? If not, why not?

In the project "Frequency Analysis of Signals" in Section 8.5.4 we develop far more efficient techniques for estimating the period and frequency of oscillatory signals like that of this spring-mass system.

4.6.4 Project: Bike Shock Absorber

Reread Examples 4.1 and 4.20 in which we modeled and did some analysis of a mountain bike front shock absorber. You may also find it helpful to do Reading Exercise 4.3.3 and Exercise 4.3.5, if you haven't already.

The goal in this project is to design a front shock absorber that, under the conditions of the examples above:

- Has a spring that is as compliant (least stiff) as possible, but yields a shock displacement of no more than 140 mm when the rider rides off a 1.5 meter drop.
- Is not excessively overdamped (which makes riding on rugged terrain feel harsh) or under-damped (which makes the bike feel too bouncy.)

Modeling Exercise 6.4.1 Suppose that the ODE that governs the shock displacement $y(t)$ is (as modeled in Example 4.1)

$$my''(t) + cy'(t) + ky(t) = -mg, \tag{4.119}$$

where $m = 46$ kg and $g = 9.8$. However, let us leave c and k undefined for the moment. These are the parameters in which we are interested. To simplify matters, let's start with a shock that is critically damped.

What choice c^* for c yields a critically damped system for the homogeneous version of (4.119)? It should depend on k . Use (4.39) to write out a general solution to the homogeneous ODE $my'' + c^*y'(t) + ky(t) = 0$.

Modeling Exercise 6.4.2 Use the method of undetermined coefficients to find a particular solution $y_p(t)$ to (4.119); this solution will depend on k . Then write out a general solution to (4.119). The general solution should also depend on k .

Modeling Exercise 6.4.3 As noted in Example 4.20, a rider who rides off a 1.5 meter drop will hit the ground at about 5.42 meters per second. Solve (4.119) with initial conditions $y(0) = 0$, $y'(0) = -5.42$, to find the displacement $y(t)$ of the shock. This is a function of t that also involves the indeterminate k .

Modeling Exercise 6.4.4 Determine, either graphically or analytically, the smallest value $k = k^*$ for k that results in the shock compressing no more than -0.14 meters. What is the corresponding value for c^* ?

Modeling Exercise 6.4.5 With the values for k^* and c^* found in Modeling Exercise 6.4.4, plot the solution $y(t)$ on the interval $0 \leq t \leq 1$. Also plot $y''(t)$ on this same time interval. What is the largest acceleration to which the rider is subjected? (It may seem large, but we haven't accounted for the shock absorption provided by the tires or the fact that the rider's legs may also act as shock absorbers.)

Modeling Exercise 6.4.6 Experiment. Can you find other values for c and k that subject the rider to less acceleration while not bottoming out the shock in a 1.5 meter drop?

4.6.5 Project: The Pendulum

In this project we derive the equation of motion for a pendulum using conservation of energy. In the project “The Pendulum 2” of Section 4.6.6 you can explore an alternate derivation based on Newton’s second law of motion; that derivation also incorporates friction, while this one does not.

Consider a pendulum of length L as depicted in Figure 4.32. As the pendulum swings back and forth it makes an angle $\theta(t)$ with respect to vertical at time t . In this exercise we will derive two nonlinear differential equations that govern the pendulum’s motion, one first-order, the other second-order, using a simple conservation of energy argument. We will then approximate the resulting nonlinear ODEs with a simpler linear second-order ODE, and compare the solutions.

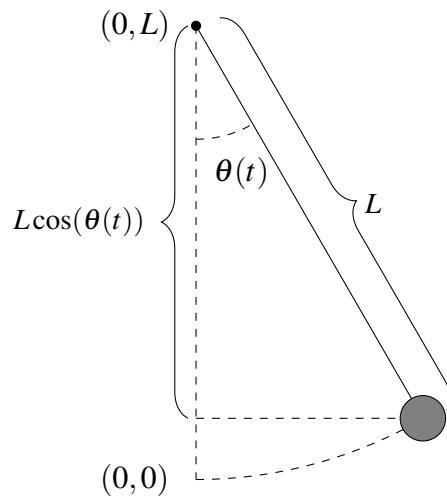


Figure 4.32: A pendulum of length L at an angle of $\theta(t)$ with respect to the vertical.

Let us take the position of the pendulum’s pivot as the point $(0, L)$ in a standard xy coordinate system. Assume that the bob (the mass at the end of the pendulum) has mass m , and that the thin rod that connects the bob to the pivot has negligible mass. The xy position of the bob at any time is

easily seen to be

$$x(t) = L \sin(\theta(t)) \quad \text{and} \quad y(t) = L - L \cos(\theta(t)). \quad (4.120)$$

If the pendulum swings without friction then its total energy, kinetic plus potential, should remain constant in time. This can be used to derive an ODE that governs the pendulum's motion.

Derivation of the Equations of Motion

Modeling Exercise 6.5.1 When the pendulum is hanging straight down ($\theta = 0$) the bob is at position $x = 0, y = 0$. Let us denote this as the position of zero potential energy. Show that if the pendulum is at an arbitrary angle $\theta(t)$ then the system has potential energy

$$U(t) = mgL(1 - \cos(\theta(t))),$$

assuming we take $g > 0$. Hint: the potential energy is the work required to lift the bob against the force of gravity.

Modeling Exercise 6.5.2 Use the equations in (4.120) (which specify the position of the bob) to show that the speed v at which the bob is moving is

$$v(t) = L|\theta'(t)|.$$

Use this to show that the kinetic energy of the pendulum at any time is

$$K(t) = \frac{1}{2}mL^2(\theta'(t))^2.$$

Modeling Exercise 6.5.3 Suppose the pendulum is pulled to an initial angle $\theta(0) = \theta_0$ and released with angular velocity $\theta'(0) = 0$. According to Modeling Exercises 6.5.1 and 6.5.2, at this moment the pendulum has potential energy $U_0 = mgL(1 - \cos(\theta_0))$ and kinetic energy $K_0 = 0$, so total energy U_0 . Energy is conserved (no friction, and gravity is a conservative force) so the total energy of the pendulum is constant in time. That is

$$U + K = U_0 \quad (4.121)$$

where U and K from Modeling Exercises 6.5.1 and 6.5.2 are functions of time. Use (4.121) to show that $\theta(t)$ obeys the first-order differential condition

$$(\theta'(t))^2 = \frac{2g(\cos(\theta(t)) - \cos(\theta_0))}{L} \quad (4.122)$$

with initial condition $\theta(0) = \theta_0$.

Modeling Exercise 6.5.4 We can solve (4.122) for $\theta'(t) = \pm\sqrt{2g(\cos(\theta(t)) - \cos(\theta_0))/L}$, using the plus sign in front of the square root for when the pendulum is swinging counterclockwise ($\theta'(t) > 0$) and the minus sign when it is swinging clockwise ($\theta'(t) < 0$). We obtain separable ODEs for these two phases of the pendulum's motion. However, the resulting ODEs have no simple analytical solution.

Instead, let's do this: Differentiate both sides of (4.122) with respect to t and assume that $\theta'(t)$ is not zero (except when the pendulum is at the extreme limits of its swing) to show that

$$\theta''(t) + \frac{g}{L} \sin(\theta(t)) = 0. \quad (4.123)$$

With $\theta(0) = \theta_0$ and $\theta'(0) = 0$ this is equivalent to (4.122) in that the same function $\theta(t)$ satisfies both.

A Linearized Approximation

Modeling Exercise 6.5.5 Unfortunately, (4.123) is nonlinear and has no simple analytical solution either. However, in many cases a useful approximation can be made that renders the ODE analytically solvable. If the pendulum has a small amplitude of motion then $|\theta(t)|$ remains close to zero for all t ; this will be the case if θ_0 is sufficiently close to zero. Use this assumption to show that equation (4.123) may be approximated with the simpler linearized second-order differential equation

$$\theta''(t) + \frac{g}{L} \theta(t) = 0. \quad (4.124)$$

Hint: look at the Taylor's series for $\sin(\theta)$ about the point $\theta = 0$.

Modeling Exercise 6.5.6 Solve (4.123) with parameters $g = 9.8$, $L = 1$, and initial conditions $\theta(0) = 0.1$ and $\theta'(0) = 0$; you'll need to use a numerical ODE solver. Plot the solution on the interval $0 \leq t \leq 5$. Repeat with equation (4.124) and compare the plots. Is the approximation accurate?

Modeling Exercise 6.5.7 Repeat Modeling Exercise 6.5.6 with initial conditions $y(0) = 1.5$ and $y'(0) = 0$. How does a larger initial angle affect the accuracy of the linearized approximation (4.124)?

4.6.6 Project: The Pendulum 2

In this project we consider a derivation of the equation of motion of a pendulum based on Newton's second law of motion, in contrast to the derivation of Section 4.6.5, which was based on arguments involving conservation of energy.

Consider a pendulum of length L , as depicted in Figure 4.32. As the pendulum swings back and forth it makes an angle $\theta(t)$ with respect to vertical at time t . In this exercise we will derive the differential equation satisfied by $\theta(t)$ using Newton's second law of motion, $\mathbf{F} = m\mathbf{a}$. We will then approximate the resulting nonlinear ODE with a simpler linear ODE, and compare the solutions.

Position, Velocity, and Acceleration of the Bob

We will take the position of the pendulum's pivot as the origin in a standard xy coordinate system. Assume that the bob (the mass at the end of the pendulum) has mass m , and that the thin rod that connects the bob to the pivot has negligible mass. A little geometry shows that the xy position (as a displacement vector $\mathbf{r}(t)$ from the origin) of the pendulum at any time is

$$\mathbf{r}(t) = L \langle \sin(\theta(t)), -\cos(\theta(t)) \rangle = L\mathbf{u}(t),$$

where $\mathbf{u}(t) = \langle \sin(\theta(t)), -\cos(\theta(t)) \rangle$, a unit vector that points radially outward from the origin toward the bob; note that $\mathbf{u}(t)$ is orthogonal to the bob's circular path.

Modeling Exercise 6.6.1 Differentiate $\mathbf{r}(t)$ with respect to t and show that the velocity of the pendulum is

$$\mathbf{v}(t) = L\theta'(t)\mathbf{u}^\perp(t) \quad (4.125)$$

where $\mathbf{u}^\perp(t) = \langle \cos(\theta(t)), \sin(\theta(t)) \rangle$. Use the dot product to verify that $\mathbf{u}(t)$ and $\mathbf{u}^\perp(t)$ are orthogonal at all times. As a result, $\mathbf{r}(t)$ and $\mathbf{v}(t)$ are also orthogonal.

Modeling Exercise 6.6.2 Differentiate $\mathbf{v}(t)$ in (4.125) with respect to t and show that the acceleration $\mathbf{a}(t)$ of the pendulum is

$$\mathbf{a}(t) = L\theta''(t)\mathbf{u}^\perp(t) - L(\theta'(t))^2\mathbf{u}(t). \quad (4.126)$$

The expression for $\mathbf{a}(t)$ above is what we will use in Newton's second law.

Forces Acting on the Bob

Now let us consider the forces acting on the bob, illustrated in a free-body diagram in Figure 4.33. These forces are the gravitational force \mathbf{F}_g and the force \mathbf{F}_T exerted by the rod connecting the bob to

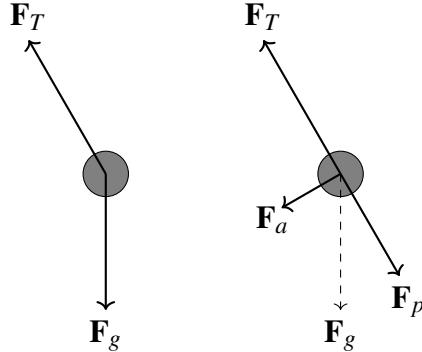


Figure 4.33: Left panel: free body diagram of pendulum bob with gravitational force \mathbf{F}_g and rod force \mathbf{F}_T . Right panel: same free-body diagram but with force \mathbf{F}_g decomposed into components \mathbf{F}_p orthogonal to path of motion, and \mathbf{F}_a , parallel to path of motion.

the pivot point, as illustrated in the left panel of Figure 4.33. The force \mathbf{F}_T exerted by the rod is parallel to the vector $\mathbf{r}(t)$ above, directed toward the origin, and so of the form $T\mathbf{u}(t)$ for some scalar T (the tension in the rod). The force of gravity on the bob is given by the vector $\mathbf{F}_g = \langle 0, -mg \rangle$, where we take $g > 0$, say $g = 9.8$ meters per second squared. This force has magnitude mg and can be expressed as a sum of two orthogonal components,

$$\mathbf{F}_g = \mathbf{F}_p + \mathbf{F}_a, \quad (4.127)$$

where \mathbf{F}_p is parallel to the rod (orthogonal to the motion of the bob) and \mathbf{F}_a is parallel to the motion of the bob (orthogonal to the rod), as illustrated in the right panel of Figure 4.33. These forces are

$$\begin{aligned} \mathbf{F}_p &= mg \cos(\theta(t)) \mathbf{u}(t) \\ \mathbf{F}_a &= -mg \sin(\theta(t)) \mathbf{u}^\perp(t). \end{aligned} \quad (4.128)$$

Modeling Exercise 6.6.3 Verify that (4.127) holds if \mathbf{F}_a and \mathbf{F}_p are as in (4.128).

Based on (4.127) and (4.128), the total force acting on the bob is

$$\mathbf{F} = -mg \sin(\theta(t)) \mathbf{u}^\perp(t) + (mg \cos(\theta(t)) - T) \mathbf{u}(t), \quad (4.129)$$

where T denotes the tension in the rod, which is the magnitude of the force that the rod exerts on the bob.

Modeling Exercise 6.6.4 Use Newton's second law of motion, $\mathbf{F} = m\mathbf{a}$, in conjunction with (4.129) and (4.126), to conclude that

$$\theta''(t) + \frac{g}{L} \sin(\theta(t)) = 0. \quad (4.130)$$

Analysis and Approximation of the Pendulum Equation

The second-order ODE (4.130) is nonlinear due to the $\sin(\theta(t))$ term. It has no simple analytical solution, but a simplifying approximation can be made, under certain conditions.

Modeling Exercise 6.6.5 If the pendulum has a small amplitude of motion then $|\theta(t)|$ remains close to zero for all t . Use this to show that equation (4.130) may be approximated with the simpler linearized second-order differential equation

$$\theta''(t) + \frac{g}{L}\theta(t) = 0. \quad (4.131)$$

Hint: look at the Taylor's series for $\sin(\theta)$ about the point $\theta = 0$.

Modeling Exercise 6.6.6 Solve (4.130) with parameters $g = 9.8$ and $L = 1$, with initial conditions $\theta(0) = 0.1$ and $\theta'(0) = 0$; you'll need to use a numerical ODE solver. Plot the solution for $0 \leq t \leq 5$. Repeat with equation (4.131) and compare plots. Is the approximation accurate?

Modeling Exercise 6.6.7 Repeat the Modeling Exercise 6.6.6 with initial conditions $\theta(0) = 1.5$, $\theta'(0) = 0$. Comment.

Adding Friction

Suppose that as the pendulum swings it experiences a frictional force with direction opposed to the velocity of the bob, and in proportion to the speed of the bob. Using (4.125), we find that this force is of the form

$$\mathbf{F}_{fric} = -c\mathbf{v} = cL\theta'(t)\mathbf{u}^\perp(t)$$

for some positive constant c . The larger the value of c , the greater the frictional force on the pendulum.

Modeling Exercise 6.6.8 Argue that (4.130) then becomes

$$\theta''(t) + c\theta'(t) + \frac{g}{L}\sin(\theta(t)) = 0. \quad (4.132)$$

This is the equation of the **damped pendulum**.

Modeling Exercise 6.6.9 Repeat Modeling Exercises 6.6.5 to 6.6.7 with (4.132) in place of (4.131), with the choice $c = 0.5$. Compare the behavior of the solution to the nonlinear ODE (4.130) with that of the linearized ODE (4.131).

5. The Laplace Transform

5.1 Discontinuous Forcing Functions

5.1.1 Motivation: Pharmacokinetics

After a major surgery patients typically have significant pain. One standard drug for providing analgesic relief to these patients is morphine sulfate. This drug is often administered intravenously (IV) as a **bolus**, that is, a discrete dose given over a short interval of time. The drug enters the bloodstream and begins working almost immediately. However, over time the drug is metabolized and excreted, so more must be given later to maintain an appropriate therapeutic amount of the drug in the body. Precisely controlling the amount present in the body can be challenging, and mathematical models have proven useful for understanding how to implement such control. Such models are an important part of the field of **pharmacokinetics**, which is concerned with modeling how the body metabolizes and excretes drugs.

The rate r_{out} at which many types of drugs are eliminated from the body is often modeled as [95, 80]

$$r_{out} = -ku(t), \quad (5.1)$$

where $u(t)$ denotes the amount (usually mass) of drug in the body at time t hours and k is a positive constant with the dimension reciprocal time T^{-1} . There is a more elaborate one-compartment model that underlies (5.1), and we will explore this later, but for now let's accept this expression for r_{out} . The constant k that governs the rate at which the drug is eliminated varies from patient to patient, but typically the amount of morphine present diminishes by a factor of about one-half every four hours. That is, morphine has a **half-life** of four hours in the body [80], and this dictates the value $k \approx 0.173$ (see Exercise 2.1.3). The constant k here has units of reciprocal hours.

An ODE Model

In addition to bolus dosing, morphine can be administered continually at some rate $r(t)$, by way of an intravenous drip or an infusion pump. The amount of drug in the body at any time can then be modeled using the same “instantaneous rate of change equals rate in minus rate out” methodology that was used in Section 1.2 for intracochlear drug delivery. By making use of (5.1) we are led to

the ODE

$$\underbrace{u'(t)}_{\text{rate of change of } u(t)} = \underbrace{r(t)}_{\text{rate in}} - \underbrace{ku(t)}_{\text{rate out}}, \quad (5.2)$$

in which $u(t)$ is the amount of morphine (mg) in the body and t is time in hours.

Consider (5.2) in the case that $r(t) = 0$, and suppose the patient is given a 10 mg bolus of morphine after surgery at time $t = 0$. The solution to (5.2) in this case is $u(t) = 10e^{-kt}$. Over the next four hours the amount of drug in the patient's system will fall to 5 mg, a level that may be too low to provide adequate pain relief. An additional dose can be given at time $t = 4$ and periodically thereafter, but another common approach is to administer an initial bolus at time $t = 0$ to immediately raise the blood concentration up to a therapeutic value and then start additional medication using an IV infusion pump, to maintain an adequate blood concentration. The pump can be programmed to administer morphine at a rate $r(t) = r_0$ for $t > 0$ for some constant r_0 , so that over time the amount of morphine in the body stabilizes at an appropriate level.

Reading Exercise 5.1.1 With $r(t) = r_0 > 0$ and $k > 0$ as some unspecified rate constant, sketch a phase line portrait for (5.2). Of course, you can confine your attention to $u \geq 0$. Show that all solutions asymptotically approach the fixed point $u = r_0/k$.

■ **Example 5.1** Suppose a 10 mg initial bolus is given at time $t = 0$ and then morphine is administered continuously using an infusion pump at a rate of 1.5 mg per hour for $t > 0$. The ODE (5.2) becomes

$$u'(t) = -ku(t) + 1.5$$

with $u(0) = 10$. The solution is $u(t) \approx 8.67 + 1.33e^{-kt}$ (recall that $k \approx 0.173$). This solution is plotted in the left panel of Figure 5.1. The amount of morphine in the body stabilizes at $1.5/k = 8.67$ mg, which may provide a sufficient concentration for the desired pain relief. ■

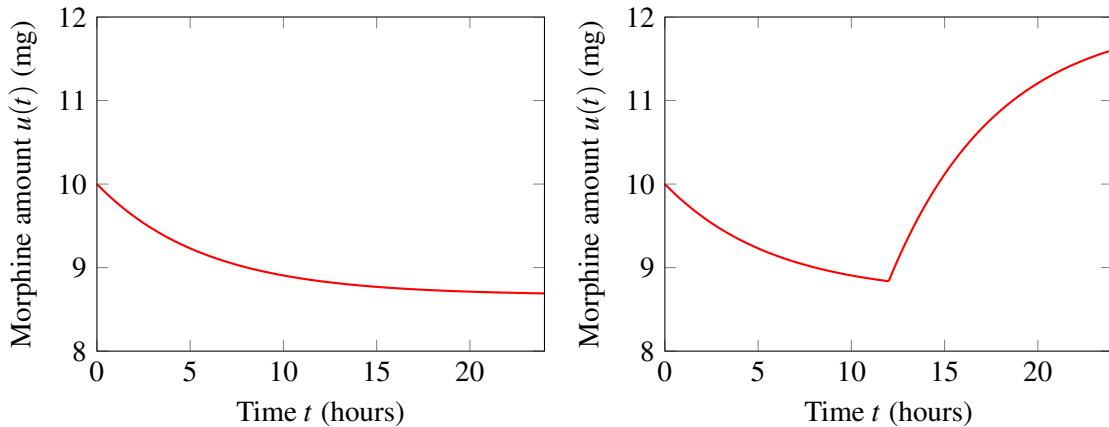


Figure 5.1: Left panel: solution to (5.2) with $u(0) = 10$ mg and $r(t) = 1.5$ mg per hour. Right panel: solution to (5.2) with $u(0) = 10$ mg and $r(t) = 1.5$ mg per hour for $t \leq 12$ and $r(t) = 2.08$ mg per hour for $t > 12$.

5.1.2 Complication: Discontinuous Forcing

What if in Example 5.1 the patient is still in pain at time $t = 12$ hours? Suppose we want to increase the amount of morphine present in the body to 12 mg and keep it there for some period of time. One

way to do this would be to increase the rate at which the infusion pump administers the medication, to a rate of $r(t) = 12k \approx 2.08$ mg per hour, starting at time $t = 12$. That is, solve (5.2) but with

$$r(t) = \begin{cases} 1.5, & 0 \leq t \leq 12 \\ 2.08, & t > 12. \end{cases} \quad (5.3)$$

The resulting amount of morphine in the body is shown in the right panel of Figure 5.1. Note the abrupt change at $t = 12$ when we begin infusing morphine at a greater rate. Solving (5.2) with the discontinuous function $r(t)$ in (5.3) isn't something we've considered so far in this text. How did we arrive at the solution shown on the right in Figure 5.1?

One approach is this: solve (5.2) with $r(t) = 1.5$ and $u(0) = 10$; this is an easy separation of variables or integrating factor computation and yields $u_1(t) \approx 8.67 + 1.33e^{-kt}$, as noted above, where we are using u_1 for the amount of morphine in the body for $t \leq 12$. At $t = 12$ the ODE (5.2) suddenly transitions to $u'(t) = -ku(t) + 2.08$, and this ODE must be solved anew for $t > 12$. Let's use $u_2(t)$ to denote the amount of morphine present when $t > 12$. The condition that ties together the solution on either side of the transition is $u_2(12) = u_1(12) \approx 8.84$. We then solve $u'_2 = -ku_2 + 2.08$ with initial condition $u_2(12) = 8.84$. The result is $u_2(t) \approx 12.0 - 3.16e^{-k(t-12)}$. The amount of morphine present is then given by $u_1(t)$ for $t \leq 12$ and $u_2(t)$ for $t > 12$. This is the function graphed in the right panel of Figure 5.1, and is given explicitly by

$$u(t) = \begin{cases} 8.67 + 1.33e^{-kt}, & 0 \leq t \leq 12 \\ 12.0 - 3.16e^{-k(t-12)}, & t > 12. \end{cases} \quad (5.4)$$

The introduction of the discontinuous change in $r(t)$ makes the solution process somewhat more tedious than the case in which $r(t)$ is given by a single simple formula, e.g., a constant. And if we make another change to $r(t)$ at a later time, we have to go through this piecemeal solution process again.

Reading Exercise 5.1.2 Suppose the amount of morphine in the patient's body is governed by (5.2) with

$$r(t) = \begin{cases} 1.5, & 0 \leq t \leq 12 \\ 2.5, & t > 12. \end{cases}$$

Find the amount of morphine $u(t)$ in the patient's body, and plot $u(t)$ for $0 \leq t \leq 24$ hours. Hint: for $t < 12$ the solution is the same as in (5.4).

5.1.3 Complication: Impulsive Forcing

Here's another twist: an examination of the right panel in Figure 5.1 shows that even with the increased administration rate, the amount of morphine rises rather slowly, taking many hours to get close to the desired 12 mg level. The patient needs pain relief now! The obvious solution is to administer an IV bolus of, say, 5 mg at time $t = 12$, then step up the infusion rate to $r(t) = 2.08$ mg per hour. How can we account for a 5 mg bolus in the framework of the ODE (5.2)? An instantaneous bolus of 5 mg at time $t = 12$ would require the dosage rate function $r(t)$ to skyrocket to an infinite value for an infinitesimal amount of time, in a way that corresponds to a 5 mg dose given in this tiny time interval. Such phenomena are called **impulsive** and in this case might be approximated as, say, $r(t) = 500$ mg per hour for $t = 11.995$ to $t = 12.005$, or $r(t) = 5000$ mg per hour for $t = 11.9995$ to $t = 12.0005$, or something similar. More generally, we can model the 5 mg bolus being delivered as

$$r(t) = \frac{5}{2\epsilon} \quad (5.5)$$

mg per hour from time $t = 12 - \varepsilon$ to $t = 12 + \varepsilon$, where $\varepsilon > 0$ is arbitrarily close to zero. This high infusion rate over a very short time would correspond to a total dose of

$$\frac{5}{2\varepsilon} \frac{\text{mg}}{\text{hour}} \times 2\varepsilon \text{ hours} = 5 \text{ mg.}$$

It would be nice if there were some way we could avoid getting involved with the value of ε , since it really shouldn't matter as long it's close to zero.

Reading Exercise 5.1.3 Suppose a patient is administered 10 mg of morphine at time $t = 0$ and then infused at a rate of 1.5 mg per hour for $t > 0$. At time $t = 12$ the patient is also administered a 5 mg bolus instantaneously. Argue that the amount of morphine in the patient's system for $t < 12$ is still $u_1(t) = 8.67 + 1.33e^{-kt}$ and that for $t > 12$ the amount $u_2(t)$ should still satisfy $u'_2(t) = -ku_2(t) + 1.5$ but with initial data $u_2(12) = u_1(12) + 5$. Then find $u_2(t)$ and plot the amount of morphine in the patient's system on the interval $0 \leq t \leq 24$.

5.1.4 Discontinuous Forcing and Transform Methods

Many physical situations like the morphine administration problem are most easily modeled by ODEs with discontinuous or impulsive forcing functions and coefficients. For example, interest rates for bank accounts may vary abruptly, and deposits result in instantaneous increases in an account's balance. Spring-mass systems may be subjected to forces that change suddenly, or, in the case of a hammer blow, can be modeled as almost infinite in magnitude for a very short period of time. A switch in a circuit may be turned from off to on or vice versa almost instantaneously. The world is full of these kinds of abrupt changes, but the traditional language of calculus involving continuous and differentiable functions isn't quite up to the task of describing this kind of change. We need to extend this language in a way that allows us to model discontinuous and impulsive phenomena.

There is a mathematical toolbox that contains a versatile set of tools for describing exactly these type of phenomena. Moreover, this toolbox provides a powerful method—the **Laplace transform**—for analyzing the resulting ODEs so obtained. The Laplace transform also has uses beyond the analysis of ODEs; it plays an important role in control theory in which one wishes to steer a physical system to a desired goal, for example, control the rate function $r(t)$ in (5.2) to obtain a desired morphine amount $u(t)$ in the body at any time. These topics are the focus of this chapter.

5.1.5 Exercises

Exercise 5.1.1 A patient is given a 5 mg bolus of morphine at time $t = 0$, followed by infusion at a rate of $r(t) = 1$ mg of morphine per hour.

- (a) Find the amount $u_1(t)$ (mg) of morphine present in the patient's body on the time interval $0 \leq t \leq 12$ hours, by solving the relevant ODE $u'_1(t) = -ku_1(t) + 1$ with $u_1(0) = 5$. Recall $k \approx 0.173$ (reciprocal hours).
- (b) From $t = 12$ to $t = 18$ hours the infusion rate is increased to $r(t) = 1.5$ mg per hour. Find the amount $u_2(t)$ of morphine in the patient's body in this time interval by solving $u'_2(t) = -ku_2(t) + 1.5$ with $u_2(12) = u_1(12)$.
- (c) At time $t = 18$ a 5 mg bolus is administered; the infusion rate is then decreased to $r(t) = 1$ mg per hour. Find the amount $u_3(t)$ of morphine present in the patient's body for $t > 18$ hours. Hint: this 5 mg bolus can be accounted for as $u_3(18) = u_2(18) + 5$.
- (d) Plot the amount of morphine in the patient's system up to time $t = 24$: $u_1(t)$ for $0 < t < 12$, $u_2(t)$ for $12 < t < 18$, and $u_3(t)$ for $18 < t < 24$. Is the graph consistent with the various

assumptions made in parts (a)-(c)?

Exercise 5.1.2 A bank account is opened with \$1000 at time $t = 0$. The account pays interest at an annual rate of 2 percent, compounded continuously, that is, the account accrues interest at a rate of $0.02p(t)$, so the account balance $p(t)$ is governed by $p'(t) = 0.02p(t)$, where t is time in years.

- Solve $p'(t) = 0.02p(t)$ with $p(0) = 1000$ to find the account balance as a function of t . How much money is in the account at time $t = 5$ years?
- Suppose that in addition to interest, money is deposited into the account on a regular basis, and frequently enough that the deposits may be considered continuous. For example, a weekly deposit of \$10 corresponds to a continuous rate of $r(t) = 520$ dollars per year, roughly. If the rate at which money is deposited is in fact given by $r(t) = 520$, argue that $p(t)$ satisfies $p'(t) = 0.02p(t) + 520$. Solve this ODE with $p(0) = 1000$. How much money is in the account at $t = 5$?
- Suppose the deposit rate is $r(t) = 520$ dollars per year from time $t = 0$ to time $t = 2$, but then drops to $r(t) = 200$ dollars per year for time $2 \leq t \leq 5$. Find the amount of money in the account as a function of time. Hint: the balance will be given by a function $p_1(t)$ for $0 \leq t < 2$ where $p_1(t)$ satisfies $p'_1(t) = 0.02p_1(t) + 520$ and a function $p_2(t)$ for $2 \leq t \leq 5$, where $p'_2(t) = 0.02p_2(t) + 200$. What is the correct initial condition for $p_2(2)$?
- Suppose that the deposit rates of part (c) still hold, but additionally at time $t = 5$ a lump sum deposit of \$1000 is made, and after this no further deposits are made ($r(t) = 0$). The interest rate remains at 2 percent. What is the balance of the account for $t > 5$, as a function of t ?

Exercise 5.1.3 An object in an environment with ambient temperature $A = 80$ degrees obeys Newton's law of cooling (2.14) with cooling constant $k = 0.05$, with time measured in minutes. The object has temperature 120 degrees at time $t = 0$. At time $t = 50$ the object is moved to an environment with ambient temperature $A = 90$ degrees; the object still obeys Newton's law of cooling with the same cooling constant $k = 0.05$. Find the temperature of the object at time $t = 70$.

Exercise 5.1.4 An undamped spring-mass system with mass $m = 2$ kg and spring constant $k = 8$ newtons per meter is at equilibrium position $u = 0$ and is not moving at time $t = 0$. No additional forces act on the mass until time $t = 10$ seconds, but for $t > 10$ a force $f(t) = 40$ newtons is applied to the mass. At the junctions $t = 10$ and $t = 15$ tie the solutions together by requiring the position and velocity of the mass to be continuous.

Exercise 5.1.5 Consider an RC circuit like that shown in Figure 2.2, with resistor $R = 10$ ohms and capacitor $C = 10^{-4}$ F. The capacitor is uncharged at time $t = 0$. Suppose the voltage source is $V(t) = 2$ volts for time $0 \leq t \leq 0.003$ seconds and then switches to $V(t) = 5$ volts for $t > 0.003$. Find the charge on the capacitor at time $t = 0.005$ seconds.

5.2 The Laplace Transform

In this section we define the Laplace transform, compute the transforms of some standard elementary functions, and then illustrate how the Laplace transform can be used to solve ODEs. We'll put aside modeling and applications in this section to concentrate on the essential mathematics, but we will use the pharmacokinetic model of the last section for inspiration. Our focus will be on homogeneous first- and second-order linear equations. In this section you'll see how the Laplace transform turns ODEs into algebra problems. It may seem that we're merely solving ODEs that we already know how to solve, and for the moment this is correct. But the next sections will illustrate how the Laplace transform facilitates analyzing ODEs with discontinuous or impulsive forcing functions and plays a central role in the subject of **control theory**.

5.2.1 Definition of the Laplace Transform

The Laplace transform is straightforward to define, but it will not be apparent at first how or why anyone came up with the definition, or why it's useful. There are various motivating arguments that can be made for why the definition is natural, but at this point they're probably more distracting than illuminating. So let's just get to it and look at motivations later.

Consider a function $f(t)$ of a real variable t with domain $t \geq 0$. The Laplace transform produces a new function F from f just as, for example, differentiation and antiderivation produce new functions f' and $\int f(t) dt$, respectively, from f . The Laplace transform of $f(t)$ is the function $F(s)$ defined as

$$F(s) = \int_0^\infty e^{-st} f(t) dt. \quad (5.6)$$

Here s will typically be a real number, although it can be complex, which is sometimes useful. It's fairly common, when given a function denoted by a lower case letter, to use the corresponding capital letter for the Laplace transform. Thus the Laplace transform of f is denoted by F , the transform of g is denoted by G , etc. Using s for the independent variable in the Laplace transform is fairly universal, although one could in principle use anything.

An alternate notation for the Laplace transform of a function f is

$$\mathcal{L}(f)(s) = \int_0^\infty e^{-st} f(t) dt. \quad (5.7)$$

Here \mathcal{L} is the Laplace transform operator that processes f into the function $\mathcal{L}(f)$ (which is F in (5.6)), a function of s , according to (5.7). We will frequently omit the independent variable s on the left in (5.7) and just write $\mathcal{L}(f)$ instead of $\mathcal{L}(f)(s)$.

There are several technical requirements f must satisfy in order to make sure it has a Laplace transform, and possibly restrictions on s , but we'll get to those shortly. Let's just start computing some Laplace transforms.

Examples

- **Example 5.2** Let's compute the Laplace transform of the function $f(t) = C$, where C is a constant. Based on (5.6) the Laplace transform $F(s)$ is given by

$$F(s) = \int_0^\infty Ce^{-st} dt. \quad (5.8)$$

This is an improper integral, and we will do it carefully. The value of the integral depends on s , and is what we'll call $F(s)$. Later on we may get a bit more casual with improper integrals, and eventually we won't even use integration to compute Laplace transforms.

Recall from Calculus 2 that an improper integral like the right side of (5.8) is computed as

$$\int_0^\infty Ce^{-st} dt = \lim_{T \rightarrow \infty} \left(\int_0^T Ce^{-st} dt \right), \quad (5.9)$$

assuming the limit on the right exists. The integral on the right in (5.9) is evaluated using the fundamental theorem of calculus, and to do so we need an antiderivative for e^{-st} with respect to t (treating s as a constant). The expression $-e^{-st}/s$, considered as a function of t , is a suitable antiderivative.

Reading Exercise 5.2.1 Verify that $-e^{-st}/s$ is an antiderivative for e^{-st} with respect to t .

We will now assume that $s > 0$, for reasons that will become apparent. With an antiderivative $-e^{-st}/s$ in hand we find that

$$\begin{aligned} \int_0^T Ce^{-st} dt &= C \int_0^T e^{-st} dt \\ &= -C \frac{e^{-st}}{s} \Big|_{t=0}^{t=T} \\ &= -C \frac{e^{-sT}}{s} + C \frac{e^{(0)s}}{s} \\ &= \frac{C}{s} - C \frac{e^{-sT}}{s}. \quad (\text{Use } e^0 = 1 \text{ above.}) \end{aligned} \quad (5.10)$$

Now use the last line from (5.10) in (5.9) to find that

$$\begin{aligned} \int_0^\infty Ce^{-st} dt &= \lim_{T \rightarrow \infty} \left(\frac{C}{s} - C \frac{e^{-sT}}{s} \right) \\ &= \lim_{T \rightarrow \infty} \frac{C}{s} - C \lim_{T \rightarrow \infty} \frac{e^{-sT}}{s} \\ &= \frac{C}{s}. \end{aligned} \quad (5.11)$$

We used the linearity of limits and the fact that $s > 0$, so the exponential $e^{-sT} \rightarrow 0$ as $T \rightarrow \infty$. Notice that C/s in the second equation in (5.11) doesn't even depend on T , so that the limit is just C/s . The final expression in (5.11) in conjunction with (5.8) shows that the Laplace transform of the constant function $f(t) = C$ is the function

$$F(s) = \frac{C}{s}$$

with domain $s > 0$. If $s \leq 0$ then the limit involving e^{-sT} on the right in the second line of (5.11) doesn't exist, since in that case the exponential grows without bound as $T \rightarrow \infty$ and so $F(s)$ is not defined when $s \leq 0$. ■

Let us do one more example with care and then consider some theoretical properties of the Laplace transform, as well as additional technical details.

■ **Example 5.3** Let $f(t) = e^{3t}$. From (5.6), the Laplace transform of this function is

$$F(s) = \int_0^\infty e^{3t} e^{-st} dt. \quad (5.12)$$

Again, this is an improper integral and we will evaluate it carefully, to illustrate an important point.

First note that from the elementary properties of exponentials,

$$e^{3t} e^{-st} = e^{3t-st} = e^{(3-s)t}.$$

The improper integral in (5.12) is computed as

$$\int_0^\infty e^{(3-s)t} dt = \lim_{T \rightarrow \infty} \left(\int_0^T e^{(3-s)t} dt \right), \quad (5.13)$$

defined only when the limit on the right exists. A suitable antiderivative for the integrand on the right in (5.13) is $e^{(3-s)t}/(3-s)$; make sure you believe this. Then

$$\begin{aligned} \int_0^T e^{(3-s)t} dt &= \frac{e^{(3-s)t}}{3-s} \Big|_{t=0}^{t=T} \\ &= \frac{e^{(3-s)T}}{3-s} - \frac{e^{(3-s)(0)}}{3-s} \\ &= \frac{1}{s-3} + \frac{e^{(3-s)T}}{3-s}. \quad (\text{Use } e^0 = 1 \text{ above.}) \end{aligned} \quad (5.14)$$

Use (5.14) in (5.13) to find that

$$\begin{aligned} \int_0^\infty e^{(3-s)t} dt &= \lim_{T \rightarrow \infty} \left(\frac{1}{s-3} + \frac{e^{(3-s)T}}{3-s} \right) \\ &= \lim_{T \rightarrow \infty} \left(\frac{1}{s-3} \right) + \lim_{T \rightarrow \infty} \left(\frac{e^{(3-s)T}}{3-s} \right). \end{aligned} \quad (5.15)$$

Here's where we have to be a little bit careful. The first term on the right in (5.15) is an easy limit with value $1/(s-3)$ ($1/(s-3)$ doesn't even depend on T). In the second term the limit of $e^{(3-s)T}$ as $T \rightarrow \infty$ exists only when $s \geq 3$, so that $3-s \leq 0$. But $s=3$ isn't allowed in that expression due to the division by $3-s$. Moreover, if $s=3$ then it's clear that the integral on the right in (5.13) that would define $F(3)$ doesn't converge. We therefore restrict our attention to $s > 3$, and in this case $e^{(3-s)T}/(3-s)$ has limit 0 as $T \rightarrow \infty$. All in all the right side of (5.15) has the value $1/(s-3)$, if $s > 3$. We conclude from (5.13) that the Laplace transform of e^{3t} is given by

$$F(s) = \frac{1}{s-3}$$

with domain $s > 3$. ■

5.2.2 What Kinds of Functions Can Be Transformed?

With the general approach of Examples 5.2 and 5.3, we're in a position to compute the Laplace transforms of all the elementary functions we know and love—polynomials, exponentials, trigonometric functions, and various combinations of these. It really is just an exercise in basic integration techniques. But first, let's briefly consider what general types of functions $f(t)$ have a Laplace transform, for not every function does. The improper integral that defines the transform must converge, and this means the function $f(t)$ can't grow too fast. The function of interest also can't be too discontinuous. These conditions ensure that the integral that defines the transform actually makes sense, and gives a firm theoretical foundation for the computations we're going to do.

In Example 5.3 the function $f(t) = e^{3t}$ had a Laplace transform $F(s) = 1/(s-3)$, and the relevant improper integral defining the transform converged only for $s > 3$. This suggests that if a function grows rapidly, the Laplace transform integral may fail to converge, and motivates the following definition.

Definition 5.2.1 A function $f(t)$ defined for $t \geq 0$ is said to be of **exponential order** if for some constant a and some constant $M > 0$ the inequality

$$|f(t)| \leq M e^{at} \quad (5.16)$$

holds for all $t \geq 0$.

Of course any exponential function $M e^{at}$ is of exponential order. But, for example, the function $f(t) = e^{t^2}$ is not of exponential order, for as t increases $f(t)$ outgrows the function $M e^{at}$ for any choice of M and a . That is, there is no choice for a and M so that $f(t) \leq M e^{at}$ for all t in this case; see Reading Exercise 5.2.2.

Only functions of exponential order play well with the Laplace transform, for otherwise the improper integral defining the Laplace transform of f won't converge for any choice of s . Happily, functions of exponential order encompass virtually everything encountered in applications.

Reading Exercise 5.2.2 Show that $f(t) = e^{t^2}$ is not of exponential order by showing that for any fixed choice of a and $M > 0$

$$\lim_{t \rightarrow \infty} \frac{e^{t^2}}{M e^{at}} = \infty.$$

Hint: it might be helpful to consider the limit of the logarithm of $e^{t^2}/(M e^{at})$ instead. Explain why this violates (5.16).

A second technical condition exists on the functions $f(t)$ that we wish to Laplace transform. Specifically, f should be piecewise continuous, which we will now define. Recall from Calculus 1 that a function $f(t)$ defined on some interval $a < t < b$ has a **jump discontinuity** at a point t_0 if the left and right handed limits L^- and L^+ defined by

$$L^- = \lim_{t \rightarrow t_0^-} f(t) \quad \text{and} \quad L^+ = \lim_{t \rightarrow t_0^+} f(t) \quad (5.17)$$

both exist but are not equal. If these limits are equal then f is continuous as $t = t_0$. At a jump discontinuity f is bounded because these one-sided limits both exist. Of course f is bounded at any point of continuity as well. We make the following definition.

Definition 5.2.2 A function $f(t)$ defined for $t \geq 0$ is **piecewise continuous** if it has finitely many jump discontinuities in any interval $0 < t \leq T$ but is continuous at all other points (including $t = 0$, in that $\lim_{t \rightarrow 0^+} f(t) = f(0)$).

This definition allows the possibility that f has infinitely many discontinuities in the interval $0 \leq t < \infty$, e.g., at every integer, which is sometimes useful to consider. The condition on the piecewise continuity of $f(t)$ is necessary because the Riemann integral is generally defined only for continuous or piecewise continuous functions.

We summarize our observations in the following theorem. The work above is not a rigorous proof of this theorem, but such a proof can be found in [24].

Theorem 5.2.1 If $f(t)$ is defined for $t \geq 0$, piecewise continuous, and of exponential order with some constant a in (5.16), then the Laplace transform $F(s)$ defined by (5.6) exists for all $s > a$.

5.2.3 Laplace Transforms of Elementary Functions

One of the most important properties of the Laplace transform is **linearity**. Specifically, if f and g are functions with Laplace transforms $F = \mathcal{L}(f)$ and $G = \mathcal{L}(g)$, then

$$\mathcal{L}(f+g)(s) = F(s) + G(s)$$

$$\mathcal{L}(cf)(s) = cF(s).$$

These facts follow from the linearity of integration and limits, specifically, if the various integrals converge then

$$\begin{aligned}\int_0^\infty e^{-st}(f(t) + g(t)) dt &= \int_0^\infty e^{-st} f(t) dt + \int_0^\infty e^{-st} g(t) dt, \\ \int_0^\infty e^{-st}(cf(t)) dt &= c \int_0^\infty e^{-st} f(t) dt.\end{aligned}$$

These two equations are precisely the above assertions concerning linearity.

Some common Laplace transforms are tabulated in Table 5.1, where $f(t)$ denotes the function and $F(s)$ its transform. In each case $F(s)$ is defined for $s > 0$ unless otherwise noted. In Table 5.1 the constants a and b can be positive or negative. The cases $f(t) = \sin(bt)$ and $f(t) = \cos(bt)$ are special cases of the last two lines (when $a = 0$), but these transforms come up often enough that we list them separately. The transform $F(s)$ in each line in Table 5.1 can be verified by evaluating the integral (5.6) that defines $F(s)$, using standard techniques from integral calculus.

Function $f(t)$	Laplace Transform $F(s)$	Comment
C	C/s	
t^n	$n!/s^{n+1}$	n an integer
e^{at}	$1/(s-a)$	$s > a$
$t^n e^{at}$	$n!/(s-a)^{n+1}$	n an integer
$\sin(bt)$	$b/(s^2+b^2)$	
$\cos(bt)$	$s/(s^2+b^2)$	
$e^{at} \sin(bt)$	$b/((s-a)^2+b^2)$	$s > a$
$e^{at} \cos(bt)$	$(s-a)/((s-a)^2+b^2)$	$s > a$

Table 5.1: Laplace transforms of elementary functions.

Reading Exercise 5.2.3 Use your favorite computer algebra system to verify the truth of any or all entries in Table 5.1, or even do the computations by hand. Depending on your facility with integration techniques, some could be tedious.

By using the information in Table 5.1 and linearity, we can transform many of the elementary functions encountered when studying ODEs.

■ **Example 5.4** Let's compute the Laplace transform of

$$f(t) = 7e^{-2t} - 3\cos(2t) + 2t^3.$$

By invoking linearity this computation can be done by transforming each term separately. In conjunction with Table 5.1 this yields

$$\begin{aligned}\mathcal{L}(f) &= 7\mathcal{L}(e^{-2t}) - 3\mathcal{L}(\cos(2t)) + 2\mathcal{L}(t^3) \\ &= \frac{7}{s+2} - \frac{3s}{s^2+4} + \frac{12}{s^4},\end{aligned}$$

where the s was left off of the \mathcal{L} expressions. ■

Reading Exercise 5.2.4 Use Table 5.1 to compute the Laplace transform of $f(t) = t^2 - 3\sin(3t) + 5$.

Remark 5.2.1 Notice that the Laplace transform integral (5.6) only involves the values of $f(t)$ for $t \geq 0$. Even if $f(t)$ is defined for $t < 0$, these values have no bearing on the Laplace transform.

There is another important entry we can add to our table of Laplace transforms, which we now discuss.

Transforming Derivatives

Suppose that $f(t)$ is differentiable for $t \geq 0$ and $f'(t)$ is piecewise continuous and of exponential order with constant a in (5.16). Then it can be shown that $f(t)$ is also piecewise continuous (in fact, continuous, since it f is differentiable by assumption) and of exponential order with the same a , so both f' and f have Laplace transforms that are defined for $s > a$. Moreover, there is an important relationship between the transform of f and that of f' . Specifically,

$$\mathcal{L}(f')(s) = sF(s) - f(0). \quad (5.18)$$

The proof is a straightforward computation that we'll do in a moment, but first let's illustrate with an example.

■ **Example 5.5** Let $f(t) = \cos(3t)$. From Table 5.1 we find $F(s) = s/(s^2 + 9)$. The derivative of f is $f'(t) = -3\sin(3t)$. From (5.18) we have

$$\mathcal{L}(f') = \mathcal{L}(-3\sin(3t)) = sF(s) - f(0) = \frac{s^2}{s^2 + 9} - 1 = -\frac{9}{s^2 + 9}.$$

The result $\mathcal{L}(f') = -9/(s^2 + 9)$ can also be obtained directly by using the relevant formula for the Laplace transform of $\sin(3t)$ in Table 5.1. ■

Proof of Equation (5.18)

To see why (5.18) holds, let's start with the definition of the Laplace transform of $f'(t)$ given by (5.6), specifically

$$\begin{aligned} \mathcal{L}(f') &= \int_0^\infty f'(t)e^{-st} dt \\ &= \lim_{T \rightarrow \infty} \left(\int_0^T f'(t)e^{-st} dt \right). \end{aligned} \quad (5.19)$$

We'll evaluate the integral on the right in (5.19) using what is arguably the single most important technique in applied mathematics, integration by parts:

$$\int_c^d u(t)v'(t) dt = u(t)v(t) \Big|_{t=c}^{t=d} - \int_c^d u'(t)v(t) dt.$$

We'll take

$$u(t) = e^{-st} \quad \text{and} \quad v'(t) = f'(t) \quad \text{so that} \quad u'(t) = -se^{-st} \quad \text{and} \quad v(t) = f(t),$$

with $c = 0$ and $d = T$. With these choices it follows that

$$\begin{aligned} \int_0^T f'(t)e^{-st} dt &= e^{-st}f(t) \Big|_{t=0}^{t=T} + s \int_0^T f(t)e^{-st} dt \\ &= e^{-sT}f(T) - f(0) + s \int_0^T f(t)e^{-st} dt. \end{aligned} \quad (5.20)$$

Now suppose that $s > a$ where a is such that $\lim_{T \rightarrow \infty} f(t)/e^{at} = 0$. Take the limit of both sides of (5.20) as $T \rightarrow \infty$; the left side approaches $\mathcal{L}(f')(s)$ (by definition) and we find

$$\begin{aligned} \mathcal{L}(f')(s) &= \lim_{T \rightarrow \infty} \int_0^T f'(t)e^{-st} dt \\ &= \underbrace{\lim_{T \rightarrow \infty} e^{-sT}f(T)}_{= 0 \text{ since } s > a} - \underbrace{\lim_{T \rightarrow \infty} f(0)}_{f(0)} + s \underbrace{\int_0^T f(t)e^{-st} dt}_{F(s)} \\ &= -f(0) + sF(s). \end{aligned}$$

This demonstrates (5.18).

5.2.4 Solving Differential Equations Using Laplace Transforms

Let's consider how the Laplace transform can be used to solve ODEs. We'll focus on examples involving first- and second-order homogeneous equations.

First-Order Homogeneous Example

Equation (5.18) is the key to using the Laplace transform to solve linear, constant-coefficient ODEs. An example is most illustrative.

■ **Example 5.6** Let's solve the ODE

$$u'(t) = 3u(t) \quad (5.21)$$

with initial condition $u(0) = 5$. We'll use $U(s)$ to denote the Laplace transform of the solution $u(t)$.

We begin by applying the Laplace transform to both sides of the ODE (5.21). From (5.18) the transform of the left side is $sU(s) - u(0)$. By linearity the transform of the right side is $3U(s)$, so the ODE (5.21) becomes

$$sU(s) - u(0) = 3U(s). \quad (5.22)$$

Notice there are no derivatives in (5.22); the $u'(t)$ in (5.21) became $sU(s) - u(0)$ after transforming. Now fill in the initial condition $u(0) = 5$ in (5.22) to obtain

$$sU(s) - 5 = 3U(s). \quad (5.23)$$

Equation (5.23) can be solved for $U(s)$ using straightforward algebra to obtain

$$U(s) = \frac{5}{s-3}. \quad (5.24)$$

The last step is to determine $u(t)$ from $U(s)$ in (5.24). The operation that takes us from $U(s)$ to $u(t)$ is called the **inverse Laplace transform** and is the subject of Section 5.2.6, but as a first example let us pursue such a computation here. Up to this point the solution process has been fairly mechanical, but now a bit of creativity may be required to deduce $u(t)$ from $U(s)$. We need to look at the right side of (5.24) and recognize what function $u(t)$ satisfies $\mathcal{L}(u) = 5/(s-3)$. A glance at the left column in Table 5.1 makes this easier. The third line down contains the expression $1/(s-a)$, and we have to contend with $5/(s-3)$. The table shows that $1/(s-3)$ corresponds to the function e^{3t} , so by linearity $5/(s-3)$ corresponds to the function $u(t) = 5e^{3t}$. We conclude that this is the solution to (5.21) with initial condition $u(0) = 5$. ■

Study Example 5.6 carefully, as it illustrates precisely how we use the Laplace transform to solve ODEs. It's easy to verify that in Example 5.6 the solution so obtained is correct, because (5.21) can be solved using techniques we've already seen. Later, however, we'll encounter situations in which the Laplace transform is used to analyze ODEs that would be more difficult to handle using previously studied methods.

Reading Exercise 5.2.5 Emulate the computation of Example 5.6 to solve the ODE $u'(t) = -2u(t)$ with $u(0) = 3$.

The General Flow for Solving ODEs Using Laplace Transforms

As in Example 5.6, the solution process for ODEs using the Laplace transform will follow the flow of Figure 5.2. A diagram like Figure 5.2 is called a **commutative diagram**. (Some mathematics books are full of them.) The idea is that to get from the ODE in the upper left corner to the solution in the lower left corner, we follow the somewhat circuitous route obtained by Laplace transforming (moves us to the upper right corner, with \mathcal{L} indicating the operation of Laplace transforming), then

performing routine algebra (moves us to the lower right corner) and then deducing $u(t)$ from $U(s)$, to arrive at the solution in the lower left corner. This last step, the inverse Laplace transform, is indicated by the notation \mathcal{L}^{-1} and is not always cut and dried, but we will give many examples shortly.

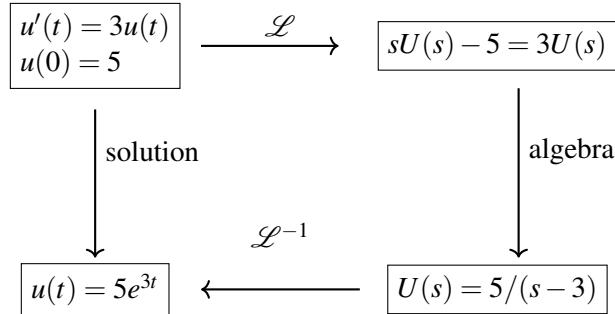


Figure 5.2: Commutative diagram for Laplace transform solution flow.

The idea behind Figure 5.2 is familiar to you from high school algebra. Suppose we want to multiply two positive real numbers x and y . If we have a finite decimal expansion or approximation for each of x and y we can just do long multiplication and grind out the product with grade school arithmetic. This is the direct path from x, y in the upper left corner to xy in the lower left corner in Figure 5.3. An alternative, if we have the means to compute logarithms and exponentials, is to compute $\ln(x)$ and $\ln(y)$ (upper right corner in Figure 5.3), then add to obtain $\ln(x) + \ln(y)$ (lower right corner), then exponentiate this sum to obtain

$$e^{\ln(x)+\ln(y)} = e^{\ln(x)} e^{\ln(y)} = xy.$$

The computation of xy comes down to doing a sum, $\ln(x) + \ln(y)$. Most people agree that addition of multi-digit numbers is simpler than multiplication, although there's no arguing that logarithms and exponentials are quite difficult to compute. Still, after transforming our multiplication problem into an addition problem, we have simplified our work. In the days before computers, this approach was commonly used, with tables for computing logarithms and exponentials. It also forms the basis for multiplication with slide rules. In the era of gigahertz computation and multi-core processors it all seems positively quaint, but the ideas are still central to mathematics.

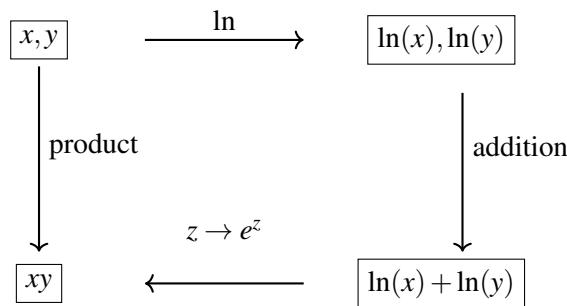


Figure 5.3: Commutative diagram for multiplication with logarithms and exponentials.

Reading Exercise 5.2.6 What would a commutative diagram for computing x^y ($x > 0$) with logarithms and exponentials look like?

Homogeneous Second-Order Examples

Before we can solve second-order equations it will be necessary to know how the Laplace transform of a function $f(t)$ is related to that of $f''(t)$. This can be done by bootstrapping off of (5.18).

Specifically,

$$\begin{aligned}\mathcal{L}(f'') &= s\mathcal{L}(f') - f'(0) && \text{(apply (5.18) to } f' \text{ instead of } f, \text{ note } f'' = (f')'') \\ &= s(sF(s) - f(0)) - f'(0) \\ &= s^2F(s) - sf(0) - f'(0),\end{aligned}\tag{5.25}$$

where as usual F denotes the Laplace transform of $f(t)$. The Laplace transform of the higher derivatives of f also have a simple relationship to F ; see Exercise 5.2.17.

■ **Example 5.7** Suppose $f(t) = te^{-2t}$. From Table 5.1 we find $F(s) = 1/(s+2)^2$. We have $f(0) = 0$, $f'(t) = -2te^{-2t} + e^{-2t}$, and so $f'(0) = 1$. From (5.25) it follows that

$$\mathcal{L}(f'')(s) = \frac{s^2}{(s+2)^2} - sf(0) - f'(0) = \frac{s^2}{(s+2)^2} - 1 = -\frac{4s+4}{(s+2)^2}.$$

■

Reading Exercise 5.2.7 Compute $f''(t)$ in Example 5.7 directly and use Table 5.1 to double-check the answer in Example 5.7.

We're now in a position to solve some second-order ODEs. In many cases some creative algebra is needed in the solution process, and this algebra strongly resembles that involved in certain integration techniques from integral calculus. Our interest is in solving the types of equations we saw in Section 4.2: linear, constant-coefficient, homogeneous, second-order ODEs. In particular, equations for overdamped, critically damped, and underdamped systems. Let's consider two examples, an overdamped system and an underdamped system.

■ **Example 5.8** Consider the ODE

$$u''(t) + 4u'(t) + 3u(t) = 0$$

with initial conditions $u(0) = 2$ and $u'(0) = 4$. We'll use $U(s)$ for the transform of $u(t)$. The general outline of the process is to Laplace transform both sides of the ODE, solve for $U(S)$ using elementary algebra, and then use this to compute $u(t)$.

The Laplace transform of the right side of the ODE is easy: $\mathcal{L}(0) = 0$. To transform the left side, use (5.25) to transform $u''(t)$ and (5.18) to transform $u'(t)$. By making use of linearity it follows that

$$s^2U(s) - su(0) - u'(0) + 4(sU(s) - u(0)) + 3U(s) = 0.$$

Be careful: the transform of $4u'(t)$ is $4(sU(s) - u(0))$; the coefficient 4 multiplies everything, including $u(0)$. Filling in the initial data $u(0) = 2$ and $u'(0) = 4$ yields

$$s^2U(s) - 2s - 4 + 4(sU(s) - 2) + 3U(s) = 0.\tag{5.26}$$

This is the step from the upper left corner in Figure 5.2 to the transformed ODE in the upper right corner, implemented by \mathcal{L} , the Laplace transform, but now applied to this second-order ODE.

The next step is to solve for $U(s)$, which moves us to the lower right corner in Figure 5.2. Collect all $U(s)$ terms on the left in (5.26) and all other terms on the right to obtain

$$(s^2 + 4s + 3)U(s) = 2s + 12.$$

Important observation: the quantity multiplying $U(s)$ on the left is exactly the characteristic polynomial for this ODE. Divide both sides above by $s^2 + 4s + 3$ to find

$$U(s) = \frac{2s+12}{s^2+4s+3}.\tag{5.27}$$

We now know that the Laplace transform $U(s)$ of the solution to the ODE is given by (5.27). The last step is to deduce $u(t)$ from $U(s)$. In order to do this we need to somehow recognize the right side of (5.27) as the transform of a function $u(t)$, but the function $U(s)$ in (5.27) doesn't look like anything in the right column of Table 5.1. A bit of creative algebra is needed to manipulate the right side of (5.27) until it matches one or more terms in the table. In integral calculus if you had to integrate the right side of (5.27) by hand with respect to s , you'd do a partial fraction decomposition. That's also exactly what works for finding $u(t)$ from $U(s)$.

Specifically, the denominator on the right in (5.27) factors as $s^2 + 4s + 3 = (s+1)(s+3)$, so following the standard rules for partial fraction decompositions we can write

$$\frac{2s+12}{s^2+4s+3} = \frac{A}{s+1} + \frac{B}{s+3} \quad (5.28)$$

for a suitable choice of constants A and B . Simplifying the right side of (5.28) by obtaining a common denominator shows that

$$\frac{A}{s+1} + \frac{B}{s+3} = \frac{(A+B)s + (3A+B)}{s^2+4s+3}. \quad (5.29)$$

Compare the left side of (5.28) to the right side of (5.29). The denominators already match; the goal is to choose A and B so that the numerators match. This occurs if $A+B=2$ (this matches the s coefficients) and $3A+B=12$ (this matches the rest). The simultaneous solution to $A+B=2$ and $3A+B=12$ is $A=5$ and $B=-3$. From (5.27) and (5.28) we obtain

$$U(s) = \frac{5}{s+1} - \frac{3}{s+3}. \quad (5.30)$$

It's now easy to deduce $u(t)$. From the third line of Table 5.1 it follows that the first term $5/(s+1)$ on the right in (5.30) corresponds to $5e^{-t}$, while $-3/(s+3)$ corresponds to $-3e^{-3t}$. From linearity it follows that $u(t) = 5e^{-t} - 3e^{-3t}$. You can easily check that this is the correct solution. Again, although we could have obtained this solution by using techniques from Chapter 4, we are practicing the Laplace transform approach in order to familiarize ourselves with the process, in anticipation of eventually tackling more challenging problems with discontinuous and impulsive forces. ■

Reading Exercise 5.2.8 Imitate the computations of Example 5.8 to solve the ODE $u''(t) + 3u'(t) + 2u(t) = 0$ with initial data $u(0) = 1$ and $u'(0) = -3$.

Let's consider an underdamped system, with an emphasis on some issues that arise such cases.

■ **Example 5.9** Consider the ODE

$$u''(t) + 4u'(t) + 20u(t) = 0$$

with initial conditions $u(0) = 2$ and $u'(0) = 4$. We use $U(s)$ for the transform of $u(t)$. Laplace transforming both sides of this ODE produces

$$s^2U(s) - su(0) - u'(0) + 4(sU(s) - u(0)) + 20U(s) = 0.$$

Fill in the initial data to obtain

$$s^2U(s) - 2s - 4 + 4(sU(s) - 2) + 20U(s) = 0. \quad (5.31)$$

Solve for $U(s)$ by first collecting all $U(s)$ terms in (5.31) on the left and everything else on the right to obtain

$$(s^2 + 4s + 20)U(s) = 2s + 12.$$

Again notice that the quantity multiplying $U(s)$ on the left is exactly the characteristic polynomial for this ODE. Solve for $U(s)$ as

$$U(s) = \frac{2s + 12}{s^2 + 4s + 20}. \quad (5.32)$$

In this case the denominator is irreducible: it does not factor, since it has no real roots.

As with the overdamped case, $U(s)$ needs a bit of algebraic preprocessing before we can employ Table 5.1. The correct course of action is dictated by techniques for evaluating integrals; if we wanted to integrate an expression like the right side of (5.32) with an irreducible denominator we'd complete the square. That's what we'll do here. Write $s^2 + 4s + 20 = (s + 2)^2 + 16 = (s + 2)^2 + 4^2$, so that

$$U(s) = \frac{2s + 12}{(s + 2)^2 + 4^2}.$$

The denominator for $U(s)$ matches the denominator of the entries in the last two lines of Table 5.1 in the case $a = -2$ and $b = 4$. From that table we see that there is a correspondence between the functions $e^{-2t} \sin(4t)$ and $e^{-2t} \cos(4t)$ and the Laplace transforms

$$\frac{4}{(s + 2)^2 + 4^2} \quad \text{and} \quad \frac{s + 2}{(s + 2)^2 + 4^2}, \quad (5.33)$$

respectively. In order to complete the complete solution process we must find $u(t)$ from $U(s)$. To do this, we recognize the numerator $2s + 12$ of $U(s)$ as a linear combination of the numerators on the right in (5.33). A little experimentation shows that $2s + 12 = 2(s + 2) + 2 \cdot 4$ so that

$$U(s) = 2 \left(\frac{s + 2}{(s + 2)^2 + 4^2} \right) + 2 \left(\frac{4}{(s + 2)^2 + 4^2} \right).$$

Now we can read $u(t)$ right off of the last two lines in the table and find

$$u(t) = 2e^{-2t} \cos(4t) + 2e^{-2t} \sin(4t).$$

■

5.2.5 The First Shifting Theorem

As you can see in the last two lines of Table 5.1, the effect of multiplying $\sin(bt)$ and $\cos(bt)$ by an exponential e^{at} is to shift the corresponding Laplace transform with respect to s , that is, replace s by $s - a$. This is true more generally and is a useful entry to our list of Laplace transform rules.

Theorem 5.2.2 — First Shifting Theorem. If $f(t)$ is piecewise continuous and of exponential order for $t \geq 0$, and if $f(t)$ has Laplace transform $F(s)$ defined for $s > c$, then

$$\mathcal{L}(e^{at}f(t)) = F(s - a),$$

and this transform is defined for $s > c + a$.

The proof is a straightforward computation:

$$\begin{aligned} \mathcal{L}(e^{at}f(t))(s) &= \int_0^\infty e^{-st}(e^{at}f(t))dt && \text{(the very definition of } \mathcal{L}(e^{at}f(t))\text{)} \\ &= \int_0^\infty e^{-(s-a)t}f(t)dt && \text{(since } e^{-st}e^{at} = e^{-(s-a)t}\text{)} \\ &= F(s - a). \end{aligned}$$

Moreover, if F is defined for $s > c$ then $F(s - a)$ is defined for $s > a - c$.

■ **Example 5.10** Let's compute $\mathcal{L}(e^{3t}t^2)$. Define $f(t) = t^2$ so the Laplace transform of f is $F(s) = 2/s^3$ from Table 5.1. From the first shifting theorem then $\mathcal{L}(e^{3t}t^2) = F(s-3) = 2/(s-3)^3$. This result could also have been obtained directly from the fourth line of Table 5.1. ■

Reading Exercise 5.2.9 Use the first shifting theorem (Theorem 5.2.2) to compute $\mathcal{L}(te^{-t})$.

5.2.6 The Inverse Laplace Transform

We can always use the definition (5.6) to compute the Laplace transform of a function $f(t)$, if we can work the integral. The more desirable approach is to use Table 5.1, which facilitates this process by tabulating the transforms of commonly encountered functions. As we've seen, solving ODEs requires that we also be able to find $u(t)$ from $U(s)$. This process is called the **inverse Laplace transform**. For the examples we've considered we used algebraic manipulation and inverse table look-up using Table 5.1. That is, we manipulate $U(s)$, often splitting it into several pieces, until we can recognize each piece in the right column of Table 5.1. You might wonder if there's an easier and more methodical to do an inverse Laplace transform. If $U(s)$ is known, is there a simple way to compute $u(t)$, perhaps by doing an integral?

But we must first ponder whether the Laplace transform is even invertible. That is, if functions $u_1(t)$ and $u_2(t)$ have the same Laplace transform $U(s)$, must u_1 and u_2 be the same function? If this isn't true then there's no point in looking for an easy inversion formula, since for a given $U(s)$ there may be no unique function $u(t)$ to go back to. Happily, it turns out that the Laplace transform is invertible for the types of functions we're using.

Theorem 5.2.3 If two piecewise continuous functions $u_1(t)$ and $u_2(t)$ of exponential order have Laplace transforms $U_1(s)$ and $U_2(s)$ respectively, with both transforms defined for $s > a$ for some constant a , and $U_1(s) = U_2(s)$ for $s > a$, then $u_1(t) = u_2(t)$ at all points where both u_1 and u_2 are continuous.

For a proof see [46]. It is possible for u_1 and u_2 to disagree at a point where either function has a jump discontinuity, but this is of no practical consequence. We'll explore this in the next section.

As we write $F(s) = \mathcal{L}(f(t))$ for the Laplace transform, we will write $f(t) = \mathcal{L}^{-1}(F(s))$ to denote the inverse Laplace transform of $F(s)$.

Inverting the Laplace Transform

Given that the Laplace transform is invertible, is there a formula or recipe for actually computing the inverse? There are at least two. The most common is the Bromwich integral which, unfortunately, involves the subject of complex analysis and is beyond the scope of this text. Another is the Post inversion formula, which involves only elementary calculus. Unfortunately, this formula is generally impractical, although modifications exist that can be used to approximate the inverse Laplace transform of a function numerically. See [28] for a discussion of this formula, examples, and an easy proof of its validity. See also Exercise 5.2.19. In general if we want to invert a Laplace transform using pencil and paper computation we must use the approach of this section, involving inspired algebraic manipulation and inverse table lookup based on Table 5.1.

In reality, after gaining an understanding of the basic ideas and flow of how the Laplace transform facilitates the analysis of ODEs, we appeal to technology and software like Maple, Mathematica, or Sage, much as we do for evaluating complicated integrals. In the following sections many of the computations that come up are not reasonable to tackle by hand. Tutorials have been posted on the book website [8] that illustrate how to use these software packages to handle Laplace and inverse Laplace transforms, and how to apply them to ODEs.

The Time Domain and the s -Domain

We'll start using a bit of terminology that is very common in science, engineering, and mathematics when applying the Laplace transform. The **time domain** or **t -domain** encompasses all those functions and objects (such as ODEs and solutions) that are expressed using t as the independent variable. Quantities that are expressed in terms of the Laplace parameter s are said to be in the **s -domain**, or the **Laplace s -domain**, or sometimes even the **frequency domain**. Each object or linear operation in the time domain has a counterpart in the s -domain and vice-versa. For example, the ODE $u'(t) = 3u(t)$ with $u(0) = 5$ has s -domain counterpart $sU(s) - 5 = 3U(s)$ (recall Example 5.6). As another example, multiplying a function by e^{ct} in the time domain shifts the Laplace transform c units to the right in the s -domain. The Laplace transform is what moves us from the time domain to the s -domain. The inverse Laplace transform takes us from the s -domain back to the time domain.

Reading Exercise 5.2.10 Given that t has the dimension time, what must be the dimension of the Laplace transform parameter s , given that e^{-st} must be computed as part of the transform? (Recall the subsection “Elementary Functions” in Section 1.5.) Do you see why s might be considered a kind of frequency?

Important Intuition for the Inverse Laplace Transform

The process of inverting Laplace transforms can be computationally tedious, but we can frequently use $F(s)$ to glean some important information about $f(t)$ in the time domain without fully inverting the transform $F(s)$. In most cases the transform $F(s)$ will be a **rational function** (a ratio of two polynomials) of the form

$$F(s) = \frac{p(s)}{q(s)},$$

where the denominator $q(s)$ is a polynomial, say of degree n , while $p(s)$ is a polynomial of degree strictly less than n ; both p and q will have real coefficients. It turns out that by merely finding the roots of $q(s)$ (solutions to $q(s) = 0$) and doing nothing else we can determine much about $f(t)$. See Appendix A for a discussion of rational functions.

■ **Example 5.11** Let's look at an example. Consider the s -domain transform

$$F(s) = \frac{2s^2 + 11s + 65}{s^3 + 4s^2 + 21s + 34}. \quad (5.34)$$

What can we deduce about $f(t) = \mathcal{L}^{-1}(F(s))$ back in the time domain? If we want to invert $F(s)$, the first step is to factor the denominator of $F(s)$ as completely as possible. This isn't always easy, but it is a fact that every polynomial with real coefficients can be factored into a product of linear pieces and irreducible quadratic pieces (see Appendix A). Here we find

$$s^3 + 4s^2 + 21s + 34 = (s + 2)(s^2 + 2s + 17).$$

The quadratic piece $s^2 + 2s + 17$ is irreducible, that is, does not factor. Based on this factorization we would next perform a partial fraction expansion on $F(s)$, of the form

$$F(s) = \frac{A_1}{s+2} + \frac{A_2s + A_3}{s^2 + 2s + 17}, \quad (5.35)$$

and then figure out A_1, A_2 , and A_3 . But even without knowing A_1, A_2 , and A_3 , we can already see from the first term on the right in (5.35) and Table 5.1 that $f(t)$ will contain a term of the form $A_1 e^{-2t}$ (unless it just so happens that $A_1 = 0$). By completing the square $s^2 + 2s + 17 = (s + 1)^2 + 4^2$ and appealing to Table 5.1 we also see that the second term on the right in (5.35) will give rise to multiples of $e^{-t} \sin(4t)$ and $e^{-t} \cos(4t)$. So $f(t)$ will consist of a linear combination of e^{-2t} , $e^{-t} \sin(4t)$, and $e^{-t} \cos(4t)$. ■

■ **Example 5.12** Suppose $F(s) = p(s)/q(s)$ where $q(s) = 2(s-1)^2(s+3)(s^2+2s+10)$. What can we deduce about $f(t)$? The $(s-1)^2$ term will give rise to terms $A_1/(s-1)$ and $A_2/(s-1)^2$ in the partial fraction decomposition of $F(s)$, which map back to multiples of e^t and te^t in the time domain (line 4 in Table 5.1). The $1/(s+3)$ term in $q(s)$ gives rise to a term $A_3/(s+3)$ in the partial fraction decomposition and this maps back to a multiple of e^{-3t} in the time domain. Finally, by completing the square, the $s^2+2s+10$ term can be written $(s+1)^2+3^2$, and so the decomposition of $F(s)$ has a term $(A_4s+A_5)/((s+1)^2+3^2)$, which maps back to multiples of $e^{-t}\sin(3t)$ and $e^{-t}\cos(3t)$ in the time domain. The function $f(t)$ will be a linear combination of terms

$$e^t, \quad te^t, \quad e^{-3t}, \quad e^{-t}\sin(3t), \quad \text{and} \quad e^{-t}\cos(3t).$$

Reading Exercise 5.2.11 If $F(s) = \frac{3s+3}{s^2+7s+10}$, what terms would you expect the inverse transform $f(t)$ to contain, given that $s^2+7s+10 = (s+2)(s+5)$? Verify your guess by actually computing the inverse transform.

Complex Roots and Poles

As usual, embracing complex numbers makes things even easier. The fundamental theorem of algebra (see Appendix A) states that any n th degree polynomial with complex coefficients factors completely over the complex numbers. That is, if

$$q(s) = A_n s^n + A_{n-1} s^{n-1} + \cdots + A_1 s + A_0,$$

then

$$q(s) = A_n (s - r_1)^{m_1} (s - r_2)^{m_2} \cdots (s - r_k)^{m_k}, \quad (5.36)$$

where r_1, \dots, r_k are the distinct roots of $q(s)$, that is, the solutions to $q(s) = 0$. The exponent m_j is called the **multiplicity** of r_j . If $q(s)$ has real coefficients then each root r_j is either a real number or one of a pair of roots that are complex-conjugate to one another (again, see Appendix A). As an example, if $q(s) = s^3 + 4s^2 + 21s + 34$ as in Example 5.11 then

$$q(s) = (s+2)(s^2+2s+17) = (s+2)(s - (-1+4i))(s - (-1-4i)).$$

The roots of $q(s)$ are $-2, -1+4i$, and $-1-4i$; each has multiplicity 1.

Here's a fact that you may not have seen before: partial fraction expansions work just fine with complex numbers. This is explored more fully in Section A.5 of Appendix A, but to provide a brief illustration, suppose we want to perform a partial fraction expansion on $F(s)$ as in (5.34). Instead of using a partial fraction expansion of the form (5.35) we try

$$F(s) = \frac{A_1}{s+2} + \frac{A_2}{s - (-1+4i)} + \frac{A_3}{s - (-1-4i)}. \quad (5.37)$$

Here's another convenient fact: The correspondence between $1/(s-b)$ in the s -domain and e^{bt} in the time domain is perfectly valid if b is a complex number. Applying this fact to (5.37) shows that $f(t)$ must be of the form

$$f(t) = A_1 e^{-2t} + A_2 e^{(-1+4i)t} + A_3 e^{(-1-4i)t}.$$

Now applying Euler's formula shows that $e^{(-1+4i)t} = e^{-t}\cos(4t) + ie^{-t}\sin(4t)$ and $e^{(-1-4i)t} = e^{-t}\cos(4t) - ie^{-t}\sin(4t)$. The upshot is that $f(t)$, which must be real-valued, will be a superposition of terms

$$e^{-2t}, \quad e^{-t}\cos(4t), \quad \text{and} \quad e^{-t}\sin(4t).$$

We can deduce all of this without working out the A_k ; all we need to do is find the roots of $q(s)$. These roots (the places where F is undefined) are called the **poles** of $F(s)$.

■ **Example 5.13** Let $F(s) = \frac{p(s)}{(s-4)^3(s^2+4s+8)}$, for some fourth degree polynomial $p(s)$. The quadratic $s^2 + 4s + 8$ polynomial has roots $s = -2 \pm 2i$, so the roots of $q(s)$ are $s = -2 + 2i$ and $s = -2 - 2i$, each with multiplicity 1, and $s = 4$ with multiplicity 3. The partial fraction expansion for $F(s)$ is of the form

$$F(s) = \frac{A_1}{s-4} + \frac{A_2}{(s-4)^2} + \frac{A_3}{(s-4)^3} + \frac{A_4}{s-(-2+2i)} + \frac{A_5}{s-(-2-2i)}.$$

From Table 5.1 the function $f(t) = \mathcal{L}^{-1}(F(s))$ will thus be a linear combination of the terms

$$e^{4t}, \quad te^{4t}, \quad t^2e^{4t}, \quad e^{-2t}\cos(2t), \quad \text{and} \quad e^{-2t}\sin(2t),$$

although the coefficient of any of these terms might turn out to be zero, depending on $p(s)$. ■

Reading Exercise 5.2.12 Given that $F(s) = (6s+2)/(s^2+4)$ has poles where $s^2 + 4 = 0$, namely $s = 2i$ and $s = -2i$, what kinds of complex exponentials would you expect to appear in the inverse transform? What types of real-valued expressions would these correspond to? Compute the actual inverse transform.

Summary: Poles and the Inverse Transform

Here's a summary of how to deduce information about a function $f(t)$ from its Laplace transform $F(s)$, without doing too much computation.

- Suppose

$$F(s) = \frac{p(s)}{q(s)}$$

is a rational function, where the polynomial $q(s)$ is of degree n and the degree of p is strictly less than n .

- Find the poles of $F(s)$, that is, the distinct roots r_1, r_2, \dots, r_k of $q(s)$, and their multiplicities; in short, factor $q(s)$ completely in the form (5.36).
- If r_j is real and has multiplicity m_j then $f(t)$ contains a superposition of terms

$$e^{r_j t}, \quad te^{r_j t}, \quad \dots, \quad t^{m_j-1}e^{r_j t}.$$

- If $r_j = \alpha_j \pm i\beta_j$ is complex and has multiplicity m_j then $f(t)$ contains a superposition of terms

$$e^{\alpha_j t} \sin(\beta_j t), \quad e^{\alpha_j t} \cos(\beta_j t), \quad \dots, \quad t^{m_j-1}e^{\alpha_j t} \sin(\beta_j t), \quad t^{m_j-1}e^{\alpha_j t} \cos(\beta_j t).$$

Once you've found a root $\alpha_j + i\beta_j$, there's no need to worry about its conjugate partner; the other root $\alpha_j - i\beta_j$ will generate the same types of terms in the time domain.

5.2.7 The Initial and Final Value Theorems

If $F(s) = \mathcal{L}(f(t))$, there is an interesting relationship between the behavior of $F(s)$ when $s \rightarrow 0^+$ and $f(t)$ as $t \rightarrow \infty$, and, conversely, between the behavior of $F(s)$ as $s \rightarrow \infty$ and of $f(t)$ as $t \rightarrow 0^+$. These relations are useful later in Section 5.6, and sometimes they're handy as a quick sanity check on whether a given transform pair $f(t)$ and $F(s)$ is correct.

The initial value theorem for the Laplace transforms relates the behavior of $f(t)$ as $t \rightarrow 0$ to that of $F(s)$ as $s \rightarrow \infty$.

Theorem 5.2.4 — Initial Value Theorem for the Laplace Transform. If $f(t)$ is piecewise continuous and of exponential order, and if $\lim_{t \rightarrow 0^+} f(t)$ exists, then

$$\lim_{t \rightarrow 0^+} f(t) = \lim_{s \rightarrow \infty} sF(s)$$

where $F(s) = \mathcal{L}(f(t))$.

For a proof see [46].

■ **Example 5.14** Let $f(t) = e^t$, so $F(s) = 1/(s - 1)$. Then

$$\lim_{t \rightarrow 0^+} f(t) = \lim_{t \rightarrow 0^+} e^t = 1$$

and

$$\lim_{s \rightarrow \infty} sF(s) = \lim_{s \rightarrow \infty} \frac{s}{s - 1} = 1.$$

■

The final value theorem is similar, but interchanges the role of f and F . It also requires an extra hypothesis on F .

Theorem 5.2.5 — Final Value Theorem for the Laplace Transform. Suppose $f(t)$ is piecewise continuous and of exponential order, and that $F(s) = \mathcal{L}(f(t))$. Suppose also that every pole of F is of the form $s = a + bi$ with $a < 0$ or if F has a pole at $s = 0$, then this pole is of multiplicity 1. Then $\lim_{t \rightarrow \infty} f(t)$ exists and

$$\lim_{t \rightarrow \infty} f(t) = \lim_{s \rightarrow 0^+} sF(s).$$

■

For a proof see [46].

■ **Example 5.15** Let $f(t) = 3 + te^{-t}$. We can compute $F(s) = \frac{3s^2 + 7s + 3}{s(s+1)^2}$, which has a pole of multiplicity 2 at $s = -1$ and a pole of multiplicity 1 at $s = 0$, so the hypotheses of Theorem 5.2.5 are satisfied. Then

$$\lim_{t \rightarrow \infty} f(t) = \lim_{t \rightarrow \infty} (3 + te^{-t}) = 3$$

and

$$\begin{aligned} \lim_{s \rightarrow 0^+} sF(s) &= \lim_{s \rightarrow 0^+} \frac{s(3s^2 + 7s + 3)}{s(s+1)^2} \\ &= \lim_{s \rightarrow 0^+} \frac{3s^2 + 7s + 3}{(s+1)^2} \\ &= 3. \end{aligned}$$

■

The hypotheses concerning the poles of F in Theorem 5.2.5 are needed. For example, if $f(t) = \sin(t)$ then $F(s) = 1/(s^2 + 1)$. Then F has two poles, one at $s = i$, another at $s = -i$, both with zero (but not negative) real part. The limit $\lim_{s \rightarrow 0^+} sF(s) = 0$, but $\lim_{t \rightarrow \infty} f(t)$ does not exist.

Reading Exercise 5.2.13 Let $f(t) = 2 + 3e^{-t}$. Compute $F(s)$ and verify that Theorem 5.2.4 and Theorem 5.2.5 hold in this case.

5.2.8 Section Summary and Remarks

The Laplace transform provides an alternative to the methods we've previously seen for solving linear, constant-coefficient ODEs, but it's time for a disclaimer: For these equations this method does not make the computations easier, as you now realize if you compared the homogeneous second-order examples of this section to the techniques of Section 4.2. The real value of the Laplace transform will be revealed in the next few sections, in conjunction with allied ideas for handling discontinuous and impulsive forcing functions. The intuition of the last subsection, about the poles of the transform $F(s)$, can also yield important insights about solutions to ODEs.

5.2.9 Exercises

Exercise 5.2.1 Use Table 5.1 to Laplace transform the following functions. There may be multiple ways to arrive at the answer.

- (a) $f(t) = 3t^2$
- (b) $g(t) = \sin(4t) + 7t - e^{2t}$
- (c) $p(t) = e^{-3t} \cos(7t)$
- (d) $f(t) = (1-t)^2$
- (e) $q(t) = t^3 e^{5t}$

Exercise 5.2.2 Use Table 5.1 to compute the inverse Laplace transform of the following functions. There may be multiple ways to arrive at the answer.

- (a) $F(s) = \frac{1}{s^2} - \frac{2}{s}$
- (b) $Q(s) = \frac{1}{s^2+4}$
- (c) $G(s) = \frac{2s+2}{s^2+4}$
- (d) $F(s) = \frac{4s}{s^2+4s+8}$
- (e) $F(s) = \frac{2}{(s+3)^3}$

Exercise 5.2.3 Suppose $F(s) = p(s)/q(s)$ with $q(s)$ as listed below, and with $p(s)$ a polynomial of degree less than the degree of $q(s)$. Find the poles of F (roots of $q(s)$) and then deduce what kinds of terms (e.g., e^{3t} , $e^{-t} \sin(5t)$, etc.) make up $f(t)$.

- (a) $q(s) = (s+1)(s+2)$
- (b) $q(s) = (s+1)^2(s+2)$
- (c) $q(s) = s^2 + 1$
- (d) $q(s) = (s^2 + 1)(s^2 + 9)$
- (e) $q(s) = (s^2 + 2s + 2)(s - 1)^3$
- (f) $q(s) = (s^2 + 5s + 4)^2$
- (g) $q(s) = (s^2 + 4s + 13)^3(s + 3)^7$

Exercise 5.2.4 Solve the following initial value problems using the method of Laplace transforms.

- (a) $u'(t) = 2u(t)$ with $u(0) = 6$
- (b) $u'(t) = -5u(t)$ with $u(0) = -4$
- (c) $u'(t) = au(t)$ with $u(0) = u_0$, where a and u_0 are constants.

Exercise 5.2.5 Solve the following initial value problems using the method of Laplace transforms.

- $u''(t) + 3u'(t) + 2u(t) = 0$ with $u(0) = 6$ and $u'(0) = 4$
- $4u''(t) + 8u'(t) + 4u(t) = 0$ with $u(0) = 5$ and $u'(0) = 3$
- $u''(t) + 2u'(t) + 10u(t) = 0$ with $u(0) = 1$ and $u'(0) = 2$
- $2u''(t) + 22u'(t) + 36u(t) = 0$ with $u(0) = 1$ and $u'(0) = 12$
- $3u''(t) + 6u'(t) + 6u(t) = 0$ with $u(0) = 1$ and $u'(0) = -2$
- $3u''(t) + 18u'(t) + 27u(t) = 0$ with $u(0) = 1$ and $u'(0) = -2$

Exercise 5.2.6 Use the fact that $\sinh(t) = (e^t - e^{-t})/2$ and $\cosh(t) = (e^t + e^{-t})/2$ to compute the Laplace transform of the hyperbolic sine and hyperbolic cosine functions. Compare to the transforms of the sine and cosine functions.

Exercise 5.2.7 Compute the Laplace transform of e^{it} (where $i = \sqrt{-1}$), treating i as a constant by taking $a = i$ in Table 5.1. Compare the result to the Laplace transform of $\cos(t) + i\sin(t)$. Try simplifying the difference. Do the rules of Table 5.1 seem to work for complex exponents?

Exercise 5.2.8 Let $H(t)$ be defined as

$$H(t) = \begin{cases} 0, & t \leq 0 \\ 1, & t > 0. \end{cases}$$

Compute $\mathcal{L}(H(t - c))(s)$, the Laplace transform of $H(t - c)$, where $c > 0$ is some constant; assume $s > 0$. Hint: break the integral of $e^{-st}H(t - c)$ up into two pieces, one from $t = 0$ to $t = c$ and one from $t = c$ to $t = \infty$, and evaluate each separately; the first integral is easy.

The function $H(t)$ is called the **Heaviside function** or **unit step function** and it plays a prominent role in the remainder of this chapter. The function $H(t)$ can be used as a kind of mathematical switch that turns from off to on at $t = 0$; $H(t - c)$ turns on at $t = c$.

Exercise 5.2.9 Let $f(t)$ be defined by

$$f(t) = \begin{cases} 3, & 0 \leq t < 5 \\ 7, & 5 \leq t < 10 \\ 0, & t \geq 10. \end{cases}$$

Compute $\mathcal{L}(f)$. Hint: break the integral of $e^{-st}f(t)$ up into three pieces, one from $t = 0$ to $t = 5$, one from $t = 5$ to $t = 10$, and one from $t = 10$ to $t = \infty$, and evaluate each separately. The last one is easy. For what arguments s does the integral that defines $\mathcal{L}(f)$ converge?

Exercise 5.2.10 Suppose $f(t)$ has Laplace transform $F(s)$. Show that the Laplace transform of $tf(t)$ is $-\frac{dF}{ds}$. Hint: start with (5.6) and differentiate both sides with respect to s . Assume you can differentiate under the integral (that is, move the s derivative inside the integral).

Exercise 5.2.11 Use the result of Exercise 5.2.10 to compute $\mathcal{L}(te^{-2t}\sin(3t))$. Hint: compute the transform of $e^{-2t}\sin(3t)$ first.

Exercise 5.2.12 Compute the Laplace transform $F(s)$ for each function below and check the assertion of Theorem 5.2.4, where applicable.

- (a) $f(t) = 1$
- (b) $f(t) = t$
- (c) $f(t) = e^t$
- (d) $f(t) = \cos(t)$
- (e) $f(t) = \sin(t)/t$; here $F(s) = \arctan(1/s)$, but that's not in our table.

Exercise 5.2.13 Compute the Laplace transform $F(s)$ for each function below and check that the hypotheses of Theorem 5.2.5 are satisfied. If so, verify the assertion of that theorem.

- (a) $f(t) = 4$
- (b) $f(t) = e^{-t}$
- (c) $f(t) = t^4 e^{-t}$
- (d) $f(t) = 2 + e^{-3t} \cos(t)$

Exercise 5.2.14 A spring-mass-damper system with mass $m = 2$ kg, damping constant $c = 8$ newtons per meter per second, and spring constant $k = 40$ newtons per meter, and no other forces on the mass, starts with initial data $x(0) = 1/2$ meter and $x'(0) = 0$ meters per second, where $x(t)$ is the displacement of the mass from equilibrium. Write out the appropriate ODE for $x(t)$ and solve it using the method of Laplace transforms.

Exercise 5.2.15 Consider a series RLC circuit with no voltage source, capacitance $C = 10^{-4}$ C, resistance $R = 2$ ohms, and inductance $L = 2 \times 10^{-4}$ H. The capacitor starts with charge $q = 0$ at time $t = 0$ and the current in the circuit at this instant is $I = 1$ ampere. Formulate the appropriate ODE for the charge $q(t)$ on the capacitor and solve this ODE using the method of Laplace transforms.

Exercise 5.2.16 What fundamental problem arises if you try to solve the logistic equation

$$u'(t) = ru(t)(1 - u(t)/K)$$

using the Laplace transform?

Exercise 5.2.17

- (a) Let $f(t)$ be a function defined for $t \geq 0$. Assume that f is three times differentiable and that $f, f', f'',$ and f''' are all piecewise continuous and of exponential order (so we can Laplace transform them). Let $F(s)$ denote the Laplace transform of f . Show that

$$\mathcal{L}(f''') = s^3 F(s) - s^2 f(0) - s f'(0) - f''(0).$$

Hint: apply (5.18) to f'' , noting that $(f'')' = f'''$.

- (b) Perform a similar computation for $f^{(4)}$ (the fourth derivative of f), assuming that $f^{(4)}$ is piecewise continuous and of exponential order.

- (c) Show that

$$\mathcal{L}(f^{(n)}) = s^n F(s) - s^{n-1} f(0) - s^{n-2} f'(0) - \cdots - s^2 f^{(n-3)}(0) - s f^{(n-2)}(0) - f^{(n-1)}(0)$$

if all derivatives are piecewise continuous and of exponential order. Thus each derivative we take in the time domain corresponds to multiplying by s in the s -domain, aside from contributions by f and its derivatives at $t = 0$.

Exercise 5.2.18 Given a continuous function $f(t)$ defined for $t \geq 0$, define a function

$$g(t) = \int_0^z f(z) dz.$$

Note that by the fundamental theorem of calculus, $g'(t) = f(t)$, and $g(0) = 0$. Solve the differential equation $g'(t) = f(t)$ using Laplace transforms to show that

$$G(s) = F(s)/s$$

where F and G are the Laplace transforms of f and g , respectively. Thus integration in the time domain corresponds to division by s in the s -domain.

Exercise 5.2.19 Let $f(t)$ be a piecewise continuous function of exponential order for $t \geq 0$ and let $F(s)$ be the Laplace transform of $f(t)$. **Post's inversion formula** (also called the **Post-Widder inversion formula**) gives a method for inverting the Laplace transform, that is, computing $f(t)$ from $F(s)$. If $F^{(k)}$ denotes the k th derivative of F then

$$f(t) = \lim_{k \rightarrow \infty} \frac{(-1)^k}{k!} \left(\frac{k}{t}\right)^{k+1} F^{(k)}\left(\frac{k}{t}\right). \quad (5.38)$$

For an elementary proof of the formula see [28].

This inversion formula would be practical, except that we have to differentiate F an arbitrarily large number of times and figure out the limit. Unfortunately, most functions become messier and messier as they are repeatedly differentiated. However, the formula is easy to use in certain simple cases.

- (a) Let $f(t) = e^{-t}$ so $F(s) = 1/(s+1)$. Show that when $k = 1$ the expression inside the limit on the right side of (5.38) equals $1/(1+t)^2$. Plot e^{-t} and $1/(1+t)^2$ for $0 \leq t \leq 5$.
- (b) Repeat part (a) but with $k = 2$. Show that the expression inside the limit on right side of (5.38) equals $1/(1+t/2)^3$. Plot e^{-t} and $1/(1+t/2)^2$ for $0 \leq t \leq 5$.
- (c) Repeat part (a) but with $k = 5$. Show that the expression inside the limit on right side of (5.38) equals $1/(1+t/5)^6$. Plot e^{-t} and $1/(1+t/5)^5$ for $0 \leq t \leq 5$.
- (d) Based on parts (a)-(c), make a conjecture for the form of the expression inside the limit on the right side of (5.38), and how it depends on t and k . Can you prove it? Can you prove that your conjectured expression approaches e^{-t} as $k \rightarrow \infty$?
- (e) Another simple case is when $f(t) = t$, so $F(s) = -1/s^2$. Show in this case that the expression inside the limit on the right side of (5.38) equals $(1 + 1/k)t$. What is the limit of this expression as $k \rightarrow \infty$?

5.3 Nonhomogeneous Problems and Discontinuous Forcing Functions

In this section we consider linear, constant-coefficient ODEs that are nonhomogeneous, with a particular focus on forcing functions that are piecewise continuous. We'll look at examples of how to implement these types of forcing functions using the Heaviside function, with many applications, and how the Laplace transform facilitates the solution process.

5.3.1 Some Nonhomogeneous Examples

All of the examples in the previous section were homogeneous linear equations, but the Laplace transform works perfectly well on linear, nonhomogeneous ODEs with constant-coefficients. To illustrate, here are some brief examples, one first-order and one second-order.

■ **Example 5.16** Let's solve the ODE $u'(t) = 3u(t) + 6$ with initial condition $u(0) = 1$. Laplace transforming both sides of the ODE yields

$$sU(s) - 1 = 3U(s) + 6/s,$$

where $U(s)$ is the transform of the solution and the initial data has been incorporated. Solving for $U(s)$ produces

$$U(s) = \frac{s+6}{s(s-3)}.$$

A partial fraction decomposition shows that

$$U(s) = -\frac{2}{s} + \frac{3}{s-3}.$$

We can read the inverse transform right off of Table 5.1 to see that

$$u(t) = -2 + 3e^{3t}.$$

■

The process for second-order equations is similar.

■ **Example 5.17** Let us solve the ODE $u''(t) + 3u'(t) + 2u(t) = 2e^{-3t}$ with initial data $u(0) = 1$ and $u'(0) = 2$. Transform both sides of the ODE to find that

$$(s^2 + 3s + 2)U(s) = 5 + s + \frac{2}{s+3},$$

after filling in the initial data, collecting all $U(s)$ terms on the left, and everything else (including the transform of $2e^{-3t}$) on the right. Solving for $U(s)$ produces

$$U(s) = \frac{s^2 + 8s + 17}{(s+1)(s+2)(s+3)}.$$

A partial fraction decomposition leads to

$$U(s) = \frac{s^2 + 8s + 17}{(s+1)(s+2)(s+3)} = \frac{5}{s+1} - \frac{5}{s+2} + \frac{1}{s+3}.$$

We can read the inverse transform off of Table 5.1 to find

$$u(t) = 5e^{-t} - 5e^{-2t} + e^{-3t}.$$

■

5.3.2 Discontinuous Forcing

The Laplace transform gives a unified framework for solving nonhomogeneous ODEs that involve forcing functions with jump discontinuities, a common occurrence in applications. To illustrate, let us return to the drug dosing problem from Section 5.1.

Morphine Administration

Recall the morphine administration problem from Section 5.1, in particular, the ODE model

$$u'(t) = r(t) - ku(t), \quad (5.39)$$

where $u(t)$ is the amount of morphine (mg) in the patient's system, t is time in hours, $k \approx 0.173$ (reciprocal hours), and $r(t)$ is the rate at which morphine is being administered in mg per hour. Suppose the patient is given a 10 mg bolus at time $t = 0$, so $u(0) = 10$. The function $r(t)$ of interest is given by (5.3), reproduced here:

$$r(t) = \begin{cases} 1.5, & 0 \leq t \leq 12 \\ 2.08, & t > 12. \end{cases} \quad (5.40)$$

This choice for $r(t)$ was motivated by a need to increase the amount of morphine in the patient's system, which was inadequate, starting at time $t = 12$. The approach used in Section 5.1 to solve (5.39) with this $r(t)$ was to break the problem up into two time intervals, $0 \leq t \leq 12$ and $t > 12$, solve on each interval separately, and then stitch the solutions together continuously at $t = 12$. This led to the solution (5.4). But Laplace transforms provide a more elegant approach. To facilitate this process we introduce the Heaviside function.

The Heaviside Function

Rather than write $r(t)$ using traditional piecewise notation, let's make use of the **Heaviside function** (named for Oliver Heaviside, 1850-1925, an influential electrical engineer, mathematician, and physicist).

Definition 5.3.1 The *Heaviside function* $H(t)$ (also known as the *unit step function*) is defined as

$$H(t) = \begin{cases} 0, & t \leq 0 \\ 1, & t > 0. \end{cases}$$

However, definitions for $H(t)$ may vary slightly from one text to another, specifically, the value assigned to $H(0)$. Our definition yields $H(0) = 0$, but some texts take $H(t) = 0$ for $t < 0$ and $H(t) = 1$ for $t \geq 0$, or $H(t) = 0$ for $t < 0$, $H(t) = 1$ for $t > 0$, and $H(0) = 1/2$. It doesn't matter, at least for any task we will undertake. From a physical perspective, the Heaviside function is supposed to model a switch being flipped instantaneously from off to on. Does it really matter whether the switch was fully off, fully on, or anywhere in between at $t = 0$? From a mathematical perspective, in any problem we encounter, the Heaviside function will ultimately appear under an integral and the value of $H(t)$ at any single point like $t = 0$ will not affect the value of the integral. Thus $H(0)$ can be taken as any convenient real number. These variations in the definition of $H(t)$ extend to software: The Maple software package leaves $H(0)$ undefined, Matlab sets $H(0) = 0.5$, and Mathematica's corresponding UnitStep command sets $H(0) = 1$.

The Heaviside function is plotted in the left panel of Figure 5.4. Some texts may use the notation $u(t)$ for the Heaviside function, especially if they call it the unit step function. For any real constant c , $H(t - c)$ is a function in which the switch from off to on occurs at $t = c$ instead of $t = 0$, and this can be a very useful way to construct functions with jump discontinuities. The function $H(t - c)$ is plotted in the right panel of Figure 5.4.

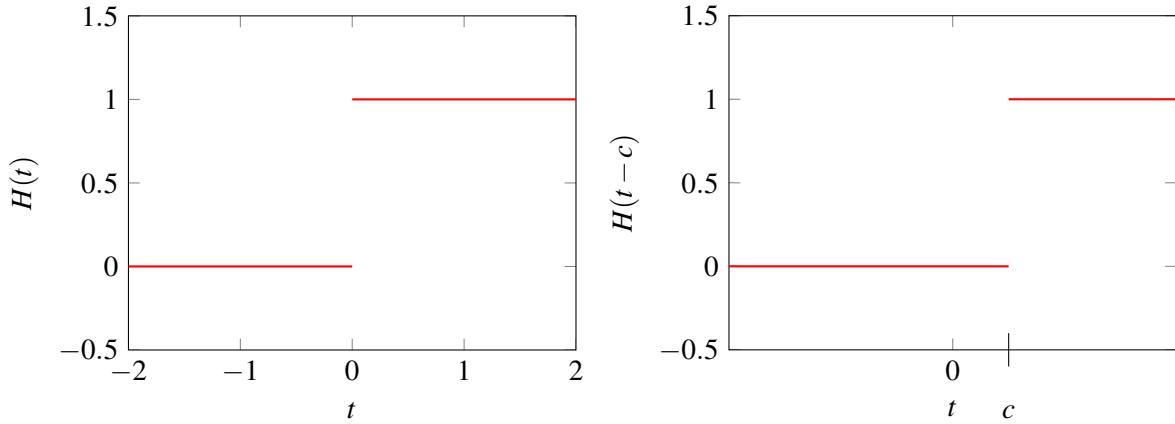


Figure 5.4: Left panel: the graph of the Heaviside function $H(t)$. Right panel: the graph of $H(t - c)$ for some $c > 0$.

■ **Example 5.18** Let's use the Heaviside function $H(t)$ to express the function

$$\phi(t) = \begin{cases} 0, & 0 \leq t \leq 2 \\ 3, & 2 < t \leq 5 \\ 0, & t > 5 \end{cases} .$$

The function $\phi(t)$ can be constructed as

$$\phi(t) = 3(H(t - 2) - H(t - 5)).$$

To see why this works, suppose $t \leq 2$. Then both $t - 2 \leq 0$ and $t - 5 \leq 0$, so both Heaviside functions are switched off (equal 0). When $2 < t \leq 5$ the function $H(t - 2)$ is on and equals 1, but $H(t - 5)$ is still off, so $\phi(t) = 3 \cdot 1 = 3$. For $t > 5$ the $H(t - 5)$ piece is also switched on, cancelling out the $H(t - 2)$ piece, and $\phi(t)$ drops back to the value 0. We should note that in an application the precise value of $\phi(t)$ at the jump discontinuities $t = 2$ and $t = 5$ would not be important. ■

■ **Example 5.19** Any piecewise defined function can be expressed using Heaviside functions. An example is probably more illuminating than any abstract formula. Consider the function

$$q(t) = \begin{cases} 0, & 0 \leq t \leq 2 \\ t^2, & 2 < t \leq 5 \\ e^t, & 5 < t \leq 7 \\ \cos(t), & t > 7 \end{cases} .$$

We can build $q(t)$ methodically using a philosophy similar to that of Example 5.18. Specifically, take

$$q(t) = (H(t - 2) - H(t - 5))t^2 + (H(t - 5) - H(t - 7))e^t + H(t - 7)\cos(t).$$

The key idea is to use the fact that if $a < b$ then $H(t - a) - H(t - b)$ equals 1 for $a < t \leq b$ and zero otherwise. ■

Reading Exercise 5.3.1 By considering t in each interval $(0, 2)$, $(2, 5)$, $(5, 7)$, and $(7, \infty)$, show that $q(t)$ in Example 5.19 works as advertised—each piece switches on or off at precisely the right time.

Back to Morphine Administration

Let's formulate the ODE (5.39) with $r(t)$ as in (5.40), using the Heaviside function. Following the flow of Example 5.19 we have

$$\begin{aligned} r(t) &= 1.5(H(t) - H(t - 12)) + 2.08H(t - 12) \\ &= 1.5 + 0.58H(t - 12). \end{aligned} \quad (5.41)$$

The last line in (5.41) follows from the line above it by using $H(t) = 1$ for $t > 0$.

With the Heaviside function at our disposal we can pose the ODE (5.39) as

$$u'(t) = -ku(t) + 1.5 + 0.58H(t - 12) \quad (5.42)$$

with initial condition $u(0) = 10$. Equation (5.42) can be solved using the Laplace transform, but first we need to know the Laplace transform of $H(t - c)$ for a constant c . You may already have done this in Exercise 5.2.8.

Laplace Transforming $H(t - c)$

We now compute the Laplace transform of $H(t - c)$ where $c \geq 0$. From the definition (5.6) or (5.7) of the Laplace transform we have

$$\begin{aligned} \mathcal{L}(H(t - c)) &= \int_0^\infty e^{-st} H(t - c) dt \\ &= \int_c^\infty e^{-st} H(t - c) dt \quad (\text{lower limit } c, \text{ since } H(t - c) = 0 \text{ for } t < c) \\ &= \int_c^\infty e^{-st} dt \quad (\text{since } H(t - c) = 1 \text{ for } t > c) \\ &= \frac{e^{-cs}}{s}. \quad (\text{routine improper integral, } s > 0) \end{aligned}$$

We have shown that

$$\mathcal{L}(H(t - c)) = \frac{e^{-cs}}{s} \quad (5.43)$$

for any $c \geq 0$. It's worth noting that the Laplace transform of $H(t)$ itself is the function $1/s$. From Table 5.1 this is the same transform as the constant function 1.

Reading Exercise 5.3.2 What is the Laplace transform of $H(t - c)$ if $c < 0$? Hint: recall Remark 5.2.1 in Section 5.2.

5.3.3 The Second Shifting Theorem

Equation (5.43) can also be expressed as

$$\mathcal{L}(H(t - c)) = e^{-cs} \mathcal{L}(H(t)),$$

where $c \geq 0$. As it turns out, a similar result holds if H is replaced by any function f .

Specifically, consider a function $f(t)$ defined for $t \geq 0$ as illustrated in the left panel of Figure 5.5. When $c > 0$ the graph of $f(t - c)$ is just the graph of $f(t)$ shifted c units to the right. The goal is to compute the Laplace transform of $f(t - c)$ for some $c > 0$. But since $f(t)$ is not defined for $t < 0$, $f(t - c)$ is not defined for $0 \leq t < c$. We need the function of interest to have some value in this region in order to compute the Laplace transform integral. We thus define $f(t - c)$ in the gap region $0 \leq t < c$ to be the zero function, which can be accomplished by using the product $H(t - c)f(t - c)$. See the right panel in Figure 5.5, in which we graph $H(t - c)f(t - c)$.

This sets the stage for the second shifting theorem:

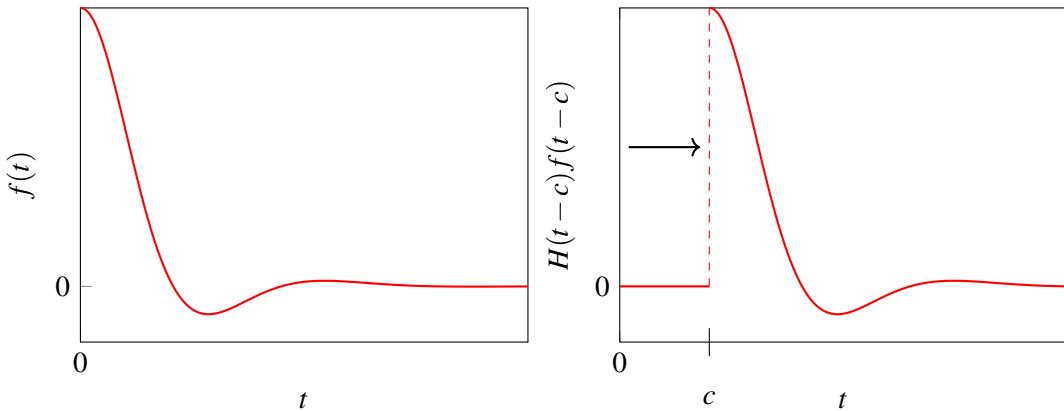


Figure 5.5: Left panel: graph of a typical (unspecified) function $f(t)$ defined for $t \geq 0$. Right panel: graph of $H(t - c)f(t - c)$ for some $c > 0$.

Theorem 5.3.1 — Second Shifting Theorem. If $f(t)$ is piecewise continuous for $t \geq 0$, of exponential order, and has Laplace transform $F(s)$ defined for $s > a$, and $c \geq 0$ is any constant, then

$$\mathcal{L}(H(t - c)f(t - c)) = e^{-cs}F(s),$$

and is defined for $s > a$.

Compare the first shifting theorem (Theorem 5.2.2) and the second shifting theorem. There is an interesting parallel in the time and s -domains. The function $e^{ct}f(t)$ in the time domain corresponds to $F(s - c)$ in the s -domain. The function $H(t - c)f(t - c)$ in the time domain corresponds to $e^{-cs}F(s)$ in the s -domain. Put another way, the operation of multiplication by e^{ct} in the time domain corresponds to a shift to the right by c units in the s -domain. The operation of multiplication by e^{-cs} in the s -domain corresponds to a shift to the right by c units in the time domain. The proof of the second shifting theorem is left as Exercise 5.3.11; it is essentially an integral calculus computation.

Examples

The second shifting theorem can be used for computing the Laplace transforms of piecewise defined functions.

■ **Example 5.20** Let $\phi(t)$ be defined as

$$\phi(t) = \begin{cases} 0, & t \leq 2 \\ t - 2, & t > 2 \end{cases}.$$

To compute $\mathcal{L}(\phi)$, first write ϕ in terms of Heaviside functions, as

$$\phi(t) = H(t - 2)(t - 2).$$

(Make sure you believe this.) Now $\phi(t)$ is perfectly set up for the second shifting theorem: define $f(t) = t$ and then $\phi(t)$ is exactly $H(t - 2)f(t - 2)$. The Laplace transform of $f(t)$ is $1/s^2$, so by the second shifting theorem the Laplace transform of $\phi(t)$ is e^{-2s}/s^2 . ■

Reading Exercise 5.3.3 Compute $\mathcal{L}(\phi)$, where

$$\phi(t) = \begin{cases} 0, & t \leq 3 \\ e^{t-3}, & t > 3. \end{cases}$$

■ **Example 5.21** Things can be a bit more subtle than Example 5.20. Here's a slight variation. Let $\phi(t)$ be defined as

$$\phi(t) = \begin{cases} 0, & t \leq 2 \\ t, & t > 2. \end{cases}$$

(The $t - 2$ in the previous example is now just t .) We will use Theorem 5.3.1 to compute $\mathcal{L}(\phi)$.

First write ϕ in terms of Heaviside functions as

$$\phi(t) = H(t - 2)t.$$

The trick now is to recognize $\phi(t)$ as $H(t - 2)f(t - 2)$ for some f . A bit of algebraic insight shows that taking $f(t) = t + 2$ works, for then $f(t - 2) = (t - 2) + 2 = t$. With this observation, we have $\phi(t) = H(t - 2)f(t - 2)$. The Laplace transform of $f(t) = t + 2$ is $F(s) = 1/s^2 + 2/s$, and so $\mathcal{L}(\phi) = e^{-2s}F(s)$ or

$$\mathcal{L}(\phi) = e^{-2s}(1/s^2 + 2/s).$$

■

Reading Exercise 5.3.4 Compute $\mathcal{L}(\phi)$ where $\phi(t) = H(t - 7)t^2$.

Example 5.21 and Reading Exercise 5.3.4 illustrate a variation on the second shifting theorem that is often useful and which we now summarize as a theorem.

Theorem 5.3.2 — Second Shifting Theorem II. Suppose $c \geq 0$ and $g(t)$ is piecewise continuous for $t \geq c$ and of exponential order. Let $f(t) = g(t + c)$. Then

$$\mathcal{L}(H(t - c)g(t)) = e^{-cs}F(s),$$

where $F(s) = \mathcal{L}(f(t))$.

The truth of this theorem follows from the observation that if $f(t) = g(t + c)$, then $g(t) = f(t - c)$ for $t \geq 0$, in which case Theorem 5.3.2 is exactly Theorem 5.3.1.

Let's do a slightly more involved example than any we've considered so far.

■ **Example 5.22** Let $\phi(t)$ be defined as

$$\phi(t) = \begin{cases} 3, & t \leq 2 \\ t, & 2 < t \leq 6 \\ 0, & t > 6 \end{cases}.$$

We will use the second shifting theorem 5.3.2 to compute $\mathcal{L}(\phi)$.

First write $\phi(t)$ in terms of Heaviside functions as

$$\phi(t) = 3(H(t) - H(t - 2)) + t(H(t - 2) - H(t - 6)).$$

It's helpful to collect similar $H(t - c)$ terms together; also note that $H(t) = 1$ for $t > 0$. Then

$$\phi(t) = 3 + H(t - 2)(t - 3) - H(t - 6)t. \quad (5.44)$$

We need to Laplace transform each piece on the right and invoke linearity.

The transform of the constant 3 on the right in (5.44) is $3/s$. For the second term $H(t - 2)(t - 3)$ on the right in (5.44) let's use Theorem 5.3.2 with $g(t) = t - 3$ and $c = 2$. In this case $f(t) = g(t + 2) = t - 1$ and then $F(s) = 1/s^2 - 1/s$, so the transform of this second term is $e^{-2s}(1/s^2 - 1/s)$. For the last term $(-t)H(t - 6)$ on the right in (5.44) we again use Theorem 5.3.2 but with $g(t) = -t$

and $c = 6$. In this case $f(t) = g(t+6) = -(t+6)$. Then $F(s) = -1/s^2 - 6/s$ and the transform of the last term is $e^{-6s}F(s) = -e^{-6s}/s^2 - 6e^{-6s}/s$. All in all, we obtain

$$\mathcal{L}(\phi) = \frac{3}{s} + \frac{e^{-2s}}{s^2} - \frac{e^{-2s}}{s} - \frac{e^{-6s}}{s^2} - \frac{6e^{-6s}}{s}.$$

■

Computing Inverse Transforms Using the Second Shifting Theorem

The second shifting theorem is useful for computing inverse Laplace transforms. A key takeaway here is that if you see a Laplace transform with an e^{-cs} in it, back in the time domain there are $H(t-c)$ functions lurking.

■ **Example 5.23** Let's compute the inverse Laplace transform of

$$P(s) = \frac{4e^{-2s}}{(s-3)^2 + 1}.$$

First, the presence of e^{-2s} means that $H(t-2)$ will figure into the answer. Begin by ignoring the e^{-2s} term and instead focus on $\frac{4}{(s-3)^2 + 1}$. Examination of the second-to-last line in Table 5.1 shows that $\frac{1}{(s-3)^2 + 1}$ corresponds to $e^{3t} \sin(t)$ in the time domain, so $\frac{4}{(s-3)^2 + 1}$ corresponds to the function $f(t) = 4e^{3t} \sin(t)$. From Theorem 5.3.1, the inverse Laplace transform of $P(s)$ is the function $H(t-2)f(t-2)$ or

$$p(t) = 4H(t-2)e^{3(t-2)} \sin(t-2).$$

■

Reading Exercise 5.3.5 Use the Theorem 5.3.2 to redo Reading Exercise 5.3.4.

Conclusion of the Morphine Administration Example

Let's return to the ODE (5.42), reproduced here:

$$u'(t) = -ku(t) + 1.5 + 0.58H(t-12) \quad (5.45)$$

with initial condition $u(0) = 10$; recall $k \approx 0.173$. You should remind yourself what the right side of (5.45) models: the rate at which morphine is being eliminated ($-ku(t)$) plus the rate at which morphine is being administered ($1.5 + 0.58H(t-12)$). We'll use the Laplace transform to reproduce the solution (5.4) in Section 5.1 that was obtained there by more piecemeal methods.

To begin, Laplace transform both sides of (5.45) and substitute in $u(0) = 10$ to obtain

$$sU(s) - 10 = -kU(s) + \frac{1.5}{s} + \frac{0.58e^{-12s}}{s}. \quad (5.46)$$

Solve (5.46) for $U(s)$ as

$$U(s) = \frac{10}{s+k} + \frac{1.5}{s(s+k)} + \frac{0.58e^{-12s}}{s(s+k)}. \quad (5.47)$$

We now inverse Laplace transform to obtain $u(t)$. The inverse transform of the first piece $10/(s+k)$ on the right in (5.47) is $10e^{-kt}$. The remaining two pieces on the right in (5.47) involve multiples of $\frac{1}{s(s+k)}$, so let's focus on inverse transforming this expression. A partial fraction expansion on $\frac{1}{s(s+k)}$ yields

$$\frac{1}{s(s+k)} = \frac{1}{ks} - \frac{1}{k(s+k)}.$$

It's convenient to name the inverse transform of this expression, say $\phi(t)$, and from Table 5.1 we find

$$\phi(t) = \frac{1}{k} - \frac{e^{-kt}}{k}.$$

The second term on the right in (5.47) has inverse transform $1.5\phi(t)$. Finally, the last piece on the right in (5.47) corresponds to $0.58H(t - 12)\phi(t - 12)$, where we use the second shifting theorem.

All in all we obtain the solution

$$\begin{aligned} u(t) &= 10e^{-kt} + 1.5\phi(t) + 0.58H(t - 12)\phi(t - 12) \\ &= 10e^{-kt} + \frac{1.5}{k} - \frac{1.5e^{-kt}}{k} + 0.58H(t - 12) \left(\frac{1}{k} - \frac{e^{-k(t-12)}}{k} \right). \end{aligned}$$

After simplifying separately on the intervals $0 < t < 12$ and $t > 12$, this is the same piecewise function defined in (5.4).

Reading Exercise 5.3.6 Solve the ODE $u'(t) = -u(t) + H(t - 1)$ with initial condition $u(0) = 1$.

5.3.4 Some More Models and Examples

Let's look at two more models from start to finish that involve the use of the Heaviside function, and how the Laplace transform aids the solution process.

■ **Example 5.24** Suppose an object has temperature $u(t)$ (degrees Fahrenheit) at time t (minutes) with $u(0) = 50$ and is in an environment with ambient temperature $A = 80$ degrees Fahrenheit. The object obeys Newton's law of cooling $u'(t) = -k(u(t) - A)$ with cooling constant $k = 0.1$ (units of reciprocal minutes). At time $t = 20$ minutes the object is moved to an environment with temperature $A = 30$ degrees, and continues to obey Newton's law of cooling with the same cooling constant $k = 0.1$. Let us find the temperature $u(t)$ of the object as a function of time.

First write the ODE $u'(t) = -k(u(t) - A)$ as $u'(t) = -ku(t) + kA$, where A is the ambient temperature that changes abruptly at time $t = 20$. We can express A as

$$A = 80 - 50H(t - 20),$$

and so the appropriate ODE here is

$$u'(t) = -0.1u(t) + 8 - 5H(t - 20) \quad (5.48)$$

with initial condition $u(0) = 50$.

To solve (5.48), Laplace transform both sides of the ODE and substitute in $u(0) = 50$ to obtain

$$sU(s) - 50 = -0.1U(s) + 8/s - 5e^{-20s}/s$$

where $U(s) = \mathcal{L}(u(t))$, then solve for $U(s)$ as

$$U(s) = \frac{50 + 8/s - 5e^{-20s}/s}{s + 0.1} = \frac{50s + 8}{s(s + 0.1)} - \frac{5e^{-20s}}{s(s + 0.1)}. \quad (5.49)$$

Each term on the right in (5.49) has denominator $s(s + 0.1)$, so it's helpful to consider a partial fraction decomposition of the form

$$\frac{A}{s} + \frac{B}{s + 0.1} = \frac{(A + B)s + 0.1A}{s(s + 0.1)}. \quad (5.50)$$

We can obtain the first term $\frac{50s+8}{s(s+0.1)}$ on the right in (5.49) by requiring $A + B = 50$ and $0.1A = 8$, which leads to $A = 80$ and $B = -30$. Substituting these values into (5.50) shows that

$$\frac{50s+8}{s(s+0.1)} = \frac{80}{s} - \frac{30}{s+0.1}.$$

The inverse Laplace transform of the expression on the right above is $80 - 30e^{-0.1t}$. This provides the inverse transform of the first term on the right side of (5.49).

For the second term on the right in (5.49), first ignore the e^{-20s} factor and seek a partial fraction decomposition of the form (5.50) for $\frac{5}{s(s+0.1)}$. This requires $A + B = 0$ and $0.1A = 5$, so $A = 50$ and $B = -50$, and so we have

$$\frac{5}{s(s+0.1)} = \frac{50}{s} - \frac{50}{s+0.1}.$$

The inverse transform of this quantity is the function $\phi(t) = 50 - 50e^{-0.1t}$, and so by Theorem 5.3.1, the inverse transform of the second term on the right in (5.49) is $H(t-20)\phi(t-20)$. All in all the inverse Laplace transform of $U(s)$ is

$$\begin{aligned} u(t) &= 80 - 30e^{-0.1t} - H(t-20)\phi(t-20) \\ &= 80 - 30e^{-0.1t} - 50H(t-20)(1 - e^{-0.1(t-20)}). \end{aligned}$$

This is the solution to the ODE (5.48), and $u(t)$ is graphed in Figure 5.6. ■

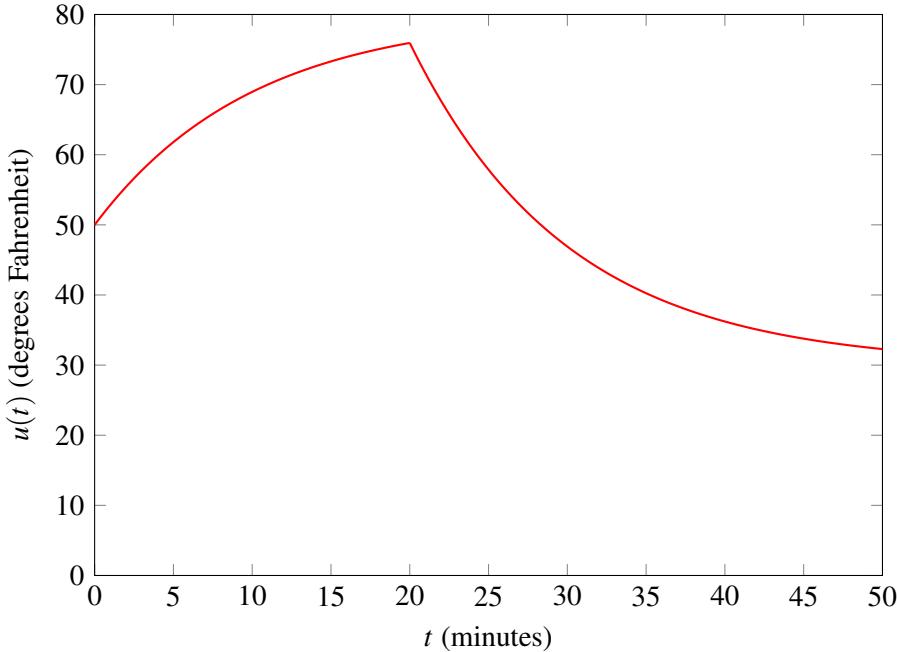


Figure 5.6: The temperature $u(t)$ of an object as modeled by (5.48), initial temperature $u(0) = 50$, ambient temperature $A = 80$ degrees Fahrenheit for $t \leq 20$, then $A = 30$ degrees Fahrenheit for $t > 20$.

Let's next analyze a second-order spring-mass-damper example, from start to finish, with discontinuous forcing.

■ **Example 5.25** A spring-mass-damper system has mass $m = 1$ kg, damping of $c = 2$ newtons per meter per second, and $k = 10$ newtons per meter. The mass is at rest at equilibrium at time $t = 0$

and no external forces act on the mass until time $t = 3$ seconds, at which time a constant force of 10 newtons acts on the mass. We seek the position $u(t)$ of the mass, displacement measured in meters.

The appropriate ODE here is

$$u''(t) + 2u'(t) + 10u(t) = 10H(t - 3), \quad (5.51)$$

with initial conditions $u(0) = u'(0) = 0$. To solve, take the Laplace transform of both sides of (5.51) and fill in the (zero) initial conditions to obtain

$$s^2U(s) + 2sU(s) + 10U(s) = \frac{10e^{-3s}}{s}.$$

Solve for $U(s)$ to find that

$$U(s) = \frac{10e^{-3s}}{s(s^2 + 2s + 10)}. \quad (5.52)$$

To find $u(t)$ the function $U(s)$ must be inverse Laplace transformed.

We focus first on inverse transforming the expression $\frac{10}{s(s^2 + 2s + 10)}$. A partial fraction decomposition is needed; the $s^2 + 2s + 10$ piece is irreducible, so try

$$\frac{10}{s(s^2 + 2s + 10)} = \frac{A}{s} + \frac{Bs + C}{s^2 + 2s + 10} = \frac{(A + B)s^2 + (2A + C)s + 10A}{s(s^2 + 2s + 10)}.$$

Matching coefficients in the various powers of s on the left and right above yields three equations:

$$A + B = 0, \quad 2A + C = 0, \quad 10A = 10.$$

The solution is $A = 1$, $B = -1$, and $C = -2$, so

$$\frac{10}{s(s^2 + 2s + 10)} = \frac{1}{s} - \frac{s + 2}{s^2 + 2s + 10}. \quad (5.53)$$

Next we need to inverse transform the expression on the right in (5.53).

The inverse transform of the first term $1/s$ on the right in (5.53) is just the constant function 1. The inverse transform of the rightmost term in (5.53) can be found by completing the square and then (as we've done before) creatively grouping terms as

$$\frac{s+2}{s^2+2s+10} = \frac{s+2}{(s+1)^2+3^2} = \frac{s+1}{(s+1)^2+3^2} + \frac{1}{(s+1)^2+3^2}. \quad (5.54)$$

The right side above is in a form whose inverse Laplace transform can be read off of the bottom two lines in Table 5.1. If we let $\phi(t)$ denote the inverse transform of $\frac{10}{s(s^2+2s+10)}$ then (5.53) and (5.54) show that

$$\phi(t) = 1 - e^{-t} \cos(3t) - \frac{1}{3}e^{-t} \sin(3t).$$

Based on (5.52) and Theorem 5.3.1, the solution to (5.51) is

$$\begin{aligned} u(t) &= H(t - 3)\phi(t - 3) \\ &= H(t - 3)(1 - e^{-(t-3)}(\cos(3(t - 3)) + \sin(3(t - 3))/3)). \end{aligned}$$

This function is graphed in Figure 5.7. ■

Reading Exercise 5.3.7 Explain why the solution to (5.51) graphed in Figure 5.7 makes perfect sense in view of the applied force $10H(t - 3)$ and initial conditions. In particular, consider $0 < t < 3$ and $t \rightarrow \infty$.

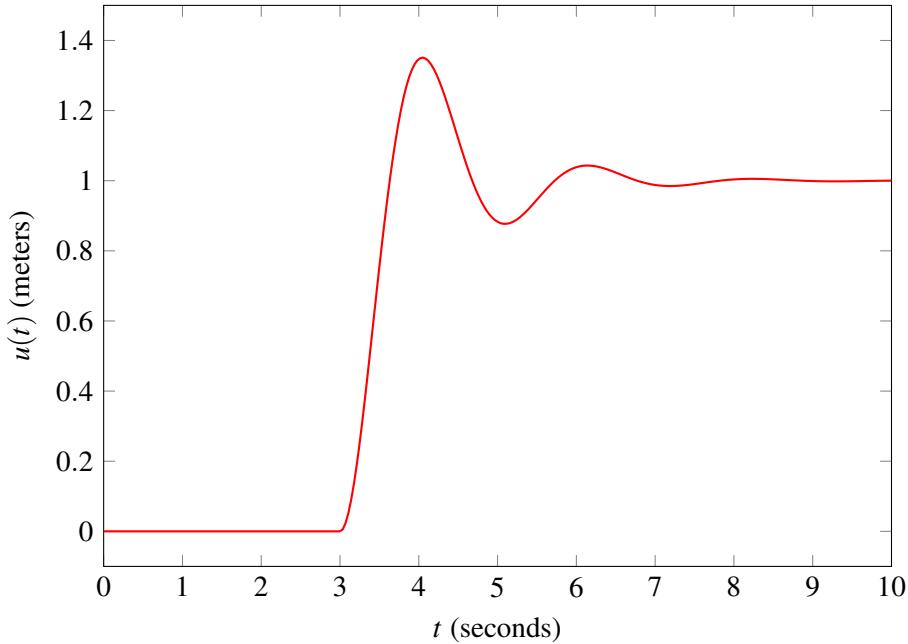


Figure 5.7: Motion of a spring-mass-damper system as modeled by (5.51), initially at rest and at equilibrium, no external forcing for $t \leq 3$, then a constant force of 10 newtons applied for $t > 3$.

5.3.5 Summary and Remarks

The examples of this section demonstrate that by using the Laplace transform we can handle piecewise discontinuous forcing functions in a unified framework, along with the smoother forcing functions like the exponentials, polynomials, sines, and cosines that were tackled in Section 4.3 using undetermined coefficients. This advantage, however, is primarily theoretical at the moment. When it is time to actually solve ODEs with discontinuous forcing functions manually, the Laplace transform computations are probably just as tedious and time consuming as breaking the problem into subintervals and solving each individually. As the problems we tackle become more complicated, we will use a computer algebra system to facilitate these of computations.

Nevertheless, it is essential to understand how to use the Heaviside function to model discontinuous phenomena, and it pays to understand in some detail how the Laplace transform is used to solve ODEs. This knowledge builds a necessary intuition about these tools so that later when we see expressions like $s(s + 2)$ or $s^2 + 2s + 8$ in the denominator of a Laplace transform, we have some idea about what to expect back in the time domain and what to expect for the system's behavior. This intuition will pay off later when we begin looking at convolution, transfer functions, and similar issues. Indeed, in many situations like those of Section 5.6 we take a time domain problem to the s -domain and never go back.

5.3.6 Exercises

Exercise 5.3.1 Express each of the piecewise continuous functions $f(t)$ below by using an appropriate combination of Heaviside functions. Assume f has domain $t \geq 0$. The precise value of f at any discontinuity doesn't matter.

(a)

$$f(t) = \begin{cases} 0, & 0 \leq t \leq 5 \\ 7, & t > 5 \end{cases}$$

(b)

$$f(t) = \begin{cases} 7, & 0 \leq t \leq 5 \\ 0, & t > 5 \end{cases}$$

(c)

$$f(t) = \begin{cases} 2, & 0 \leq t \leq 3 \\ 5, & 3 < t \leq 6 \\ -3, & t > 6 \end{cases}$$

(d)

$$f(t) = \begin{cases} 1, & 0 \leq t \leq 3 \\ t, & 3 < t \leq 6 \\ e^{-t}, & t > 6 \end{cases}$$

(e)

$$f(t) = \begin{cases} 0, & 0 \leq t \leq 3 \\ e^t, & 3 < t \leq 6 \\ e^{2t}, & 6 < t \leq 10 \\ 4, & t > 10 \end{cases}$$

Exercise 5.3.2 Compute the Laplace transform of each function in Exercise 5.3.1.

Exercise 5.3.3 Compute the inverse Laplace transform of each of the following expressions.

$$(a) F(s) = \frac{2e^{-3s}}{s^2}$$

$$(b) Q(s) = \frac{e^{-s}}{s^2 + 16}$$

$$(c) G(s) = \frac{(3s+2)e^{-5s}}{s^2 + 4}$$

$$(d) F(s) = \frac{e^{-2\pi s}s}{s^2 + 6s + 25}$$

$$(e) F(s) = \frac{12e^{-3s}}{(s+2)^4}$$

Exercise 5.3.4 Solve the following first-order ODEs using the method of Laplace transforms and plot the solution on the interval $0 \leq t \leq 10$.

$$(a) u'(t) = -2u(t) + 4H(t-5) \text{ with } u(0) = 1$$

$$(b) u'(t) = -3u(t) + 3H(t-3) - 6H(t-5) \text{ with } u(0) = 1$$

$$(c) u'(t) = -u(t) + tH(t-1) - H(t-2) \text{ with } u(0) = 2$$

Exercise 5.3.5 Solve the following second-order ODEs using the method of Laplace transforms and plot the solution on the interval $0 \leq t \leq 10$.

- (a) $u''(t) + 4u'(t) + 3u(t) = H(t - 1)$ with $u(0) = u'(0) = 0$
- (b) $u''(t) + 16u(t) = H(t - 3)$ with $u(0) = u'(0) = 0$
- (c) $u''(t) + 4u'(t) + 4u(t) = 4 + 8H(t - 3)$ with $u(0) = 1$ and $u'(0) = 2$
- (d) $u''(t) + 16u(t) = H(t - \pi) \cos(4t)$ with $u(0) = 0$ and $u'(0) = -1$. Hint: the result of Exercise 5.2.10 may be useful, with $F(s) = 1/(s^2 + 16)$.

Exercises 5.3.6 to 5.3.10 are variations on (or identical to) Exercises 5.1.1-5.1.5.

Exercise 5.3.6 A patient is given a 5 mg bolus of morphine at time $t = 0$, followed by an infusion at a rate of $r(t) = 1$ mg of morphine per hour. From $t = 12$ to $t = 48$ hours the infusion rate is increased to $r(t) = 1.5$ mg per hour. Assume the drug amount is governed by (5.2). Formulate an appropriate ODE and initial condition and solve using the method of Laplace transforms. Plot the solution on the interval $0 \leq t \leq 48$.

Exercise 5.3.7 A bank account is opened with \$1000 at time $t = 0$ years. The account pays interest at an annual rate of 2 percent, compounded continuously; that is, the account accrues interest at a rate of $0.02p(t)$. Suppose the deposit rate is $r(t) = 520$ dollars per year from time $t = 0$ to time $t = 2$, but then drops to $r(t) = 200$ dollars per year for time $t \geq 2$. Formulate an appropriate ODE with an initial condition, and solve using the Laplace transform. Plot the solution for time $0 \leq t \leq 10$.

Exercise 5.3.8 An object in an environment with ambient temperature $A = 80$ degrees obeys Newton's law of cooling (2.14) with cooling constant $k = 0.05$. The object has temperature 120 degrees at time $t = 0$. At time $t = 50$ the object is moved to an environment with ambient temperature $A = 90$ degrees. The object still obeys Newton's law of cooling with the same cooling constant $k = 0.05$. Formulate an appropriate ODE with an initial condition and solve using the Laplace transform. Plot the object's temperature for $0 \leq t \leq 100$.

Exercise 5.3.9 An undamped spring-mass-damper system with mass $m = 2$ kg and spring constant $k = 8$ newtons per meter is at equilibrium position $u = 0$ and is not moving at time $t = 0$. No additional forces act on the mass until time $t = 10$ seconds, but for $t > 10$ a force $f(t) = 40$ newtons is applied to the mass. At time $t = 15$ the force drops to zero. Find the position of the mass for $t > 0$ by formulating an appropriate ODE with an initial conditions and solving with the Laplace transform. Plot the solution on the interval $0 \leq t \leq 25$.

Exercise 5.3.10 Consider an RC circuit like that shown in Figure 2.2, with resistor $R = 10$ ohms and capacitor $C = 10^{-4}$ F. The capacitor is uncharged at time $t = 0$. Suppose the voltage source is $V(t) = 2$ volts for time $0 \leq t \leq 0.003$ seconds and then switches to $V(t) = 5$ volts for $t > 0.003$. Formulate an appropriate ODE for the charge $q(t)$ on the capacitor, with initial condition, and solve using the Laplace transform. Plot the solution on the interval $0 \leq t \leq 0.01$.

Exercise 5.3.11 The proof of the second shifting theorem (Theorem 5.3.1) is a straightforward integral calculus computation.

- (a) From the definition of the Laplace transform we have

$$\mathcal{L}(H(t-c)f(t-c)) = \int_0^\infty e^{-st} H(t-c)f(t-c) dt.$$

Argue that this leads to

$$\mathcal{L}(H(t-c)f(t-c)) = \int_c^\infty e^{-st} f(t-c) dt. \quad (5.55)$$

Hint: what is the value of $H(t-c)$ for $0 < t < c$? What is the value of $H(t-c)$ for $t > c$?

- (b) Make a substitution $w = t - c$ (so $t = w + c$ and $dt = dw$) in (5.55). Don't forget that the limits of integration change too. Show that the new integral yields

$$\mathcal{L}(H(t-c)f(t-c)) = e^{-cs} \int_0^\infty e^{-ws} f(w) dw.$$

Why does this prove Theorem 5.3.1?

Exercise 5.3.12 Flesh out the details necessary to prove Theorem 5.3.2 based on Theorem 5.3.1.

5.4 The Dirac Delta Function

The linear, constant-coefficient ODEs we've considered in this chapter involved forcing functions that are piecewise continuous (this includes continuous functions, of course) and of exponential order. These types of functions were discussed at some length in Section 5.2, since they are guaranteed to have meaningful Laplace transforms. Now we're going to break those rules. But we'll do this with some care, and for a good reason: it will facilitate modeling impulsive phenomena in the Laplace transform framework we've been developing. These phenomena would be cumbersome to analyze with a more traditional approach. The essential mathematical object of interest in this section is the Dirac delta function, popularized by physicist Paul Dirac in his work on quantum mechanics, though the ideas go back much further. The Dirac delta function isn't a function at all in the conventional sense. Mathematicians would call it a *distribution* or a *generalized function*, or a *measure*.

Reading Exercise 5.4.1 Have you been paying attention? Is the Dirac delta function a function?

5.4.1 Motivational Examples

In the morphine dosing example of Section 5.1 we discussed the possibility of modeling the administration of a 5 mg bolus of morphine to a patient at precisely time $t = 12$ hours, by using the infusion rate $r(t)$ given by (5.5). This is our first example of an impulsive forcing function in an ODE: the delivery of a finite amount of something at a very high rate over a very short time period. This is the type of phenomenon that can be modeled using a Dirac delta function.

Let's consider two additional examples that illustrate the need for what the Dirac delta function provides.

Instantaneous Deposits in a Bank Account

A savings or investment account is opened with a \$10,000 balance at time $t = 0$. The account earns interest at a constant rate of 2 percent annually, compounded continuously. This means that in the absence of any additional deposits the balance $p(t)$ would grow according to $p'(t) = 0.02p(t)$ with

initial condition $p(0) = 10000$. However, if deposits are made continuously at a rate of $r(t)$ dollars per year then the balance obeys

$$p'(t) = 0.02p(t) + r(t). \quad (5.56)$$

If the deposit rate $r(t)$ is piecewise continuous then (5.56) can be solved using Laplace transforms, if we can compute the appropriate Laplace and inverse transforms.

But what if instead of continuous deposits we make deposits in lump sums? Suppose the only deposit after the initial \$10,000 is a lump sum of \$5,000 at time $t = 3$ years. Can this be accommodated in the model (5.56)? We could define $r(t) = 5000$ for $2.5 \leq t \leq 3.5$ and $r(t) = 0$ elsewhere, which models \$5,000 deposited continuously over the course of a year. This really isn't an instantaneous lump sum deposit, but such an $r(t)$ would be piecewise constant and easy to handle with Laplace transforms. Or we could try to simulate the situation a bit more accurately, say take $r(t) = 50000$ for $t = 2.95$ to $t = 3.05$ (still a total deposit of \$5,000). This piecewise constant $r(t)$ could be tackled with Laplace transforms, but again this isn't an instantaneous deposit.

More generally, we could set $r(t) = 5000/(2\epsilon)$ dollars per year for 2ϵ years, from $t = 3 - \epsilon$ to $t = 3 + \epsilon$ with $\epsilon > 0$. This corresponds to a total deposit of $\frac{5000}{2\epsilon} \frac{\text{dollars}}{\text{year}} \times 2\epsilon \text{ years} = \$5,000$, and the smaller we take ϵ , the more closely this models an instantaneous deposit. In this case the deposit rate is

$$r(t) = \frac{5000}{2\epsilon}(H(t - 3 + \epsilon) - H(t - 3 - \epsilon)). \quad (5.57)$$

For future reference, the fact that $\frac{5000}{2\epsilon} \frac{\text{dollars}}{\text{year}} \times 2\epsilon \text{ years} = \$5,000$ can also be expressed by noting that

$$\int_0^\infty r(t) dt = 5000 \quad (5.58)$$

for any $\epsilon > 0$, since this integral computes the area under the graph of $r(t)$, a rectangle with height $5,000/(2\epsilon)$ and base width 2ϵ . With this $r(t)$ as in (5.57) the ODE (5.56) becomes

$$p'(t) = 0.02p(t) + \frac{5000}{2\epsilon}(H(t - 3 + \epsilon) - H(t - 3 - \epsilon)) \quad (5.59)$$

with initial condition $p(0) = 10000$. The solution to this ODE, obtained via Laplace transforms, is plotted in Figure 5.8 for $\epsilon = 0.5$, $\epsilon = 0.25$, and $\epsilon = 0.05$.

It's pretty clear that as ϵ approaches zero the solution stabilizes on some underlying function that has a jump discontinuity at $t = 3$. It would be nice to figure out what this function is, without the annoyance of dealing with ϵ or getting hung up on how long it took to deposit \$5,000. The Dirac delta function allows us to do this.

A Hammer Blow to a Spring-Mass System

Consider a mass-spring-damper system with $m = 1$ kg, $c = 2$ newtons per meter per second, and $k = 10$ newtons per meter. The mass is at equilibrium at time $t = 0$, and at rest. Of course, nothing will happen unless an external force is applied. That external force comes in the form of a hammer blow at time $t = 1$. A very large force is applied to the mass for a very short time, a fairly common type of force encountered in physics.

Consider, for example, a force $f(t)$ of 100 newtons applied for a time interval of $1/10$ of a second, from time $t = 0.95$ to time $t = 1.05$ seconds. The product of the force times the interval of duration is called the total **impulse** of the blow, which in this case is $100 \text{ newtons} \times 0.1 \text{ seconds} = 10 \text{ newton-seconds}$. This has the same dimension as momentum and as we shall see it is the total momentum imparted to the mass by the blow. The relevant ODE here is

$$u''(t) + 2u'(t) + 10u(t) = 100(H(t - 0.95) - H(t - 1.05)),$$

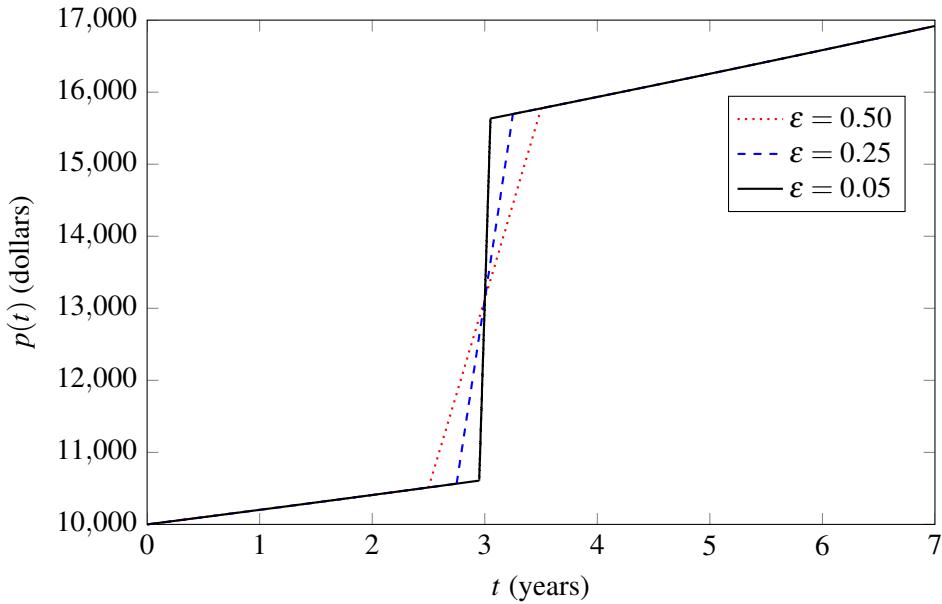


Figure 5.8: Graph of the solution to (5.59) modeling a bank account balance with initial \$10,000 balance at time $t = 0$ and \$5,000 deposit near time $t = 3$, made over a time interval $3 - \epsilon < t < 3 + \epsilon$, with $\epsilon = 0.5$ (dotted red), $\epsilon = 0.25$ (dashed blue), and $\epsilon = 0.05$ (solid blue).

with $u(0) = u'(0) = 0$. The solution is easily obtained using Laplace transforms and is shown in the left panel of Figure 5.9.

But how do we know the hammer blow lasted 0.1 seconds? The system response to a sharper blow of magnitude 1000 newtons for $t = 0.995$ to $t = 1.005$ seconds (also with total impulse $1000 \text{ newtons} \times 0.01 \text{ seconds} = 10 \text{ newton-seconds}$) is shown as the solid black curve in the right panel of Figure 5.9, which is barely distinguishable from the response in the left panel. Even the response to a mushy hammer blow of magnitude 20 newtons from $t = 0.75$ to $t = 1.25$ seconds (again, a total impulse of 10 newton-seconds), shown as the dashed red curve in the right panel of Figure 5.9, doesn't look much different. If we want to model a hammer blow it seems the short duration of the impact is not relevant; the total impulse is what matters.

It makes sense to model a hammer blow with a total impulse of 10 newton-seconds and brief duration from time $t = t_0 - \epsilon$ to $t = t_0 + \epsilon$ as a constant force of $10/(2\epsilon)$ newtons over the time interval of length 2ϵ seconds, where $\epsilon > 0$. This kind of force can be written in terms of Heaviside functions as

$$f(t) = \frac{10}{2\epsilon}(H(t - t_0 + \epsilon) - H(t - t_0 - \epsilon)). \quad (5.60)$$

The area under the graph of $f(t)$ is a rectangle of base width 2ϵ and height $\frac{10}{2\epsilon}$, so the fact that the total impulse is 10 newton-seconds can also be expressed as

$$\int_0^\infty f(t) dt = 10 \quad (5.61)$$

for any $\epsilon > 0$, in exactly the same way that the integral (5.58) expressed a deposit of \$5,000. The appropriate model for the motion of the spring-mass system is then

$$u''(t) + 2u'(t) + 10u(t) = f(t). \quad (5.62)$$

A truly instantaneous hammer blow would correspond to $\epsilon = 0$, a duration of zero seconds, but of course this would mean the applied force is infinite, in just the right way to make the impulse

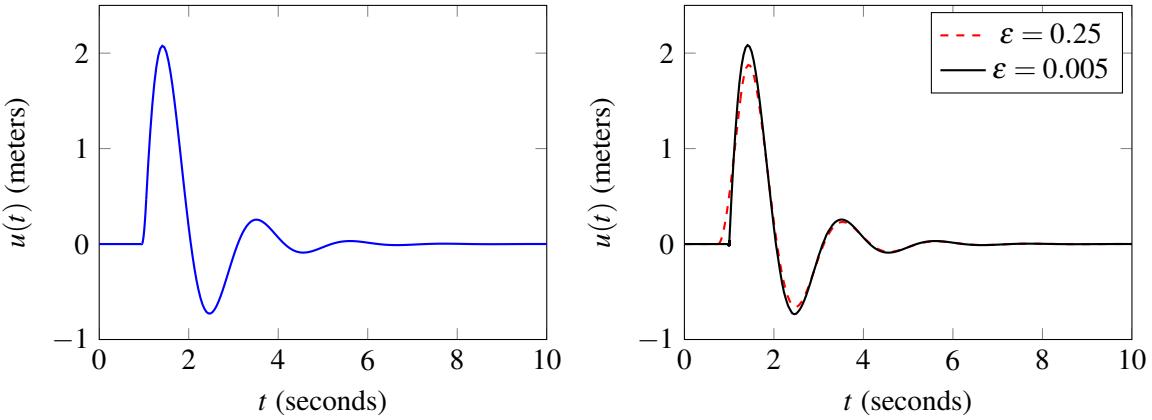


Figure 5.9: Left panel: graph of solution to (5.62) with forcing function (5.60) modeling a spring-mass system with hammer blow lasting 0.1 seconds ($\varepsilon = 0.05$), total impulse 10 newton-seconds. Right panel: Same, but with hammer blow duration 0.01 seconds ($\varepsilon = 0.005$, solid black) and duration 0.5 seconds ($\varepsilon = 0.25$, dashed red).

integral 10 newton-seconds. This all sounds fairly suspect, but the Dirac delta function lets us make this approach rigorous.

5.4.2 Definition of the Dirac Delta Function

We'll start with an intuitive view of the Dirac delta function, then put it on a firmer foundation.

Intuition

Look back at $r(t)$ in (5.57) and $f(t)$ in (5.60). Each is a function of the form

$$\phi_\varepsilon(t) = \frac{A}{2\varepsilon}(H(t - t_0 + \varepsilon) - H(t - t_0 - \varepsilon)) \quad (5.63)$$

for some constant A and time t_0 , where the subscript ε on ϕ_ε explicitly indicates the dependence of the right side of (5.63) on the parameter $\varepsilon > 0$. The parameter A in (5.63) quantifies the total impulse of the event, which is finite, e.g., the total deposit in the case of an interest computation, or the total impulse of a hammer blow.

In Figure 5.10 the function $\phi_\varepsilon(t)$ is graphed with $t_0 = 1$, $A = 2$, and each of $\varepsilon = 0.5, 0.2, 0.1$, and 0.05 . In each case the area under the graph is a rectangle of base width 2ε and height $\frac{A}{2\varepsilon}$. Note that in each case the total impulse of the event is the area under the graph of $\phi_\varepsilon(t)$. In this example that area is 2, but more generally one can compute that the area under the curve is

$$\int_a^b \phi_\varepsilon(t) dt = \int_{t_0-\varepsilon}^{t_0+\varepsilon} \frac{A}{2\varepsilon} dt = (2\varepsilon) \times \left(\frac{A}{2\varepsilon} \right) = A, \quad (5.64)$$

as long as the limits of integration a and b on the left in (5.64) satisfy $a \leq t_0 - \varepsilon$ and $b \geq t_0 + \varepsilon$ (so the limits encompass the entire base interval on which $\phi_\varepsilon(t) > 0$). For example, in (5.58) the total impulse was $A = 5000$ dollars, while in (5.61) the impulse was $A = 10$ newton-seconds. Limits $a = -\infty$ to $b = \infty$ in (5.64) are common.

Reading Exercise 5.4.2 Argue that if $a < b < t_0 - \varepsilon$ or $t_0 + \varepsilon < a < b$ in (5.64) then the integral in (5.64) equals zero.

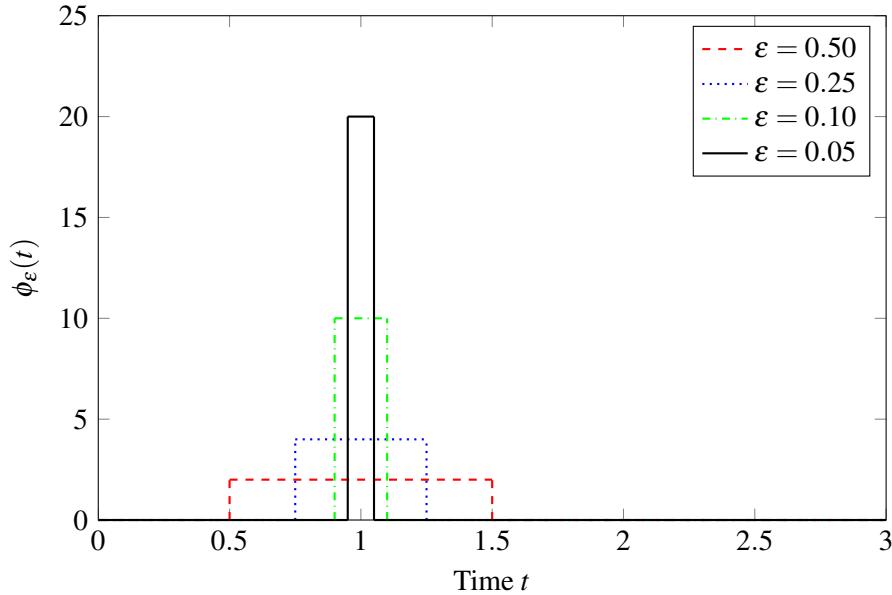


Figure 5.10: Graph of $\phi_\epsilon(t)$ in (5.63) to approximate an impulsive event with duration $1 - \epsilon < t < 1 + \epsilon$, total impulse $A = 2$, with $\epsilon = 0.5$ (dashed red), $\epsilon = 0.25$ (dotted blue), $\epsilon = 0.1$ (green), $\epsilon = 0.05$ (solid black).

The Dirac Delta Function: Initial Definition

Let's focus on the simple base case in which $t_0 = 0$ and $A = 1$, an impulsive event at time $t = 0$, of total impulse 1. In this case

$$\phi_\epsilon(t) = \frac{H(t + \epsilon) - H(t - \epsilon)}{2\epsilon}.$$

The right side is a difference quotient one would use in Calculus 1 to compute $H'(t)$ by letting $\epsilon \rightarrow 0$. Moreover, $\epsilon \rightarrow 0$ is exactly how we are modeling instantaneous impulses. This leads to the **Dirac delta function**, which might be defined as

$$\begin{aligned} \delta(t) &= \lim_{\epsilon \rightarrow 0^+} \phi_\epsilon(t) \\ &= \lim_{\epsilon \rightarrow 0^+} \left(\frac{H(t + \epsilon) - H(t - \epsilon)}{2\epsilon} \right) \\ &= H'(t). \end{aligned} \tag{5.65}$$

You may notice one drawback to this definition: it's absolute nonsense.

The limit in (5.65) is an intuitive take on the Dirac delta function and can be visualized in a manner similar to Figure 5.10, a sort of infinite spike, except in this case at $t = 0$ and with an area of 1 under its graph. This is not mathematically rigorous though, since for any $t \neq 0$ you can check that the limit in (5.65) is 0, while if $t = 0$ the limit does not exist. That is, $\delta(t) = 0$ for $t \neq 0$, while $\delta(0)$ is undefined. The function $\delta(t)$ defined by (5.65) makes no sense, and yet the process that led to it, as in the examples involving the instantaneous deposit of money into a bank account and the hammer blow to a spring-mass system, is clearly of value.

How do we reconcile the need for $\delta(t)$ with its slippery definition?

A More Careful Approach to the Dirac Delta Function

The integration in (5.64) is the key to handling impulsive events with greater mathematical rigor. Let's first consider the simplest case: how to interpret $\delta(t)$:

- Whenever $\delta(t)$ appears in a computation, it will at some point be integrated. The integration may not occur immediately—it might even be implicit—but it's there, often in the form of a Laplace transform.
- When $\delta(t)$ is integrated, we will interpret the integral by using (5.64) and taking a limit $\varepsilon \rightarrow 0$. For example, in (5.64) with $A = 1$ and $t_0 = 0$ we find that if $a < 0$ and $b > 0$ then

$$\int_a^b \delta(t) dt = \lim_{\varepsilon \rightarrow 0^+} \int_{-\varepsilon}^{\varepsilon} \frac{1}{2\varepsilon} dt = \lim_{\varepsilon \rightarrow 0^+} 1 = 1.$$

More examples of this type of reasoning, and generalizations, are shown below.

Reading Exercise 5.4.3 What is $\int_a^b \delta(t) dt$ if $0 < a < b$ or $a < b < 0$? Hint: consult Reading Exercise 5.4.2.

The General Dirac Delta Function

Impulsive forcing may occur at time $t = t_0$ instead of $t = 0$, so it's appropriate to consider $\delta(t - t_0)$. Moreover, an impulse may have magnitude A instead of 1 ($A < 0$ is allowed.) so we'll want to consider $A\delta(t - t_0)$. We handle this more general Delta function as before, with integration, but the Delta function will also sometimes appear under an integral with another function $g(t)$. This will occur, for example, when we Laplace transform $A\delta(t - t_0)$. So let's make sense of an integral like

$$\int_a^b A\delta(t - t_0)g(t) dt. \quad (5.66)$$

We will assume that g is continuous, at least near the point $t = t_0$. What value should we assign to this integral? We'll take the same approach we used for $\delta(t)$, which motivates the following definition.

Definition 5.4.1 The value of the integral in (5.66) is defined to be

$$\int_a^b A\delta(t - t_0)g(t) dt = \lim_{\varepsilon \rightarrow 0} \int_a^b \frac{A}{2\varepsilon} (H(t - t_0 + \varepsilon) - H(t - t_0 - \varepsilon))g(t) dt, \quad (5.67)$$

if this limit exists.

Equation (5.67) provides a more rigorous definition of the **Dirac delta function**, by specifying exactly how it should be handled when it appears in an integral.

We can compute the limit of the integral on the right in (5.67) explicitly. It may be helpful to refer to Figure 5.11, where we assume that $a < t_0$ and $b > t_0$ (that is, the interval of integration contains the time at which the impulsive event occurs).

Theorem 5.4.1 If $g(t)$ is continuous and $a < t_0 < b$ then

$$\lim_{\varepsilon \rightarrow 0} \int_a^b \frac{A}{2\varepsilon} (H(t - t_0 + \varepsilon) - H(t - t_0 - \varepsilon))g(t) dt = Ag(t_0). \quad (5.68)$$

As a result of Theorem 5.4.1, we set

$$\int_a^b A\delta(t - t_0)g(t) dt = Ag(t_0) \quad (5.69)$$

whenever we encounter a Dirac delta function under an integral, provided $a < t_0 < b$ and g is continuous at $t = t_0$. Equation (5.69) is sometimes called the **sifting property** of the Dirac delta function. The integration of $g(t)$ against $A\delta(t - t_0)$ sifts out the value of g at $t = t_0$, multiplied by A . We defer the proof to look at some examples.

■ **Example 5.26** Let's compute

$$\int_1^3 \delta(t-2)e^t \sin(t) dt.$$

A straightforward application of (5.69) with $t_0 = 2$, $A = 1$, $a = 1$, and $b = 3$ shows that this integral equals $e^2 \sin(2)$. ■

Reading Exercise 5.4.4 Take $t_0 = 2$ and $g(t) = t^2 + t$ (or any other continuous function you like that has a simple antiderivative). Compute the integral

$$\int_0^5 \frac{1}{2\epsilon} (H(t-2+\epsilon) - H(t-2-\epsilon)) g(t) dt$$

as a function of ϵ (we're using $A = 1$, $a = 0$, and $b = 5$). Then take the limit as $\epsilon \rightarrow 0^+$. Compare your answer to $g(2)$.

Proof of Theorem 5.4.1

To see why Theorem 5.4.1 is true, note that, as illustrated in Figure 5.11, if $a < t_0 < b$ then once ϵ is sufficiently small we have $a < t_0 - \epsilon < t_0 + \epsilon < b$. In this case the limits in the integral on the left in (5.68) can be taken as $t = t_0 - \epsilon$ to $t = t_0 + \epsilon$, since $H(t-t_0+\epsilon) - H(t-t_0-\epsilon) = 0$ outside these limits, and so the integrand is zero there. Between $t = t_0$ and $t = t_0 + \epsilon$ the function $H(t-t_0+\epsilon) - H(t-t_0-\epsilon) = 1$, so the expression on the right in (5.67) simplifies as

$$\lim_{\epsilon \rightarrow 0^+} \int_a^b \frac{A}{2\epsilon} (H(t-t_0+\epsilon) - H(t-t_0-\epsilon)) g(t) dt = \lim_{\epsilon \rightarrow 0^+} \frac{A}{2\epsilon} \int_{t_0-\epsilon}^{t_0+\epsilon} g(t) dt \quad (5.70)$$

after moving the constant $A/2\epsilon$ out in front of the integral. To evaluate the integral on the right in (5.70), let $G(t)$ be an antiderivative for $g(t)$, so $G'(t) = g(t)$. Then by the fundamental theorem of calculus

$$\int_{t_0-\epsilon}^{t_0+\epsilon} g(t) dt = G(t) \Big|_{t=t_0-\epsilon}^{t=t_0+\epsilon} = G(t_0+\epsilon) - G(t_0-\epsilon). \quad (5.71)$$

The use of (5.70) and (5.71) in (5.67) leads to

$$\lim_{\epsilon \rightarrow 0^+} \int_a^b \frac{A}{2\epsilon} (H(t-t_0+\epsilon) - H(t-t_0-\epsilon)) g(t) dt = A \lim_{\epsilon \rightarrow 0^+} \frac{G(t_0+\epsilon) - G(t_0-\epsilon)}{2\epsilon}.$$

But the limit on the right is exactly $G'(t_0)$, which equals $g(t_0)$ (if g is continuous at $t = t_0$). We have shown that

$$\lim_{\epsilon \rightarrow 0^+} \int_a^b \frac{A}{2\epsilon} (H(t-t_0+\epsilon) - H(t-t_0-\epsilon)) g(t) dt = Ag(t_0),$$

which is exactly the assertion of Theorem 5.4.1.

Reading Exercise 5.4.5 Argue that if $a < b < t_0$ or $t_0 < a < b$ then

$$\int_a^b A \delta(t-t_0) g(t) dt = 0. \quad (5.72)$$

Hint: refer to Figure 5.11 and note that the integrand is zero in the limit that ϵ approaches zero.

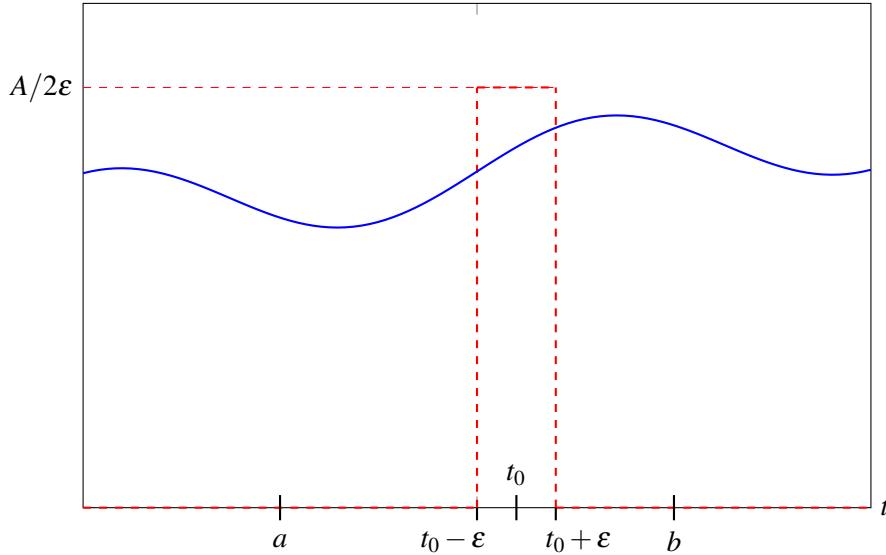


Figure 5.11: Interpreting the Dirac delta function. Graph of $\frac{A}{2\epsilon}(H(t - t_0 + \epsilon) - H(t - t_0 - \epsilon))$ (dashed, red) and graph of a continuous function $g(t)$ (solid black).

The Dirac Mass and Interval Endpoints

A Dirac delta function of the form $A\delta(t - t_0)$ is often said to have its **mass** at the point $t = t_0$, and such a function is sometimes called a **Dirac mass**. When we integrate a product $A\delta(t - t_0)g(t)$ over an interval $a \leq t \leq b$ and the Dirac mass is strictly inside the interval (so $a < t_0 < b$) then (5.69) holds. When the mass is strictly outside the interval ($t_0 < a$ or $t_0 > b$) then (5.72) holds.

But what if $t_0 = a$ or $t_0 = b$? This question is a bit delicate. Using the definition (5.67) yields the value $\frac{A}{2}g(a)$ for $t_0 = a$ or $\frac{A}{2}g(b)$ for $t_0 = b$ if g is continuous at the relevant point (see Reading Exercise 5.4.6). In some sense these results consider half of the Dirac mass to lie in the interval of integration and half outside. But there are variations on the definition (5.67) that yield different answers when t_0 is an endpoint, though they all agree with our definition if t_0 is strictly inside or outside the interval of integration.

We won't often encounter Dirac functions $\delta(t - t_0)$ where t_0 coincides with the end of the interval of integration. On those occasions that we do, for example, if $t_0 = a$ on the interval of integration $a \leq t \leq b$, it will be most convenient to set

$$\begin{aligned} \int_a^b A\delta(t - a)g(t) dt &= \lim_{t_0 \rightarrow a^+} \int_a^b A\delta(t - t_0)g(t) dt \\ &= \lim_{t_0 \rightarrow a^+} Ag(t_0). \end{aligned} \tag{5.73}$$

If g is continuous at $t = a$ the final limit of (5.73) is exactly $Ag(a)$.

Reading Exercise 5.4.6 Take $t_0 = 0$ and $g(t) = t^2 + t$ (or any other continuous function you like that has a simple antiderivative).

(a) Compute the integral

$$\int_2^5 \frac{1}{2\epsilon} (H(t - 2 + \epsilon) - H(t - 2 - \epsilon))g(t) dt$$

and take the limit as $\epsilon \rightarrow 0^+$. Compare the result found here to $g(2)$. This would be the definition of $\int_2^5 \delta(t - 2)g(t) dt$ according to (5.67). Hint: the limits on the integral are effectively $t = 2$ to $t = 2 + \epsilon$.

- (b) Assume $2 < t_0 < 5$. Compute $\int_2^5 \delta(t - t_0)g(t) dt$ according to (5.69).
(c) Use the result of part (b) to compute

$$\lim_{t_0 \rightarrow 2^+} \int_2^5 \delta(t - t_0)g(t) dt.$$

This is the definition of $\int_2^5 \delta(t - 2)g(t) dt$ according to (5.73), which we will adopt. Compare to $g(2)$ and the answer from parts (a) and (b).

5.4.3 Three Models: Money, Masses, and Medication

Let's return to the three models with impulsive forcing that we've already encountered, and reinterpret them using the Dirac delta function.

■ **Example 5.27** Recall the example at the start of this section, in which \$10,000 is invested at a 2 percent interest rate, compounded continuously. If additional deposits are made at a rate $r(t)$ then the account balance $p(t)$ obeys the ODE $p'(t) = 0.02p(t) + r(t)$; this was (5.56). There we considered how to model a lump sum deposit of \$5,000 at time $t = 3$. This might be considered as a deposit consisting of a very large deposit rate $r(t) = 5000/(2\varepsilon)$ for a very short window of time, $t = 3 - \varepsilon$ to $t = 3 + \varepsilon$. With the Dirac delta function in our arsenal we merely write

$$p'(t) = 0.02p(t) + 5000\delta(t - 3).$$

The initial condition is still $p(0) = 10000$. There is no longer any need to get emotionally involved with how long it took to deposit \$5,000. ■

■ **Example 5.28** Recall the example from earlier in this section involving a mass-spring-damper system with $m = 1$ kg, $c = 2$ newtons per meter per second, and $k = 10$ newtons per meter. The mass is at equilibrium and at rest at time $t = 0$. At time $t = 1$ the mass is subjected to a hammer blow with total impulse 10 newton-seconds. The relevant ODE is $u''(t) + 2u'(t) + 10u(t) = f(t)$, where $f(t)$ is the force of the hammer blow. We could consider $f(t)$ as having a very large value, $f(t) = 10/(2\varepsilon)$ for a very short time window $t = 1 - \varepsilon$ to $t = 1 + \varepsilon$. But a simpler model is obtained using the Dirac delta function,

$$u''(t) + 2u'(t) + 10u(t) = 10\delta(t - 1)$$

with initial conditions $u(0) = u'(0) = 0$. ■

■ **Example 5.29** Finally, let's return to the morphine administration problem of Example 5.1. To recap, after surgery a patient is given a 10 mg bolus of morphine at time $t = 0$. The morphine is metabolized and excreted with a half-life of 4 hours, and if additional morphine is given continuously at rate $r(t)$ mg per hour then we have the ODE

$$u'(t) = -ku(t) + r(t) \quad (5.74)$$

with initial condition $u(0) = 10$ and $k \approx 0.173$. In Section 5.3 we considered the situation in which $r(t)$ equals 1.5 mg per hour for $0 \leq t \leq 12$, but is then increased to 2.08 mg per hour for $t > 12$, in an attempt to provide better pain control. This was equation (5.45), in which the rate used in (5.74) was $r(t) = 1.5 + 0.58H(t - 12)$. The solution to the resulting ODE was shown in the right panel of Figure 5.1. That solution indicates that the amount of morphine in the patient's system increases toward the desired level too slowly.

To provide rapid pain relief, in addition to increasing the rate at which morphine is infused, suppose the patient is also given a 5 mg bolus at time $t = 12$. This could be modeled as a very high

infusion rate, $r(t) = 5/(2\varepsilon)$ for a short time window $t = 12 - \varepsilon$ to $t = 12 + \varepsilon$, or more directly, by adding $5\delta(t - 12)$ to $r(t)$. The amount of morphine is modeled using (5.74) with this $r(t)$, so we have

$$u'(t) = -ku(t) + 1.5 + 0.58H(t - 12) + 5\delta(t - 12)$$

and initial condition $u(0) = 10$. ■

Examples 5.27 to 5.29 contain ODEs that model impulsive phenomena using Dirac delta functions on their right sides, a nice conceptual and notational simplification compared to expressions like $\frac{1}{2\varepsilon}(H(t + \varepsilon) - H(t - \varepsilon))$. But how do we actually solve these types of ODEs? With the Laplace transform. But first, we need to compute the Laplace transform of the Dirac delta function.

5.4.4 The Laplace Transform of the Dirac Delta Function

The Laplace transform of the Dirac delta function $\delta(t - t_0)$ is

$$\mathcal{L}(\delta(t - t_0)) = \int_0^\infty e^{-st} \delta(t - t_0) dt. \quad (5.75)$$

When $t_0 > 0$ the value assigned to this integral is easy to deduce from (5.69) with the choice $A = 1$ and $g(t) = e^{-st}$, and we obtain

$$\mathcal{L}(\delta(t - t_0)) = e^{-st_0}. \quad (5.76)$$

When $t_0 = 0$ the Laplace transform is given by the integral

$$\mathcal{L}(\delta(t)) = \int_0^\infty e^{-st} \delta(t) dt$$

in which the Dirac mass lies at the left end of the interval of integration. This puts us precisely in the situation of (5.73) with $a = 0$, $b = \infty$, and $g(t) = e^{-st}$. In this case we interpret

$$\begin{aligned} \mathcal{L}(\delta(t)) &= \lim_{t_0 \rightarrow 0^+} \int_0^\infty e^{-st} \delta(t - t_0) dt \\ &= \lim_{t_0 \rightarrow 0^+} \mathcal{L}(\delta(t - t_0)) \\ &= \lim_{t_0 \rightarrow 0^+} e^{-st_0} \\ &= 1. \end{aligned}$$

where the second line follows from (5.75), the third from (5.76), and the last from the fact that e^{-st} is continuous in t . In summary,

$$\mathcal{L}(\delta(t)) = 1.$$

We are now in a position to solve ODEs with impulsive forcing functions.

5.4.5 Solving ODEs with Dirac Delta Functions

Let's return to Examples 5.27 - 5.29 and solve each relevant ODE.

■ **Example 5.30** In Example 5.27 the ODE of interest was

$$p'(t) = 0.02p(t) + 5000\delta(t - 3), \quad (5.77)$$

with $p(0) = 10000$. Laplace transforming both sides of (5.77) and making use of (5.76) and $p(0) = 10000$ yields

$$sP(s) - 10000 = 0.02P(s) + 5000e^{-3s},$$

where $P(s) = \mathcal{L}(p(t))$. We can solve for $P(s)$ as

$$P(s) = \frac{10000}{s - 0.02} + \frac{5000e^{-3s}}{s - 0.02}.$$

The inverse Laplace transform of the first term on the right is $10000e^{0.02t}$. The inverse transform of the second term on the right can be obtained by inverse transforming $5000/(s - 0.02)$ to obtain $5000e^{0.02t}$ and then using Theorem 5.3.1. The inverse transform of this second term is $5000H(t - 3)e^{0.02(t-3)}$. All in all then, the inverse transform of $P(s)$ yields the solution

$$p(t) = 10000e^{0.02t} + 5000H(t - 3)e^{0.02(t-3)}.$$

The first term on the right above quantifies the effect of the interest on the initial balance of \$10,000. The second term, which is not active until $t = 3$, reflects the \$5,000 deposit at $t = 3$ and interest on that deposit for $t > 3$. ■

■ **Example 5.31** In Example 5.28 the ODE of interest was

$$u''(t) + 2u'(t) + 10u(t) = 10\delta(t - 1) \quad (5.78)$$

with initial conditions $u(0) = u'(0) = 0$. Laplace transforming both sides of (5.78) and making use of (5.76) and $u(0) = u'(0) = 0$ yields

$$s^2U(s) + 2sU(s) + 10U(s) = 10e^{-s},$$

where $U = \mathcal{L}(u(t))$. We can solve for $U(s)$ as

$$U(s) = \frac{10e^{-s}}{s^2 + 2s + 10}.$$

To inverse transform, first consider $1/(s^2 + 2s + 10)$. Completing the square shows that $1/(s^2 + 2s + 10) = 1/((s + 1)^2 + 3^2)$, and from Table 5.1 we find the inverse transform of this quantity is $e^{-t} \sin(3t)/3$. From linearity and Theorem 5.3.1 it follows that the inverse Laplace transform of $U(s)$ is then

$$u(t) = \frac{10}{3}H(t - 1)e^{-(t-1)} \sin(3(t - 1)). \quad (5.79)$$

The formula for $u(t)$ reflects the fact that the mass remains at rest until time $t = 1$, when the hammer blow lands. ■

Reading Exercise 5.4.7 The momentum of the mass in Example 5.31 just after the hammer blow lands is $\lim_{t \rightarrow 1^+} mu'(t)$ (mass times velocity) where $m = 1$. Use (5.79) to compute this limit and compare it to the impulse delivered by the hammer blow, including the units on both quantities, assuming all units are SI. See also Exercise 5.4.10.

■ **Example 5.32** In Example 5.29 the ODE of interest was

$$u'(t) = -ku(t) + 1.5 + 0.58H(t - 12) + 5\delta(t - 12) \quad (5.80)$$

with initial condition $u(0) = 10$. Laplace transforming both sides of (5.80) and using $u(0) = 10$ yields

$$sU(s) - 10 = -kU(s) + \frac{1.5}{s} + \frac{0.58e^{-12s}}{s} + 5e^{-12s}.$$

We can solve for $U(s)$ to find

$$U(s) = \frac{10}{s+k} + \frac{1.5}{s(s+k)} + \frac{0.58e^{-12s}}{s(s+k)} + \frac{5e^{-12s}}{s+k}. \quad (5.81)$$

To find $u(t)$ we inverse transform each term on the right in (5.81).

The inverse transform of $1/(s+k)$ is e^{-kt} , so that the inverse transform of the first term on the right in (5.81) is $10e^{-kt}$. The inverse transform of the fourth term, $5e^{-12s}/(s+k)$, follows from $\mathcal{L}^{-1}(1/(s+k)) = e^{-kt}$ and Theorem 5.3.1, and is $5H(t-12)e^{-k(t-12)}$. The inverse transform of $\frac{1}{s(s+k)}$ can be obtained from the partial fraction decomposition

$$\frac{1}{s(s+k)} = \frac{1}{ks} - \frac{1}{k(s+k)}$$

and is given by the function $\phi(t) = 1/k - e^{-kt}/k$. As a result the inverse transform of the second term $\frac{1.5}{s(s+k)}$ on the right in (5.81) is $1.5\phi(t)$. The inverse transform of the third term, $\frac{0.58e^{-12s}}{s(s+k)}$, follows from Theorem 5.3.1 and is $0.58H(t-12)\phi(t-12)$. Putting all of this together shows that the solution to (5.80) is

$$u(t) = 10e^{-kt} + 1.5\phi(t) + 0.58H(t-12)\phi(t-12) + 5H(t-12)e^{-k(t-12)}$$

where $\phi(t) = 1/k - e^{-kt}/k$ and $k \approx 0.173$. This function is graphed in Figure 5.12. At time $t = 12$ we see a jump in the value of $u(t)$. This jump is precisely 5 mg. ■

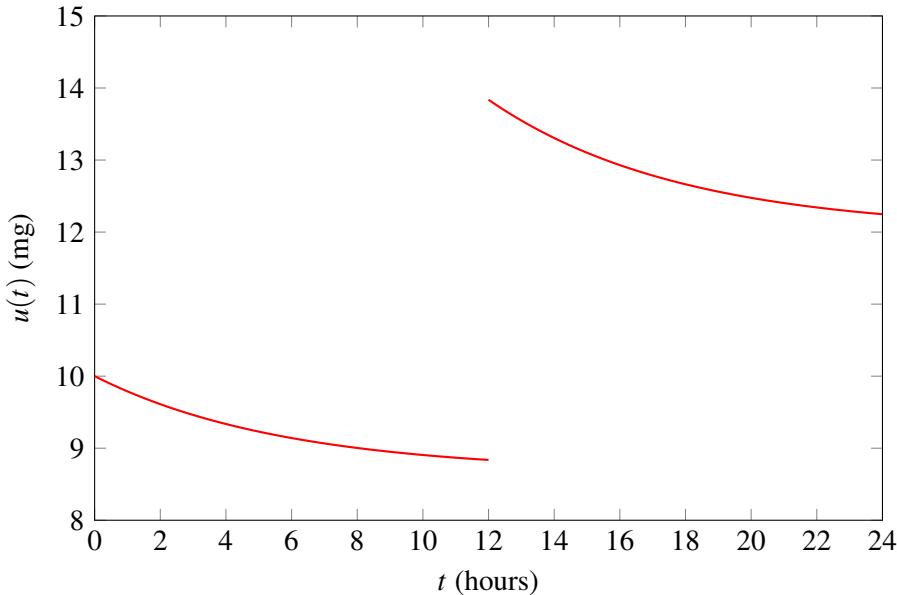


Figure 5.12: Graph of the solution to (5.80) with $u(0) = 10$, to model the amount of morphine $u(t)$ (mg) in a patient's body with 5 mg bolus administered at time $t = 12$.

5.4.6 Summary and a Few Remarks

The Dirac delta function is a mathematical idealization of impulsive phenomena that provides a convenient framework for modeling, and simplifies the computations needed to solve the ODEs that arise. It has many uses in physics; for example, to represent point masses and point charges.

Here is one issue that you may not have noticed in our treatment of the Dirac delta function $\delta(t)$ and the Heaviside function $H(t)$. Consider the simple ODE

$$u'(t) = -1 + 2H(t - 1)$$

with $u(0) = 1$. The solution can be obtained using Laplace transforms and is $u(t) = 1 - t + 2(t - 1)H(t - 1)$, which is entirely equivalent to $u(t) = |t - 1|$. There's just one problem: $u'(t)$ doesn't exist when $t = 1$, for the graph of $|t - 1|$ has a corner there. How can we call $u(t)$ a solution to a differential equation if $u(t)$ is not differentiable at some point in its domain? It might seem like failing to be differentiable at just one point is no big deal, but by that reasoning the Heaviside function $H(t)$ satisfies the ODE $H'(t) = 0$, except at $t = 0$, and so we might be tempted to conclude that $H(t)$ is constant. The careful resolution of this complication leads to the theory of distributions and generalized functions, something usually encountered in an advanced ODE course. See [132] for more information.

5.4.7 Laplace Transform Table

For convenience, here is a more complete table of Laplace transforms that includes the Heaviside function, shifting theorems, and Dirac delta function, and the result of Exercises 5.2.10 and 5.2.18.

Function	Laplace Transform	Comment
C	C/s	
t^n	$n!/s^{n+1}$	n an integer
e^{at}	$1/(s - a)$	$s > a$
$t^n e^{at}$	$n!/(s - a)^{n+1}$	n an integer
$\sin(bt)$	$b/(s^2 + b^2)$	
$\cos(bt)$	$s/(s^2 + b^2)$	
$e^{at} \sin(bt)$	$b/((s - a)^2 + b^2)$	$s > a$
$e^{at} \cos(bt)$	$(s - a)/((s - a)^2 + b^2)$	$s > a$
$e^{at} f(t)$	$F(s - a)$	(first shifting theorem)
$-tf(t)$	$F'(s)$	(Exercise 5.2.10)
$H(t - c)$	e^{-cs}/s	(H is the Heaviside function)
$H(t - c)f(t - c)$	$e^{-cs}F(s)$	(second shifting theorem)
$H(t - c)g(t)$	$e^{-cs}F(s)$	(second shifting theorem II, $f(t) = g(t + c)$)
$\delta(t - c)$	e^{-cs}	($\delta(t)$ is the Dirac delta function)
$\int_0^t f(\tau) d\tau$	$F(s)/s$	(Exercise 5.2.18)

Table 5.2: Time domain and s -domain Laplace transform pairs.

5.4.8 Exercises

Exercise 5.4.1 Solve the following first-order ODEs using the method of Laplace transforms. In each case plot the solution on the interval $0 \leq t \leq 10$.

- (a) $u'(t) = -2u(t) + 4\delta(t - 5)$ with $u(0) = 1$.

- (b) $u'(t) = -3u(t) + 3\delta(t-3) - 6H(t-5)$ with $u(0) = 1$.
 (c) $u'(t) = -u(t) + tH(t-1) - 3\delta(t-2)$ with $u(0) = 2$.

Exercise 5.4.2 Solve the following second-order ODEs using the method of Laplace transforms.

In each case plot the solution on the interval $0 \leq t \leq 10$.

- (a) $u''(t) + 4u'(t) + 3u(t) = \delta(t-1)$ with $u(0) = u'(0) = 0$.
 (b) $u''(t) + u(t) = \delta(t-3)$ with $u(0) = u'(0) = 0$.
 (c) $u''(t) + 4u'(t) + 4u(t) = 1 + 5\delta(t-2)$ with $u(0) = 1$ and $u'(0) = 2$.
 (d) $u''(t) + 4u(t) = \cos(2t) - 20\delta(t-3)$ with $u(0) = 1$ and $u'(0) = 0$.

Exercise 5.4.3 Define a function $\phi(t)$ by the definite integral

$$\phi(t) = \int_{-\infty}^t \delta(z) dz$$

where δ is the Dirac delta function. Compute $\phi(t)$ explicitly for $t < 0$ and then for $t > 0$. How does the answer compare to $H(t)$? Why is this answer consistent with the nonsensical statement (5.65)?

Exercise 5.4.4 A spring-mass-damper system has mass $m = 4$ kg, damping constant $c = 16$ newtons per meter per second, and spring constant $k = 116$ newtons per meter. The mass is at rest and equilibrium at time $t = 0$, and no other forces acts on the mass until time $t = 5$, when a hammer blow strikes the mass with total impulse 20 newton-seconds.

- (a) Model the situation using an appropriate ODE and initial conditions.
 (b) Solve the ODE using the Laplace transform and plot the solution on the interval $0 \leq t \leq 10$. Comment on what you see—does it make sense?

Exercise 5.4.5 A salt tank contains 100 liters of pure water at time $t = 0$, when salty water begins flowing into the tank at 2 liters per minute. The incoming liquid contains a concentration of 0.1 kg of salt per liter. The well-stirred liquid flows out of the tank at 2 liters per minute.

- (a) Model the situation with a first-order ODE, with $x(t)$ as the mass of salt in the tank at time t . Solve this ODE using the Laplace transform.
 (b) Suppose that at time $t = 20$ minutes 5 kg of salt is dumped into the tank and dissolves instantaneously. Modify the ODE from part (a) appropriately (Hint: at $t = 20$ salt enters the tank at a high rate, for a very brief period.) Solve the resulting ODE using the Laplace transform. Plot the solution to make sure it's sensible.

Exercise 5.4.6 A patient is given a 10 mg bolus of morphine at time $t = 0$, followed by a 5 mg bolus at times $t = 4, 8$, and 12 hours. Formulate an appropriate ODE to model this situation using (5.2) with $k = 0.173$. Solve the ODE and plot the solution for $0 \leq t \leq 24$. Is this solution consistent with given information?

Exercise 5.4.7 An investment account is opened with \$1000 at time $t = 0$. The account earns interest at an annual rate of 5 percent, compounded continuously, that is, the account accrues interest at a rate of $0.05p(t)$. Suppose the deposit rate is $r(t) = 500$ dollars per year for $t > 0$.

- Formulate an appropriate ODE with initial condition, and solve using the Laplace transform. Plot the solution for time $0 \leq t \leq 10$.
- Suppose that in addition to the deposit rate of 500 dollars per year, a lump sum deposit of \$1000 is made at time $t = 2$. Formulate and solve an appropriate ODE. What is the account balance at time $t = 10$?
- Redo part (b) but under the assumption that the \$1000 lump sum deposit is made at time $t = 8$ (not $t = 2$), and compute the account balance at time $t = 10$. Why is the balance in part (b) higher?

Exercise 5.4.8 An unforced spring-mass-damper system obeys $2u''(t) + 4u'(t) + 52u(t) = 0$ where $u(t)$ is the position of the mass, with initial conditions $u(0) = 1$ and $u'(0) = -1$.

- Solve the ODE to find the position $u(t)$ of the mass.
- The system is underdamped, so the mass repeatedly passes through equilibrium. Find the second positive time $t = t_2$ at which the mass passes through equilibrium.
- Suppose that at time $t = t_2$ a hammer blow of impulse A is to be applied to the mass to bring it to a dead stop. What should A equal? (A can be negative.) Hint: it should counteract the momentum of the mass.
- Solve the ODE $2u''(t) + 4u'(t) + 52u(t) = A\delta(t - t_2)$ with initial conditions $u(0) = 1$ and $u'(0) = -1$, and A and t_2 as from parts (b) and (c). Plot the solution on the range $0 \leq t \leq 2$.

Exercise 5.4.9 An investment account pays 4 percent interest, compounded continuously, and has a balance of \$2000 at time $t = 0$. Lump sum deposits of A dollars are to be made at times $t = 2$ and $t = 4$, where A is to be determined.

- Formulate an appropriate ODE and initial condition.
- Solve the ODE in part (a). The solution should contain the parameter A .
- Suppose we want to choose these deposits so the account has \$10,000 at time $t = 10$. What value of A gives the desired account balance?

Exercise 5.4.10 Suppose a particle of mass m is at rest somewhere on the x -axis, for all times $t < t_0$, where t_0 is a positive time. At $t = t_0$ a hammer blow of total impulse A newton-seconds is applied to the particle, which sets it in motion; no other forces act on the particle for $t > t_0$ (not even friction). Let $v(t)$ denote the velocity of the particle. From $F = ma$ with $F = A\delta(t - t_0)$ and $a = v'$ it follows that

$$mv'(t) = A\delta(t - t_0). \quad (5.82)$$

- Solve (5.82) for $v(t)$, using the initial condition $v(0) = 0$.
- Show that the momentum $mv(t)$ of the particle after the hammer blow is equal to A , precisely the same as the impulse of the blow.

5.5 Input-Output, Transfer Functions, and Convolution

Many of the physical systems we've seen in this text can be viewed from the point of view of input and output, with the relevant linear ODE governing the mathematics of how input is processed into output. In the time domain this is often quantified by a mathematical process known as *convolution*, in which a pair of functions are combined to form a new function. In the *s*-domain, however, the mathematics becomes much simpler and is governed by something called the *transfer function* of the process. In this section we develop the appropriate mathematics and then apply it to the problem of system identification, a form of parameter estimation. In the next section we'll look at how these tools can be applied to the subject of control theory.

5.5.1 A System Identification Problem

In Section 3.4 we considered methods for estimating unknown parameters in ODEs that model physical phenomena. For example, one might want to estimate the cooling constant in the Newton-cooling ODE by using temperature data collected over time. Or one may wish to estimate the damping or spring constants in a spring-mass-damper system by periodically measuring the displacement of the mass. In previous cases we solved the estimation problem using time domain data and the time domain solution to the ODE.

To motivate some of the mathematics in this section, let's revisit the problem of estimating the parameters in a spring-mass-damper system. This time we'll look at the problem in the *s*-domain. Consider a driven spring-mass-damper system governed by

$$mu''(t) + cu'(t) + ku(t) = f(t) \quad (5.83)$$

with some initial conditions, which we will take to be $u(0) = u'(0) = 0$. The parameters m , c , and k are unknowns, and the goal is to estimate these parameters. To do this we apply a stimulus force $f(t)$ of our choosing to the system, measure the response $u(t)$, and from this information try to determine m , c , and k . Think of $f(t)$ as an input to the system and the mass response $u(t)$ as an output.

As a specific example, suppose the forcing function we apply in (5.83) is $f(t) = H(t - 1) \cos(t - 1)$, and suppose the mass responds according to

$$u(t) = H(t - 1) \left(\frac{\cos(t - 1)}{20} + \frac{\sin(t - 1)}{10} - \frac{e^{-(t-1)}}{8} + \frac{3e^{-3(t-1)}}{40} \right). \quad (5.84)$$

What information does this give us about m , c , and k ? This kind of problem, in which unknown system parameters must be identified from input-output data is often called **system identification**; it's a form of parameter estimation. We'll return to this problem shortly in Example 5.33.

5.5.2 Input-Output Systems

The input-output model for a system or process is one of the most common paradigms in science and engineering. We'll use the driven spring-mass-damper example above as a concrete example, which encompasses many of the real situations we've looked at, including the earthquake model, the bike shock absorber, the vibration isolation table, and the RLC circuit of Section 4.1. The corresponding models were all of the general form of (5.83). We may think of $f(t)$ as a stimulus or **input** to the system and $u(t)$ as a response or **output**, as illustrated in Figure 5.13. The input $f(t)$ is processed by the spring-mass-damper system with parameters m , c , and k , and turned into the output response $u(t)$; naturally, the output of this process depends on the values of m , c , and k . The initial conditions also figure into the computation, so for simplicity we fix those at $u(0) = u'(0) = 0$.

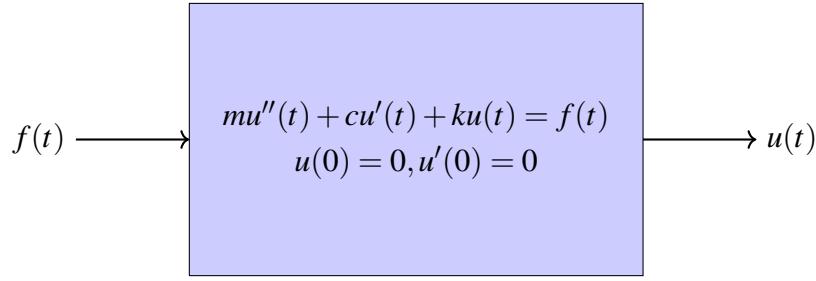


Figure 5.13: Spring-mass as an input-output system in the time domain, with input force $f(t)$ and output response $u(t)$.

It would be nice if the process by which $f(t)$ is turned into $u(t)$ was simple, for example, if $u(t) = f'(t)$ held, but this is rarely the case. To obtain $u(t)$ from $f(t)$ we have to solve the ODE (5.83), and the resulting dependence of $u(t)$ on $f(t)$ isn't very explicit. But if we use the Laplace transform to move the problem into the s -domain, things become much easier, both conceptually and computationally. Laplace transforming both sides of the ODE (5.83) and substituting in the initial conditions yields

$$(ms^2 + cs + k)U(s) = F(s) \quad (5.85)$$

where $U = \mathcal{L}(u(t))$ and $F = \mathcal{L}(f(t))$. The s -domain version of Figure 5.13 is shown in Figure 5.14.

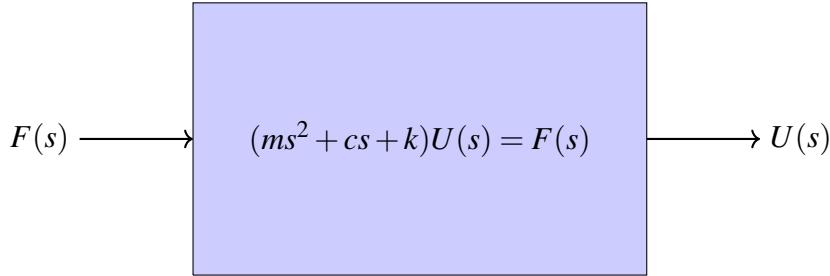


Figure 5.14: Spring-mass as an input-output system in the s -domain, with input $F(s)$ and output $U(s)$.

The input-output process in the s -domain is governed by (5.85), which makes the manner by which $F(s)$ is turned into $U(s)$ completely transparent. We can easily see that

$$U(s) = G(s)F(s), \quad (5.86)$$

where

$$G(s) = \frac{1}{ms^2 + cs + k}. \quad (5.87)$$

The function $G(s)$ is called the **transfer function** for this spring-mass-damper system.

Reading Exercise 5.5.1 Show that $G = \mathcal{L}(g_\delta(t))$ where $g_\delta(t)$ is the solution to

$$mg_\delta''(t) + cg_\delta'(t) + kg_\delta(t) = \delta(t)$$

with $g_\delta(0) = g_\delta'(0) = 0$.

The function $g_\delta(t)$ in Reading Exercise 5.5.1 is called the **unit impulse response** or just **impulse response** of the system governed by the ODE $mu'' + cu' + ku = f(t)$. It is the response of the system when $f(t) = \delta(t)$, a hammer blow of total impulse 1 at time $t = 0$.

We will revisit the idea of transfer functions and impulse responses shortly, but let us emphasize now that for this analysis the system must be governed by a linear, constant-coefficient ODE. The input or forcing function $f(t)$ appears as the right hand side of the nonhomogeneous version of the equation. The relevant ODE may be first- or second-order, or higher. We also assume, for now, that the system has zero initial conditions.

With these ideas we can already solve the system identification problem posed at the end of Section 5.5.1.

■ **Example 5.33** The spring-mass-damper system at the end of Section 5.5.1 had input $f(t) = H(t - 1)\cos(t - 1)$ with initial data $u(0) = u'(0) = 0$ and the system responded with $u(t)$ as given in (5.84). We can think of $u(t)$ as describing measured data. The goal is to determine m, c , and k from knowledge of what went in— $f(t)$ —and what came out— $u(t)$.

We begin with the relation (5.86) and

$$U(s) = \mathcal{L}(u(t)) = \frac{se^{-s}}{2(s^2 + 1)(s + 1)(s + 3)},$$

computed from our knowledge of $u(t)$ in (5.84). We also know $F(s) = \mathcal{L}(f(t)) = \frac{se^{-s}}{s^2 + 1}$. Substitute this information into (5.86) to find

$$\underbrace{\frac{se^{-s}}{2(s^2 + 1)(s + 1)(s + 3)}}_{U(s)} = G(s) \underbrace{\left(\frac{se^{-s}}{s^2 + 1}\right)}_{F(s)}.$$

Some obvious cancellations show that

$$\frac{1}{2(s + 1)(s + 3)} = \frac{1}{2s^2 + 8s + 6} = G(s), \quad (5.88)$$

and we conclude that the system transfer function is $G(s) = 1/(2s^2 + 8s + 6)$. But of course we also know from (5.87) that the transfer function for this spring-mass system is of the form $G(s) = 1/(ms^2 + cs + k)$. This makes it clear that $m = 2, c = 8$, and $k = 6$. In short, with knowledge of the input $f(t)$ and output $u(t)$, we can determine the transfer function $G(s)$ and therefore also the system parameters. ■

5.5.3 Convolution

Equation (5.86) quantifies how an input $F(s)$ is processed into an output $U(s)$ and is exceedingly straightforward, a simple multiplication of two functions G and F . We've previously noted how many operations in the time domain have a parallel in the s -domain and vice versa. It turns out there is also a time domain counterpart to the operation that transforms the input $F(s)$ to the output $U(s)$ in the s -domain via the relation $U(s) = G(s)F(s)$. The time domain operation is called **convolution**. It plays an important role in the analysis of many engineering systems. It's a strange-looking beast and it comes in a variety of forms, but you've been doing it since you were in the 5th grade.

Grade School Convolution

Let's take a brief excursion. Suppose we are given two positive integers, e.g., 237 and 461 (in base 10, although that really doesn't matter), and want to compute the product 237×461 . The grade school procedure you learned for long multiplication is a variation on writing

$$237 = 2 \cdot 10^2 + 3 \cdot 10^1 + 7 \cdot 10^0$$

$$461 = 4 \cdot 10^2 + 6 \cdot 10^1 + 1 \cdot 10^0$$

and then multiplying every term in the expansion of 237 by every term in the expansion of 461, then adding up the whole mess. The product will contain a 10^4 term stemming from the $10^2 \cdot 10^2$ product, and a 10^3 term stemming from the $10^1 \cdot 10^2$ and $10^2 \cdot 10^1$ cross terms, and a 10^2 term, stemming from the $10^2 \cdot 10^0$, $10^1 \cdot 10^1$, or $10^0 \cdot 10^2$ cross terms, and so on. Collecting the like powers of 10 in the product together shows that

$$\begin{aligned} 237 \cdot 461 &= (2 \cdot 4)10^4 + (2 \cdot 6 + 4 \cdot 3)10^3 + (2 \cdot 1 + 3 \cdot 6 + 7 \cdot 4)10^2 \\ &\quad + (3 \cdot 1 + 7 \cdot 6)10^1 + (7 \cdot 1)10^0 \\ &= (8) \cdot 10^4 + (24) \cdot 10^3 + (48) \cdot 10^2 + (45) \cdot 10^1 + (7) \cdot 10^0. \end{aligned} \quad (5.89)$$

That's the heart of long multiplication, although at this point it would be traditional to perform a carry in (5.89), for example, by noting that $(45) \cdot 10^1 = 4 \cdot 10^2 + 5 \cdot 10^1$ so that the 4's digit can be carried into the 10^2 term (which is then subject to its own carry). The operation of carrying normalizes the expansion on the right in (5.89) so that all digits are in the range 0 to 9, and has the advantage that the product is expressed in its unique base 10 representation. But if we don't worry about carrying, it would be perfectly legitimate to write something like

$$237 \cdot 461 = (8, 24, 48, 45, 7)$$

to express the right side of (5.89), where the comma delineates the powers of 10 in $(8, 24, 48, 45, 7)$, and with the understanding that the far right entry (in this case, 7) is the 10^0 term. With this notation we could also write the number 237 as $(2, 3, 7)$ and 461 as $(4, 6, 1)$.

This procedure can be used to multiply arbitrary three digit positive integers $a = (a_2, a_1, a_0)$ and $b = (b_2, b_1, b_0)$ (understood to mean $a = a_2 10^2 + a_1 10^1 + a_0 10^0$ and $b = b_2 10^2 + b_1 10^1 + b_0 10^0$) and find that

$$ab = (a_2 b_2, a_2 b_1 + a_2 b_2, a_2 b_0 + a_1 b_1 + a_0 b_2, a_1 b_0 + a_0 b_1, a_0 b_0).$$

More generally, you can convince yourself that if we have integers a and b with

$$\begin{aligned} a &= (a_n, a_{n-1}, \dots, a_1, a_0) \\ b &= (b_m, b_{m-1}, \dots, b_1, b_0) \end{aligned} \quad (5.90)$$

then the product $c = ab$ can be written as

$$c = (c_{m+n}, c_{m+n-1}, \dots, c_1, c_0) \quad (5.91)$$

where

$$c_k = \sum_{j=0}^k a_j b_{k-j}, \quad (5.92)$$

though the c_k are not necessarily digits in the traditional 0 to 9 range. The integer sequence with entries c_k in (5.91) where c_k is given by (5.92) is called the **convolution** of the integer sequences that defined a and b in (5.90). Convolution is usually denoted with a “*”. One notation might be to write

$$(c_{m+n}, c_{m+n-1}, \dots, c_1, c_0) = (a_n, a_{n-1}, \dots, a_1, a_0) * (b_m, b_{m-1}, \dots, b_1, b_0).$$

We might even write just $c = a * b$. This is a form of convolution that's common for integer sequences.

Looking at 5th grade arithmetic in this way might start to convince you that convolution is an operation that arises rather naturally in many areas of mathematics, science, and engineering. The corresponding convolution operation appropriate to the study of ODEs has many parallels to this integer version.

Reading Exercise 5.5.2 Suppose $a = (2, 3, 0, 4)$ and $b = (1, 3, 7)$ in the notation above. Use (5.92) to compute c_0 through c_5 of the product ab (note that $n = 3$ and $m = 2$ here). Then perform the carries and verify that you got the correct product for 2304×137 .

Convolution for Functions

Let's just get right to it.

Definition 5.5.1 Let $f(t)$ and $g(t)$ be piecewise continuous functions defined on $0 \leq t < \infty$. The **convolution** of f and g is the function $(f * g)$ defined by

$$(f * g)(t) = \int_0^t f(\tau)g(t - \tau) d\tau. \quad (5.93)$$

It's somewhat conventional to parenthesize the convolution and write $(f * g)$ instead of $f * g$. Then $(f * g)(t)$ means the function $(f * g)$ applied to t . You may also see the notation $f(t) * g(t)$. Because f and g in (5.93) are assumed to be piecewise continuous and the integral above is over a finite interval, this integral will converge for any $t \geq 0$, and so $(f * g)(t)$ is defined for all $t \geq 0$. The choice of τ as the dummy variable of integration is common.

In Reading Exercise 5.5.4 you will explore similarities between (5.92) and (5.93), but first let's look at an example.

■ **Example 5.34** Let's compute the convolution of $f(t) = t$ and $g(t) = t^2$. From (5.93) we have (using $f(\tau) = \tau$ and $g(t - \tau) = (t - \tau)^2$)

$$\begin{aligned} (f * g)(t) &= \int_0^t \tau(t - \tau)^2 d\tau \\ &= \int_0^t (t^2\tau - 2t\tau^2 + \tau^3) d\tau \\ &= \left(\frac{t^2\tau^2}{2} - \frac{2t\tau^3}{3} + \frac{\tau^4}{4} \right) \Big|_{t=0}^{t=\tau} \\ &= \frac{t^4}{12}, \end{aligned}$$

after simplifying. ■

Reading Exercise 5.5.3 Compute the convolution of $f(t) = e^{-t}$ and $g(t) = e^{-2t}$.

Reading Exercise 5.5.4 Compare (5.92) with (5.93) using the correspondences $t \leftrightarrow k$, $\tau \leftrightarrow j$, $f \leftrightarrow a$, $g \leftrightarrow b$, and $q \leftrightarrow c$, and with the sum in (5.92) replaced by the integral in (5.93). Convince yourself that they have the same general structure.

Properties of Convolution

Convolution has some important algebraic properties. If f_1 , f_2 , and g are piecewise continuous on $0 \leq t < \infty$ and a and b are any scalars then convolution satisfies the properties listed below.

Commutativity: $f_1 * g = g * f_1$.

Distributivity: $(af_1 + bf_2) * g = af_1 * g + bf_2 * g$ and $g * (af_1 + bf_2) = ag * f_1 + bg * f_2$.

Associativity: $(f_1 * f_2) * g = f_1 * (f_2 * g)$.

The proofs are fairly routine manipulations of the integral that defines convolution, or they can be proved using Laplace transform; see Exercise 5.5.12. In Exercise 5.5.13 you'll show that if the functions f_1 and f_2 are of exponential order then $f_1 * f_2$ is of exponential order.

Here is one of the most important and useful properties concerning convolution and how it interacts with the Laplace transform.

Theorem 5.5.1 — The Convolution Theorem. Let $f_1(t)$ and $f_2(t)$ be piecewise continuous functions of exponential order defined on $0 \leq t < \infty$ and let $q = f_1 * f_2$. Let $F_1(s) = \mathcal{L}(f_1(t))$, $F_2(s) = \mathcal{L}(f_2(t))$, and $Q(s) = \mathcal{L}(q(t))$. Then

$$Q(s) = F_1(s)F_2(s). \quad (5.94)$$

Also, if $F_1(s)$ is defined for $s > a$ and $F_2(s)$ is defined for $s > b$, then $Q(s)$ is defined for $s > \max(a, b)$.

Equation (5.94) might also be written as $\mathcal{L}(f_1 * f_2) = \mathcal{L}(f_1)\mathcal{L}(f_2)$. The proof of Theorem 5.5.1 is a straightforward change of variable in a double integral and is given in Section 5.5.5. Let's look at a few examples where we apply Theorem 5.5.1.

■ **Example 5.35** In Example 5.34 with $f_1(t) = t$ and $f_2(t) = t^2$ we computed that $q(t) = (f_1 * f_2)(t) = t^4/12$. We can compute that the corresponding Laplace transforms are given by $F_1(s) = 1/s^2$, $F_2(s) = 2/s^3$, and $Q(s) = 2/s^5$. We see that $Q(s) = F_1(s)F_2(s)$, in accordance with the Theorem 5.5.1. ■

■ **Example 5.36** Theorem 5.5.1 can be used to find inverse Laplace transforms. For example, suppose $Q(s) = \frac{1}{s(s+1)^2}$. The function $Q(s)$ can be split as $Q(s) = F_1(s)F_2(s)$ where $F_1(s) = 1/(s+1)^2$ and $F_2(s) = 1/s$ (there are other ways to split Q). Then from Table 5.2 we have $f_1(t) = te^{-t}$ and $f_2(t) = H(t)$, and so by Theorem 5.5.1 we have $q = f * g$, or

$$\begin{aligned} q(t) &= \int_0^t f_1(\tau)f_2(t-\tau)d\tau \\ &= \int_0^t \tau e^{-\tau} H(t-\tau)d\tau \\ &= \int_0^t \tau e^{-\tau} d\tau \quad (\text{since } H(t-\tau) = 1 \text{ for } 0 \leq \tau \leq t) \\ &= -(\tau+1)e^{-\tau} \Big|_{\tau=0}^{\tau=t} \\ &= 1 - e^{-t}(1+t). \end{aligned}$$

Convolution with Heaviside Functions

In Example 5.36 we computed the convolution of a specific function f with the Heaviside function. More generally, the convolution of a function $f(t)$ with a Heaviside function $H(t)$ is the function $(f * H)(t)$ given by the integral

$$\begin{aligned} (f * H)(t) &= \int_0^t f(\tau)H(t-\tau)d\tau \\ &= \int_0^t f(\tau)d\tau. \end{aligned} \quad (5.95)$$

Note that (5.95) along with the fundamental theorem of calculus shows that $(f * H)(t)$ is an antiderivative for $f(t)$ that satisfies $(f * H)(0) = 0$. Since $\mathcal{L}(H(t)) = 1/s$, Theorem 5.5.1 shows that

$$\mathcal{L}(f * H) = \mathcal{L}(f(t))\mathcal{L}(H(t)) = \frac{F(s)}{s}, \quad (5.96)$$

which was the result of Exercise 5.2.18.

Convolution with Dirac Delta Functions

Consider the convolution of a continuous function f with a Dirac delta function. Let $\delta_{t_0}(t) = \delta(t - t_0)$ where $t_0 > 0$, so δ_{t_0} is a Dirac delta function with its mass at $t = t_0$. We define the convolution of a function f with $\delta_{t_0}(t)$ by using (5.69) (with $A = 1$, $g = f$, and variable of integration τ instead of t),

$$\begin{aligned} (\delta_{t_0} * f)(t) &= \int_0^t \delta_{t_0}(\tau) f(t - \tau) d\tau \\ &= \int_0^t \delta(\tau - t_0) f(t - \tau) d\tau. \end{aligned} \tag{5.97}$$

The integral in (5.97) can be evaluated as follows: If $0 < t_0 < t$ then the mass of the Dirac delta function is in the range of integration. By (5.69) the integral equals $f(t - t_0)$. If $t < t_0$ then the Dirac mass lies outside the interval of integration and the integral in (5.97) is zero. This can be expressed as

$$(\delta_{t_0} * f)(t) = H(t - t_0) f(t - t_0). \tag{5.98}$$

In the special case that $t_0 = 0$ the convolution of $\delta(t)$ with f can be computed using (5.73) with $A = 1, a = 0$, and $b = \infty$ to find

$$(\delta * f)(t) = \int_0^\infty \delta(\tau) f(t - \tau) d\tau = f(t). \tag{5.99}$$

Equation (5.99) is consistent with Theorem 5.5.1, for if we Laplace transform the left side of (5.99) we should obtain $\mathcal{L}(\delta * f) = \mathcal{L}(\delta)\mathcal{L}(f(t)) = 1 \cdot \mathcal{L}(f(t)) = \mathcal{L}(f(t))$, which is, of course, exactly the Laplace transform of the right side of (5.99).

Reading Exercise 5.5.5 Use the second shifting theorem (Theorem 5.3.1) to compute $\mathcal{L}(H(t - t_0)f(t - t_0))$ and verify that the result is $\mathcal{L}(\delta_{t_0}(t))\mathcal{L}(f(t))$, in agreement with Theorem 5.5.1.

5.5.4 The Impulse Response and Convolution

Equation (5.87) shows that the transfer function $G(s)$ of a system governed by $mu''(t) + cu'(t) + ku(t) = f(t)$ is given by $G(s) = 1/(ms^2 + cs + k)$. In conjunction with $U(s) = G(s)F(s)$ (this is (5.86)) the transfer function provides a conceptually and algebraically straightforward method to see how an input $f(t)$ is processed into the output system response $u(t)$, in the s -domain.

The equation $U(s) = G(s)F(s)$ has a time domain counterpart. To understand this counterpart, recall that in Reading Exercise 5.5.1 you showed that $G(s) = \mathcal{L}(g_\delta)$ (equivalently, $g_\delta(t) = \mathcal{L}^{-1}(G(s))$) where g_δ is the impulse response of the system. The function $g_\delta(t)$ is the solution to

$$mg_\delta''(t) + cg_\delta'(t) + kg_\delta(t) = \delta(t) \tag{5.100}$$

with $g_\delta(0) = g_\delta'(0) = 0$. The physical interpretation of $g_\delta(t)$ is that it is the motion of the mass in response to a hammer blow of total impulse 1 at time $t = 0$. The relations $G(s) = \mathcal{L}(g_\delta)$ or equivalently, $g_\delta(t) = \mathcal{L}^{-1}(G(s))$, show that the impulse response $g_\delta(t)$ and transfer function $G(s)$ are time domain and s -domain counterparts to one another.

If we know the impulse response $g_\delta(t)$ for a spring-mass system—that is, the response of the system to hammer blow at time $t = 0$ —then we can compute the response of the system to any input forcing function $f(t)$, at least with zero initial conditions. To show this, consider the ODE $mu''(t) + cu'(t) + ku(t) = f(t)$ with $u(0) = u'(0) = 0$. Laplace transforming both sides of this ODE and using the initial conditions leads to the conclusion $U(s) = G(s)F(s)$, which is exactly (5.86). By the convolution theorem (Theorem 5.5.1), the multiplication $G(s)F(s)$ in the s -domain

is corresponds to the operation $g_\delta * f$ in the time domain. Since $u(t)$ corresponds to $U(s)$, we have shown that the solution to $mu''(t) + cu'(t) + ku(t) = f(t)$ with initial data $u(0) = u'(0) = 0$ is

$$u(t) = (g_\delta * f)(t). \quad (5.101)$$

where g_δ is the unit impulse response of the system. Equation (5.101) is the time domain counterpart to $U(s) = G(s)F(s)$. For any input forcing function $f(t)$ we can compute the system's response $u(t)$ with initial data $u(0) = u'(0) = 0$ by performing the convolution (5.101).

■ **Example 5.37** Consider the driven harmonic oscillator $2u''(t) + 4u'(t) + 4u(t) = t$ with $u(0) = u'(0) = 0$. The transfer function for this system is given by

$$G(s) = \frac{1}{2s^2 + 4s + 4}.$$

An inverse Laplace transform shows that the impulse response is $g_\delta(t) = \mathcal{L}^{-1}(G(s))$ or

$$g_\delta(t) = \frac{1}{2}e^{-t} \sin(t).$$

With $f(t) = t$ for the forcing function, we have $F(s) = 1/s^2$. The Laplace transform of the solution is

$$U(s) = G(s)F(s) = \frac{1}{s^2(2s^2 + 4s + 4)}.$$

The actual time domain solution $u(t)$ can be found by inverse transforming $U(s)$. Alternatively, $u(t)$ can be computed as the convolution

$$\begin{aligned} u(t) &= g_\delta(t) * t \\ &= \int_0^t \frac{1}{2}e^{-\tau} \sin(\tau)(t - \tau) d\tau \\ &= \frac{1}{4}(t - 1 + e^{-t} \cos(t)). \end{aligned}$$

■

Impulse Response for First-Order ODEs

The notions of impulse response and transfer function are not limited to second-order equations. The ideas work just as well for linear first-order (and higher-order) equations, provided the equations are constant-coefficient. For example, consider a system governed by the first-order ODE $au'(t) + bu(t) = f(t)$ with initial condition $u(0) = 0$. A Laplace transform shows that the solution's Laplace transform $U(s)$ is given by $U(s) = G(s)F(s)$ where $F(s) = \mathcal{L}(f)$ and

$$G(s) = \frac{1}{as + b}. \quad (5.102)$$

Here $G(s)$ is the transfer function for this first-order equation. The impulse response of this first-order system is the solution $g_\delta(t)$ to

$$ag'_\delta(t) + bg_\delta(t) = \delta(t) \quad (5.103)$$

with $g_\delta(0) = 0$. Laplace transforming both sides of (5.103) makes it easy to see that $\mathcal{L}(g_\delta) = G(s)$, or equivalently, $g_\delta = \mathcal{L}^{-1}(G) = e^{-bt/a}/a$. From Theorem 5.5.1 we see that the time domain counterpart to $U(s) = G(s)F(s)$ is $u(t) = (g_\delta * f)(t)$. See Exercise 5.5.2 for examples.

5.5.5 System Identification with Impulsive Input

We considered a system identification problem in Section 5.5.1, then solved it using the Laplace transform and an input-output philosophy in Example 5.33. The input forcing function there was $f(t) = H(t - 1)\cos(t - 1)$, but it is quite common in practice to use impulsive inputs for system identification. Let's consider an example.

■ **Example 5.38** Suppose a spring-mass system has mass m , damping constant c , and spring constant k . Each of these parameters is to be determined. To do this we apply a known forcing function $f(t)$ to the system, in this case an impulse force $f(t) = 6\delta(t - 1)$. We assume the initial condition are $u(0) = u'(0) = 0$. By measuring the response $u(t)$ of the system to this input we can determine m, c , and k . This is most easily done in the s -domain.

Suppose the response of the system is $u(t) = H(t - 1)e^{-(t-1)} \sin(2t - 2)$. In the s -domain the relation between the input force $f(t)$ and output response $u(t)$ is $U(s) = G(s)F(s)$ where $G(s)$ is the system's transfer function. From (5.87) we know that $G(s) = 1/(ms^2 + cs + k)$. Because $f(t)$ and $u(t)$ are known, we can compute $F(s) = \mathcal{L}(f(t)) = 6e^{-s}$ and $U(s) = \mathcal{L}(u(t)) = 2e^{-s}/(s^2 + 2s + 5)$. From $U(s) = G(s)F(s)$ we conclude that

$$\frac{2e^{-s}}{s^2 + 2s + 5} = G(s)6e^{-s}. \quad (5.104)$$

Divide both sides in (5.104) by $6e^{-s}$ to conclude that $G(s) = 1/(3s^2 + 6s + 15)$. From $G(s) = 1/(ms^2 + cs + k)$ it is now apparent that $m = 3$, $c = 6$, and $k = 15$. ■

Examples 5.33 and 5.38 were both noise-free and we assumed that we had a functional form for both the input $f(t)$ and output response $u(t)$. These ideals cannot always be attained. For more realistic estimation problems see Exercises 5.5.7 or 5.5.8, and the projects in Section 5.7.

Remark 5.5.1 In the parameter estimation problems of Examples 5.33 and 5.38, we transformed the problems into the s -domain and found the answers we needed there. This is common in many applications, parameter estimation and otherwise. The Laplace transform is used to take the problem into the s -domain and perform all analysis there, without ever inverse transforming back to the time domain. Section 5.6 explores another topic in which all essential analysis takes place in the s -domain.

Proof of the Convolution Theorem

The proof of Theorem 5.5.1 is a fairly straightforward computation. With $F_1(s) = \mathcal{L}(f_1(t))$ and $F_2(s) = \mathcal{L}(f_2(t))$ set $p(t) = (f_1 * f_2)(t)$; that is,

$$p(t) = \int_0^t f_1(\tau)f_2(t - \tau) d\tau.$$

If f_1 and f_2 are of exponential order, then so is $p(t)$, so we can compute $\mathcal{L}(p)$; see Exercise 5.5.13. The Laplace transform $P(s) = \mathcal{L}(p(t))$ is given by the integral

$$\begin{aligned} P(s) &= \int_0^\infty e^{-st} p(t) dt \\ &= \int_0^\infty e^{-st} \left(\int_0^t f_1(\tau)f_2(t - \tau) d\tau \right) dt \\ &= \int_0^\infty \int_0^t e^{-st} f_1(\tau)f_2(t - \tau) d\tau dt. \end{aligned} \quad (5.105)$$

Thus $P(s)$ is defined by the double integral on the right in (5.105), an improper double integral that converges absolutely for sufficiently large s . We will evaluate this double integral as an iterated

integral, but switch the order of integration from $d\tau dt$ to $dt d\tau$. As is always the case when reversing the order of integration in a double integral, a clear sketch of the region of integration is invaluable. The region of integration in the $t\tau$ plane for the integral in (5.105) is defined by the inequalities $0 \leq \tau \leq t$ and $0 \leq t < \infty$ and is depicted as the shaded region in Figure 5.15. The region consists of all points below the line $\tau = t$.

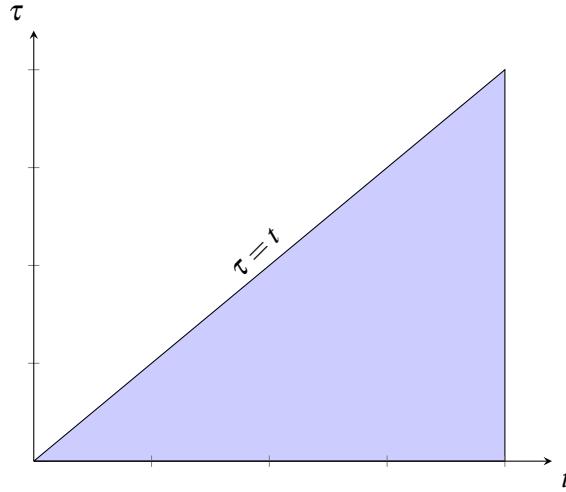


Figure 5.15: Region of integration for double integral in (5.105).

This region can also be defined by the inequalities $\tau \leq t < \infty$ and $0 \leq \tau < \infty$, and so the iterated integral (5.105) can be evaluated in the order $dt d\tau$ to find

$$\begin{aligned} P(s) &= \int_0^\infty \int_0^t e^{-st} f_1(\tau) f_2(t-\tau) d\tau dt \\ &= \int_0^\infty \int_\tau^\infty e^{-st} f_1(\tau) f_2(t-\tau) dt d\tau. \end{aligned} \quad (5.106)$$

Let us now make a substitution $w = t - \tau$ in the inside integral in (5.106) that is with respect to t , so $t = w + \tau$ and $dt = dw$. The limits of integration for the dt integral are $w = 0$ to $w = \infty$ and we have

$$\begin{aligned} P(s) &= \int_0^\infty \int_0^\infty e^{-s(w+\tau)} f_1(\tau) f_2(w) dw d\tau \\ &= \int_0^\infty \int_0^\infty e^{-sw} e^{-s\tau} f_1(\tau) f_2(w) dw d\tau \\ &= \underbrace{\left(\int_0^\infty e^{-s\tau} f_1(\tau) d\tau \right)}_{F_1(s)} \underbrace{\left(\int_0^\infty e^{-sw} f_2(w) dw \right)}_{F_2(s)}, \end{aligned} \quad (5.107)$$

where in the last line we use the fact that for a double integral with separable integrand $f(x)g(y)$ we can evaluate the double integral by performing each integration separately, namely as

$$\int_a^b \int_c^d f(x)g(y) dx dy = \left(\int_c^d f(x) dx \right) \left(\int_a^b g(y) dy \right).$$

Equation (5.107) shows that $P(s) = F_1(s)F_2(s)$. Since $P(s) = \mathcal{L}(f_1 * f_2)$ we have shown that $\mathcal{L}(f_1 * f_2) = \mathcal{L}(f_1)\mathcal{L}(f_2)$, which completes the proof of Theorem 5.5.1.

5.5.6 Exercises

Exercise 5.5.1 For each pair of functions $f_1(t)$ and $f_2(t)$ defined for $t \geq 0$:

- Compute the Laplace transforms $F_1(s) = \mathcal{L}(f_1)$ and $F_2(s) = \mathcal{L}(f_2)$, then compute and simplify the product $F_1(s)F_2(s)$.
 - Compute the convolution $p(t) = (f_1 * f_2)(t)$ and the Laplace transform $P(s) = \mathcal{L}(p(t))$. Verify that $P(s) = F_1(s)F_2(s)$.
- (a) $f_1(t) = t, f_2(t) = t$
 (b) $f_1(t) = t, f_2(t) = 1$
 (c) $f_1(t) = t, f_2(t) = e^t$
 (d) $f_1(t) = e^{at}, f_2(t) = e^{bt}$, a and b constants
 (e) $f_1(t) = \sin(t), f_2(t) = \sin(t)$
 (f) $f_1(t) = t^2, f_2(t) = \delta(t)$. Take note of (5.99).
 (g) $f_1(t) = t + 3, f_2(t) = \delta(t - 2)$

Exercise 5.5.2 Use the method of Laplace transforms to solve each ODE with zero initial conditions and find the unit impulse response of the system.

- (a) $u'(t) + 4u(t) = \delta(t)$
 (b) $u'(t) - 2u(t) = \delta(t)$
 (c) $u'(t) = \delta(t)$
 (d) $u''(t) + 3u'(t) + 2u(t) = \delta(t)$
 (e) $u''(t) + u(t) = \delta(t)$
 (f) $2u''(t) + 4u'(t) + 10u(t) = \delta(t)$
 (g) $u''(t) + 4u'(t) + 4u(t) = \delta(t)$
 (h) $u''(t) = \delta(t)$

Exercise 5.5.3 Let $g(t) = \sin(t)$, $p(t) = (2t + 2)e^{-t} - 2\cos(t)$, and let $f(t)$ be an unknown function that satisfies the convolutional integral equation

$$\int_0^t g(t - \tau) f(\tau) d\tau = p(t).$$

Find f . In other words, solve the equation $f * g = p$, where g and p are known and f is unknown. This is an example of a **deconvolution** problem. Hint: use Theorem 5.5.1 to move the problem into the s -domain.

Exercise 5.5.4 A system is governed by a first-order ODE $au'(t) + bu(t) = f(t)$ for some constants a and b . When $f(t) = H(t)$ (the Heaviside function) and the initial data is $u(0) = 0$, we are given that the response of the system is $u(t) = (1 - e^{-5t})/5$. Determine a and b . (For an input $f(t) = H(t)$ the response of a system is called the **step response**.) Hint: work in the s -domain.

Exercise 5.5.5 A system is governed by a first-order ODE $au'(t) + bu(t) = f(t)$ for some constants a and b . When $f(t) = \delta(t - 3)$ and the initial data is $u(0) = 0$ the response of the system is $u(t) = H(t - 3)e^{-2(t-3)}$. Determine a and b .

Exercise 5.5.6 A spring-mass-damper system at equilibrium and at rest at time $t = 0$ is subjected to a hammer blow at time $t = 5$, with total impulse 4 newton-seconds. The system responds as $u(t) = H(t - 5)(e^{-(t-5)} - e^{-5(t-5)})$. Determine the mass, the damping constant, and the spring constant for the system.

Exercise 5.5.7 A spring-mass-damper system is at rest and at equilibrium at time $t = 0$. At time $t = 1$ the mass is subject to an impulsive blow $3\delta(t - 1)$. The position of the mass is approximately

$$u(t) \approx 1.56H(t - 1)e^{-0.092(t-1)} \sin(1.61(t - 1)).$$

Estimate m , c , and k . There may be no perfect choice for the parameters, due to the rounding to three significant figures. Report your estimates to three significant figures.

Exercise 5.5.8 The charge $q(t)$ on the capacitor in an RLC series circuit with voltage source $V(t)$ is governed by the equation

$$Lq''(t) + Rq'(t) + q(t)/C = V(t).$$

The circuit has an 8 ohm resistor, but with unknown values for L and C . Prior to time $t = 0$ the capacitor is uncharged $q(0) = 0$ and no current flows ($q'(0) = 0$). The input voltage $V(t)$ is zero for $t < 0$ and then $V(t) = 5$ for $t > 0$; think of a switch closed at $t = 0$ that completes the circuit, with a 5 volt source. The voltage across the resistor is measured and is given by $V_R(t) = 2.383e^{-285.7t} \sin(1199t)$, rounded to four significant figures. Estimate the values of L and C . The measured voltage may not be perfectly consistent with any choice for L and C , due to the rounding error; just find the best choices you can, to three significant figures. Hint: the current through the resistor is $q'(t)$, and so $v_r(t) = 8q'(t)$ by Ohm's law. Work in the s -domain.

Exercise 5.5.9 A salt tank is filled with 100 liters of pure water at time $t = 0$. After time $t = 0$ water with 0.1 kg of salt per liter begins flowing into the tank through a pipe at a rate of 2 liters per minute. The well-stirred solution exits the tank through another pipe, at 2 liters per minute. At time $t = 111$ an unknown amount A kg of salt is dumped into the tank. For $t > 111$ the concentration of salt in the exit pipe is monitored and found to be $0.1 + 0.195e^{-0.02t}$ kg per liter. Estimate the time t_0 and amount A . Hint: solve the appropriate ODE and work in the time domain.

Exercise 5.5.10 Suppose a system is governed by an ODE $ay''(t) + by'(t) + cy(t) = f$, with $y(0) = y'(0) = 0$. Assume a, b , and c are all positive constants. An unknown input $f(t)$ is applied and the response of the system is

$$y(t) = 2H(t - 2)(e^{-4(t-2)} + e^{-2(t-2)} - 2e^{-3(t-2)}).$$

Can you find choices for a, b , and c , and also the input function f that are consistent with the given information? If so, do it. More challenging: are the values for a, b, c and f uniquely determined by the given information? If not, what can be uniquely determined?

Exercise 5.5.11 Consider the unforced harmonic oscillator $mu''(t) + cu'(t) + ku(t) = 0$ with initial data $u(0) = 0$, $u'(0) = v_0$.

- (a) Show that the Laplace transform $U(s)$ of the solution can be expressed as

$$U(s) = mv_0G(s),$$

where $G(s) = 1/(ms^2 + cs + k)$ is the system transfer function.

- (b) Use (a) to show that the solution $u(t)$ is given by

$$u(t) = mv_0g_\delta(t),$$

where g_δ is the impulse response.

Exercise 5.5.12 Prove the following properties of convolution:

Commutativity: $f_1 * g = g * f_1$.

Distributivity: $(af_1 + bf_2) * g = af_1 * g + bf_2 * g$ and $g * (af_1 + bf_2) = ag * f_1 + bg * f_2$.

Associativity: $(f_1 * f_2) * g = f_1 * (f_2 * g)$.

Hint: in each case you can just appeal to Theorem 5.5.1.

Exercise 5.5.13 Suppose $f(t)$ is defined for $t \geq 0$ and of exponential order, so that for some constants M_1 and a we have $|f(t)| \leq M_1 e^{at}$ for all $t \geq 0$. Suppose that $g(t)$ is also defined for $t \geq 0$ and of exponential order, so that for some constants M_2 and b we have $|g(t)| \leq M_2 e^{bt}$ for all $t \geq 0$. Show that $f * g$ is of exponential order by showing that for any choice of $d > \max(a, b)$ there is some constant K such that $|(f * g)(t)| \leq K e^{dt}$, for all $t \geq 0$, by following these steps.

- (a) Start with the definition of convolution (5.93) and take the absolute value of both sides to conclude

$$\begin{aligned} |(f * g)(t)| &= \left| \int_0^t f(\tau)g(t - \tau) d\tau \right| \\ &\leq \int_0^t |f(\tau)||g(t - \tau)| d\tau. \end{aligned}$$

You may find it useful to recall the integral calculus fact that $\left| \int_a^b \phi(t) dt \right| \leq \int_a^b |\phi(t)| dt$.

- (b) Let $c = \max(a, b)$. Argue that since, by assumption, $|f(t)| \leq M_1 e^{at}$ and $|g(t)| \leq M_2 e^{bt}$ we must also have $|f(t)| \leq M_1 e^{ct}$ and $|g(t)| \leq M_2 e^{ct}$ for all $t \geq 0$.

- (c) Use the result of part (b) to argue that

$$|(f * g)(t)| \leq M_1 M_2 t e^{ct}$$

for $t \geq 0$.

- (d) Show that for any $d > c$ there is a constant K such that $M_1 M_2 t e^{ct} \leq K e^{dt}$ for all $t \geq 0$. Hint: this last inequality is equivalent to $M_1 M_2 t e^{(c-d)t} \leq K$; show the quantity $M_1 M_2 t e^{(c-d)t}$ attains a maximum value of $\frac{M_1 M_2}{e^{(d-c)}}$ for $t \geq 0$, and so we can take this maximum value as our choice of K . In conjunction with part (c), how does this show that $f * g$ is of exponential order?

5.6 A Taste of Control Theory

5.6.1 The Need for Control

For many of the physical systems we've modeled in this text, our goal has been to induce a certain behavior in the system or to obtain some desired outcome. For example, for the intracochlear drug delivery of Section 1.2 we want a specific dose or concentration of medication in the cochlea. For the morphine administration example of Section 5.1, we sought to maintain a certain therapeutic level of the drug in the patient's system. In Section 1.3 the desired outcome was that the fish population remain above a minimum level. For the vibration isolation table of Section 4.1, we wanted the tabletop to remain motionless. Situations like these abound in engineering and science, and life more generally. You have a thermostat to maintain the temperature of your house or apartment at a specified level, right?

In many cases a quantitative ODE model is merely a prelude to understanding how to steer the system to a desired end by adjusting an input to the system. In the intracochlear drug delivery model we can control the rate at which the drug is delivered, and we have similar control in the morphine administration problem. For the fish population model we can control the rate at which fish are harvested. The vibration isolation table model lacks any obvious control, but by the addition of an actuator that can exert specified forces, we gain the possibility of countering floor vibration by the appropriate use of the actuator, something that we will consider in the projects at the end of this section.

The appropriate control to achieve these desired ends is not always obvious. The analysis necessary to achieve these end leads to the aptly-named subject of **control theory**, which addresses the problem of effectively controlling the types of systems we've been considering. Control theory is a huge subject with many different techniques. In this section we'll focus on a small portion of the theory concerning the control of a system modeled by a linear, constant-coefficient differential equation using **proportional-integral-derivative** (or **PID**) control techniques. The goal here is not a thorough treatment of PID control, but rather a series of illustrative examples that show how Laplace transform techniques really begin to pull their weight in this common and important application.

Reading Exercise 5.6.1 Look back at some of the other systems we've modeled in this text. Is there an obvious desired outcome for each system? What control could we exert on the system to achieve this end?

5.6.2 Modeling an Incubator

Incubators are common in laboratories that grow or maintain biological materials like bacteria, plants, or even animals. Incubators are used to maintain specimens at carefully controlled temperatures in ambient environments that may vary in unpredictable ways.

Consider a typical tabletop incubator that consists of an insulated cabinet to hold specimens. The incubator also includes a heat source, and may also have cooling ability, which we assume is the case for this example. A typical such tabletop incubator is shown in Figure 5.16

Let $y(t)$ denote the temperature inside the incubator, where time t is measured in hours and temperature in degrees Celsius; assume this temperature is uniform throughout the interior of the incubator (which has a circulation fan). Let $a(t)$ denote the ambient temperature outside the incubator cabinet. Suppose that in the absence of any active heating or cooling element, the incubator temperature would change according to Newton's law of cooling,

$$y'(t) = -k(y(t) - a(t)), \quad (5.108)$$

for some positive constant k , which has units of reciprocal hours. The closer k is to zero, the better insulated the incubator is from the outside environment. The most desirable situation would be $k = 0$, but this is unattainable.



Figure 5.16: A tabletop incubator.

The incubator has a heating/cooling element that can be used to control the incubator's interior temperature, so let us now account for this input. Suppose that this element is controlled according to some function $u(t)$. For example, $u(t)$ could be a voltage that allows us to adjust the intensity of the heating or cooling. Here $u(t)$ is called the **control function**. We will assume that the rate at which heat energy is added to or extracted from the incubator is proportional to $u(t)$, so that if $u(t) > 0$, the heating element is pumping heat energy into the cabinet at a rate $J(t) = K_1 u(t)$ joules per hour for some positive constant K_1 , or removing heat if $u(t) < 0$. The constant K_1 depends on how $u(t)$ is normalized. For example, it might be that $u(t) = 1$ corresponds to $J(t) = 1000$ joules per hour, which is about 0.278 watts.

A reasonable and typical model for how the heat energy input affects the temperature $y(t)$ is based on the assumption that if heat energy is pumped into the incubator cabinet at a rate of $J(t)$ joules per hour then this raises the temperature of the interior of the incubator at a rate proportional to $J(t)$, in the absence of any heat losses elsewhere. That is, $y'(t) = K_2 J(t) = K_1 K_2 u(t)$ for some positive constant K_2 , if no heat is lost elsewhere. If we also account for the change in $y(t)$ due to the heat lost to the environment, quantified as $-k(y(t) - a(t))$ in (5.108), then we arrive at a model

$$y'(t) = -k(y(t) - a(t)) + Ku(t), \quad (5.109)$$

where $K = K_1 K_2$ is some positive constant. Equation (5.109) is a common variation of Newton's law of cooling that incorporates a heat source or sink. With initial condition $y(0) = y_0$, (5.109) has a unique solution.

For any incubator the constants k and K could be estimated from data using techniques from Section 3.4. Thus k and K will be considered as known constants, at least for now. For the moment we'll assume that $u(t)$ can take any real value, so that the heating and cooling element can supply any rate of heating or cooling. The goal here is to choose $u(t)$ so that the interior of the incubator tracks a desired target temperature $r(t)$ degrees. In the language of control theory, $r(t)$ is called the

setpoint or reference signal and $y(t)$ is the **process variable**.

In what follows we will view $u(t)$ as the input to a system and $y(t)$ as the output, in the spirit of Section 5.5.2.

5.6.3 Open-Loop Control

Let us consider an approach to controlling the incubator temperature that's based on assumptions that, as we'll see later, leave something to be desired. Assume we have accurate values for the constants k and K in the model (5.109), and that $a(t)$ is known. The assumption concerning $a(t)$ isn't totally absurd if the incubator sits in a laboratory that is kept at a constant or otherwise controlled temperature. For mathematical clarity, let's start with the assumption that $a(t)$ is constant and even more, that $a(t) = 0$. In this case $y(t)$ may be interpreted as the incubator's temperature relative to the ambient environment, while $r(t)$ is the desired temperature profile over time, relative to the ambient temperature. With these assumptions the ODE (5.109) that governs the incubator temperature becomes

$$y'(t) = -ky(t) + Ku(t). \quad (5.110)$$

The goal is to choose the control function $u(t)$ so that $y(t) = r(t)$ for all $t \geq 0$. This means we must take initial condition $y(0) = y_0 = r(0)$, so the initial temperature is dictated by $r(t)$. If this condition isn't met the temperature $y(t)$ will still approach $r(t)$ asymptotically, as will be shown. As one further simplifying assumption, let us assume that $r(0) = 0$. That is, the initial desired temperature is the ambient temperature; of course over time $r(t)$ can change to any desired temperature.

The Open-Loop Control Solution

It's straightforward to choose the control $u(t)$ so that $y(t) = r(t)$ for all $t \geq 0$. This can be done in the time domain or the Laplace s -domain. The time domain solution is left for Exercise 5.6.1. To get into the spirit of how Laplace transforms can be used in control theory, let's approach the problem in the s -domain.

Laplace transforming both sides of (5.110) and using $y(0) = 0$ shows that $sY(s) = -kY(s) + KU(s)$, where $Y = \mathcal{L}(y(t))$ and $U = \mathcal{L}(u(t))$. We can then solve for $Y(s)$ as

$$Y(s) = G_p(s)U(s), \quad (5.111)$$

where

$$G_p(s) = \frac{K}{s+k}. \quad (5.112)$$

Here $G_p(s)$ is the transfer function in the s -domain model (5.111) that governs how the input control $u(t)$ is turned into the output process variable $y(t)$, the incubator temperature.

To obtain $y(t) = r(t)$ for all $t \geq 0$ we need $Y(s) = R(s)$, where $R(s) = \mathcal{L}(r(t))$. Substituting $Y(s) = R(s)$ into (5.111) shows that $U(s)$ must satisfy

$$U(s) = G_c(s)R(s), \quad (5.113)$$

where

$$G_c(s) = \frac{1}{G_p(s)} = \frac{s+k}{K}. \quad (5.114)$$

Equation (5.113), along with (5.114), is the complete prescription, in the s -domain, for how to turn the setpoint function $r(t)$ into the control function $u(t)$ that will give the desired incubator temperature $y(t) = r(t)$. For a given $R(s)$ we can use (5.113) to compute $U(s)$, then inverse transform to find $u(t)$. See Examples 5.39 and 5.40.

Control Block Diagram

The entire computation (5.111) to (5.114) can be neatly summarized with a block diagram as shown in Figure 5.17, which depicts the control process in the s -domain. In the language of control theory, the physical system we want to control is called the **plant**. The plant here is the incubator. In the time domain the input to the controller-plant system is the setpoint $r(t)$ and the output is the process variable $y(t)$, but in the s -domain the input is $R(s)$, and the output is $Y(s)$. The transfer function for the entire process, in our case (5.111) to (5.114), is the product $G_c(s)G_p(s)$. Since $G_c(s)G_p(s) = 1$ here (this follows from (5.114)) we have $Y(s) = R(s)$, and so $y(t) = r(t)$ as desired. In the time domain the mapping from $r(t)$ to $u(t)$ or $u(t)$ to $y(t)$ would be quantified by a convolution, but in the s -domain they are simple function multiplications.

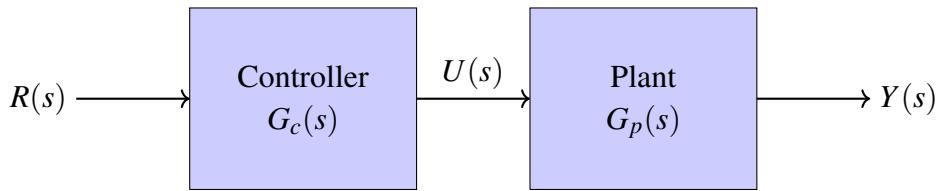


Figure 5.17: An s -domain block diagram for open-loop control.

Open-Loop Control Examples

■ **Example 5.39** Suppose that for a particular incubator the parameter values are $k = 0.05$ (reciprocal hours), $K_1 = 1000$ (units joules per hour per volt, if the control function is a voltage), and $K_2 = 0.0005$ (units degrees per joule), so that $K = K_1K_2 = 0.5$ (units degrees per hour per volt). We choose a setpoint temperature of $r(t) = 10(1 - e^{-t})$ degrees Celsius in the incubator, relative to the lab ambient temperature. From (5.114) we find

$$G_c(s) = \frac{s + 0.05}{0.5} = 2s + 0.1.$$

A Laplace transform of $r(t)$ shows that $R(s) = \frac{10}{s(s+1)}$, and therefore from (5.113) we get $U(s) = \frac{10(2s+0.1)}{s(s+1)}$. An inverse transform shows that the appropriate control function is

$$u(t) = 1 + 19e^{-t}.$$

Under these conditions the control function $u(t)$ will asymptotically raise the temperature $y(t)$ from 0 to 10 degrees above ambient, as $y(t) = 10(1 - e^{-t})$. ■

Reading Exercise 5.6.2 Solve (5.110) with $u(t) = 1 + 19e^{-t}$ and constants as in Example 5.39, with $y(0) = 0$, and verify that $y(t) = 10(1 - e^{-t})$.

Reading Exercise 5.6.3 Solve (5.110) with $u(t) = 1 + 19e^{-t}$ and constants as in Example 5.39, but with general initial condition $y(0) = y_0$. Show that the solution approaches 10 degrees for any value of y_0 .

■ **Example 5.40** Suppose that in the setting of Example 5.39 the desired temperature is $r(t) = 5 \sin(2\pi t/24)$, which oscillates between -5 and 5 degrees relative to the ambient temperature with a 24 hour cycle, perhaps to emulate a daily temperature rhythm for the incubator specimens. A Laplace transform shows that

$$R(s) = \frac{60\pi}{\pi^2 + 144s^2},$$

and from (5.111) and (5.114) it follows that

$$U(s) = \frac{60\pi(2s+0.1)}{\pi^2 + 144s^2}.$$

An inverse transform yields the control function

$$u(t) = \frac{5\pi}{6} \cos(\pi t/12) + \frac{1}{2} \sin(\pi t/12).$$

With this control function the solution to (5.110) with $y(0) = 0$ is $y(t) = r(t)$.

Even if $y(0) \neq 0$, the solution $y(t)$ asymptotically approaches $r(t)$; see Reading Exercise 5.6.4 and Figure 5.18, in which we show the solution $y(t)$ with $y(0) = r(0) = 0$ (so $y(t) = r(t)$), the solution $y(t)$ with $y(0) = -6$, and the solution $y(t)$ with $y(0) = 20$. No matter what initial condition we start with, if we wait long enough the controlled temperature profile approaches the target curve, although in Figure 5.18 it takes the better part of 100 hours. ■

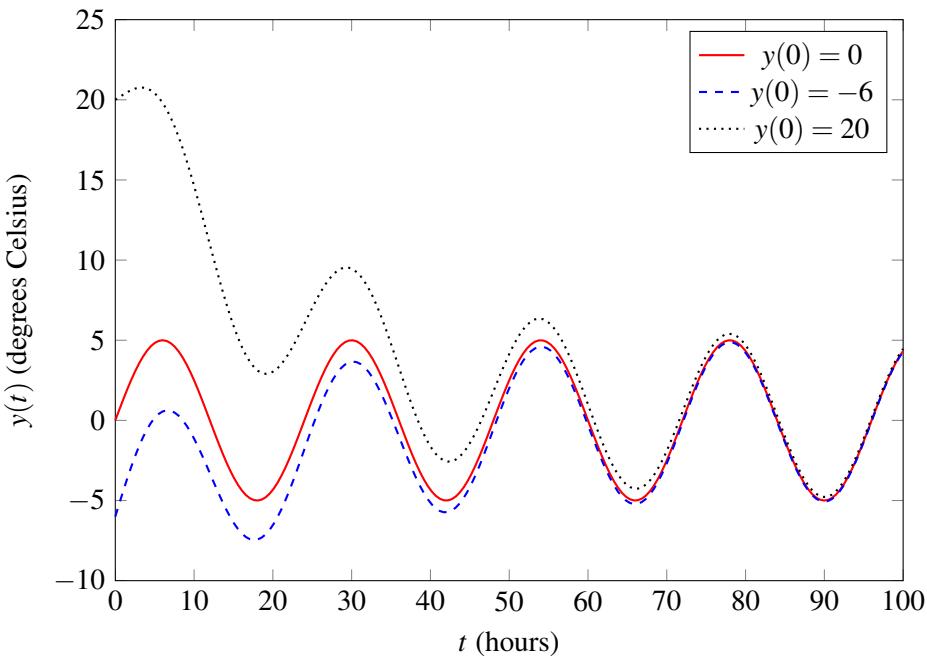


Figure 5.18: Temperature profiles for open-loop control applied to incubator governed by (5.110) with setpoint $r(t) = 5 \sin(2\pi t/24)$ and initial conditions $y(0) = 0$ (solid red), $y(0) = -6$ (dashed blue), $y(0) = 20$ (dotted black).

Reading Exercise 5.6.4 Solve (5.110) with $u(t) = \frac{5\pi}{6} \cos(\pi t/12) + \frac{1}{2} \sin(\pi t/12)$ and constants as in Example 5.40 with general initial condition $y(0) = y_0$. Show that the solution approaches $r(t) = 5 \sin(2\pi t/24)$ as $t \rightarrow \infty$ for any choice of y_0 .

The control methodology outlined here is an example of **open-loop control**. We begin with an accurate system model, so the constants k and $K = K_1 K_2$ must be known, because as (5.114) illustrates, they are needed to compute the control function. The ambient temperature $a(t)$ must also be known, for our interpretation of $r(t)$ is relative to $a(t)$. In this case we can design a control function $u(t)$ so that the system tracks the desired temperature profile $r(t)$. And even if the system starts off with an arbitrary initial temperature the system will still asymptotically approach the desired temperature profile $r(t)$, as Reading Exercise 5.6.4 illustrates. In this open-loop strategy the control function $u(t)$ is fixed for all t , right at the start. If all of our assumptions hold and there are no surprises, the control $u(t)$ will give us what we want.

Shortcomings

Things don't always go according to plan. What if a careless student leaves the window in the lab open and the ambient lab temperature $a(t)$ falls 5 degrees? An open-loop controller doesn't adapt to this, but goes merrily on with the assumption that $a(t)$ hasn't changed. Moreover, there's no guarantee that the values for k , K_1 , and K_2 will remain constant. As anyone who has owned a refrigerator knows, door seals eventually leak, and this would be expected to increase the cooling constant k . The constant K_1 that quantifies the amount of heat generated by the heating element as a function of control voltage $u(t)$ may change over time as the heater ages, corrodes, or gets dusty. And K_2 , which quantifies how the incubator temperature changes in response to a given amount of input thermal energy, will depend on how full the incubator is. If it's packed with specimens, then more thermal energy should be needed to change the temperature by a given amount than when the incubator is empty, thereby increasing K_2 . Even the simple act of opening the incubator door will upset the delicate temperature balance. In control theory events like these are called **disturbances**. The open-loop strategy cannot adapt to these disturbances, so some poor person will be endlessly fiddling with a knob to maintain the proper temperature.

Reading Exercise 5.6.5 Consider the situation of Example 5.39 with the same constants k , K_1 , and K_2 , with ambient temperature $a(t) = 0$. Suppose the desired temperature setpoint is $r(t) = 10 - 10e^{-t}$. The open-loop control $u(t) = 1 + 19e^{-t}$ in that example was designed under the assumption that $K_2 = 0.0005$. But now the incubator has been filled with specimens and $K_2 = 0.0004$, giving $K = K_1 K_2 = 0.4$. Solve the controlled ODE

$$y'(t) = -ky(t) + 0.4u(t),$$

with $y(0) = 0$ and $u(t) = 1 + 19e^{-t}$. Compare the solution to the desired setpoint temperature $r(t)$. How well does the incubator temperature $y(t)$ track the desired setpoint $r(t)$?

5.6.4 Closed-Loop Control

The goal in this section is to develop a more robust control algorithm that can adapt to disturbances in the environment or in the system itself. To do this we will use **closed-loop control**, in which the output process variable of the system is monitored and that information is used to adjust the control function $u(t)$. For example, with the incubator we will monitor the interior temperature $y(t)$ and use that to control the heater. Closed-loop control is also known as **feedback control**, for the output of the system is fed back into the control strategy to change $u(t)$ when we aren't hitting the target. For a truly robust control strategy the function $u(t)$ should not depend on the constants k , K_1 , or K_2 , since these may not be known precisely, and they may change over time. The control function should not depend on the possibly unpredictable ambient temperature either.

Proportional Control

Let's consider the simplest form of closed-loop control, called **proportional control**. If the incubator temperature is too cold, then we should take $u(t) > 0$ to add heat energy; if the incubator temperature is too warm, then we should take $u(t) < 0$ to extract heat energy. One way to do this is to set

$$u(t) = K_p e(t) \tag{5.115}$$

in (5.110), where $e(t) = r(t) - y(t)$ is the error between the setpoint and current temperature and K_p is a positive constant, called the **gain** for the controller, or the **proportional gain**, since this is proportional control. This is a form of feedback control, for the current temperature $y(t)$, in relation to the desired temperature $r(t)$, dictates what the heating element should do. The control function $u(t)$ depends on the desired temperature $r(t)$, the current temperature $y(t)$, and nothing else. This is

desirable. This control function might continue to work well even if the parameters k , K , or $a(t)$ change or are unknown.

Reading Exercise 5.6.6 Consider the case in which $a(t) = 0$, and so $y(t)$ satisfies (5.110). Suppose $u(t)$ is given by (5.115). Show that

$$y'(t) = -(k + KK_p)y(t) + KK_p r(t). \quad (5.116)$$

Then take $k = 0.05$, $K = 0.5$, and $K_p = 1$ with setpoint $r(t) = 10 - 10e^{-t}$. Solve the ODE (5.116) with $y(0) = 0$ and plot $y(t)$ and $r(t)$ for $0 \leq t \leq 20$. The desired temperature $r(t)$ approaches 10. What does $y(t)$ approach? Experiment with larger or smaller values for K_p . How do they affect the performance of this control?

Reading Exercise 5.6.7 Repeat Reading Exercise 5.6.6 with $k = 0.05$, $K = 0.5$, and $K_p = 1$, and change $r(t)$ to $r(t) = 5 \sin(2\pi t/24)$. Plot and compare $r(t)$ to $y(t)$ over the time interval $0 \leq t \leq 200$. Does $y(t)$ approach $r(t)$ as $t \rightarrow \infty$?

Proportional Control Analysis in the s -Domain

The constant K_p in (5.115) is the only flexibility available in proportional control. How should K_p be chosen? As will become clear, for this type of controller no choice of K_p will yield $y(t) = r(t)$. The analysis of feedback control in general is much easier if we work in the s -domain, where important operations become simple function multiplication, rather than the time domain where these operations are convolutions. Figure 5.19 gives a view of the general flow of the feedback control process in the s -domain.

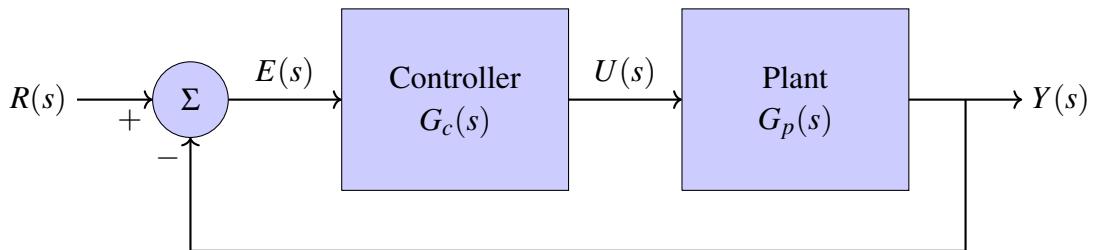


Figure 5.19: An s -domain block diagram for feedback or closed-loop control.

Our first goal is to explicitly determine how the setpoint $r(t)$ is transformed into the process variable $y(t)$ and how this process depends on the parameters k , K , and K_p . This will allow us to choose K_p so that $y(t)$ is close to $r(t)$, the desired outcome. As noted above, this analysis is easier in the s -domain, so we work with $R(s) = \mathcal{L}(r(t))$ instead of $r(t)$ and $Y(s) = \mathcal{L}(y(t))$ instead of $y(t)$. We will find a simple and explicit relation between $R(s)$ and $Y(s)$.

We begin with the plant transfer function, which has not changed since our open-loop analysis, and so the relation $Y(s) = G_p(s)U(s)$ with $G_p(s) = K/(s+k)$ as in (5.112) still holds, assuming $y(0) = 0$. This is embodied by the rectangle labeled “Plant” on the right in Figure 5.19, with $U(s)$ as the input to the plant and $Y(s)$ as the output.

For the controller transfer function use (5.115) to compute

$$U(s) = G_c(s)E(s), \quad (5.117)$$

where $E(s) = \mathcal{L}(e(t))$ and

$$G_c(s) = K_p,$$

which is the controller transfer function, a constant function in this case. Equation (5.117) relates the controller input $e(t)$ to the controller output $u(t)$, but in the s -domain. This relationship is embodied by the left rectangle labelled “Controller” in Figure 5.19.

The output or process variable $Y(s)$ from the plant is fed back to the start and combined with $R(s)$ to form $E(s) = R(s) - Y(s) = \mathcal{L}(r(t) - y(t))$; this operation occurs at the circle labeled “ Σ .” In the s -domain the transform $E(s)$ becomes the input to the controller and $U(s)$ the output. Figure 5.19 should make it clear why this is called feedback control or closed-loop control, as the path from $Y(s)$ back to Σ closes the control loop. Contrast this to Figure 5.17.

To find the transfer function that maps $R(s)$ to $Y(s)$ note that

$$\begin{aligned} Y(s) &= G_p(s)U(s) \\ &= G_p(s)G_c(s)E(s) \\ &= G_p(s)G_c(s)(R(s) - Y(s)). \end{aligned} \quad (5.118)$$

Solving (5.118) for $Y(s)$ yields

$$Y(s) = \underbrace{\frac{G_p(s)G_c(s)}{1 + G_p(s)G_c(s)}}_{G(s)} R(s). \quad (5.119)$$

In (5.119) the function $G(s)$ is the **closed-loop transfer function** for the whole system that specifies in the s -domain how an input setpoint $r(t)$ is transformed into temperature output $y(t)$.

■ **Example 5.41** Consider the situation of Reading Exercise 5.6.6 in which $y(t)$ satisfies (5.110) with $k = 0.05$, $K = 0.5$, $u(t)$ is given by (5.115) with $K_p = 1$, and $r(t) = 10 - 10e^{-t}$. In this case the plant transfer function is $G_p(s) = 0.5/(s + 0.05)$ and the control transfer function is $G_c(s) = 1$. From (5.119) it follows that $Y(s) = G(s)R(s)$ where

$$G(s) = \frac{10}{20s + 11}$$

after simplifying. Since $R(s) = \mathcal{L}(r(t)) = \frac{10}{s(s+1)}$ we compute that

$$Y(s) = \frac{100}{s(s+1)(20s+11)}.$$

An inverse Laplace transform shows that in the time domain the incubator temperature is

$$y(t) = \frac{100}{11} + \frac{100}{9}e^{-t} - \frac{2000}{99}e^{-11t/20}. \quad (5.120)$$

The setpoint $r(t)$ and temperature $y(t)$ are plotted in Figure 5.20. For comparison we also plot the controlled temperature obtain by taking $K_p = 5$ and $K_p = 0.1$. ■

Observations on Proportional Control

For proportional control with controller transfer function $G_c(s) = K_p$ and incubator transfer function $G_p(s) = K/(s+k)$ we can use (5.119) to compute that in general the **closed-loop transfer function** is

$$G(s) = \frac{KK_p}{KK_p + s + k}. \quad (5.121)$$

Since $Y(s) = G(s)R(s)$ and the goal is $Y(s) = R(s)$ for all $s \geq 0$, it would be ideal if $G(s) = 1$ for all $s \geq 0$. Unfortunately, no choice for K_p makes this work since the denominator for $G(s)$ is always strictly larger than the numerator if $k > 0$. However, for any fixed value of s , $\lim_{K_p \rightarrow \infty} G(s) = 1$, so larger values of K_p should give better results, an observation that is supported by the graphs in Figure 5.20. The main issue with large values for K_p is that since the control is $u(t) = K_p(r(t) - y(t))$,

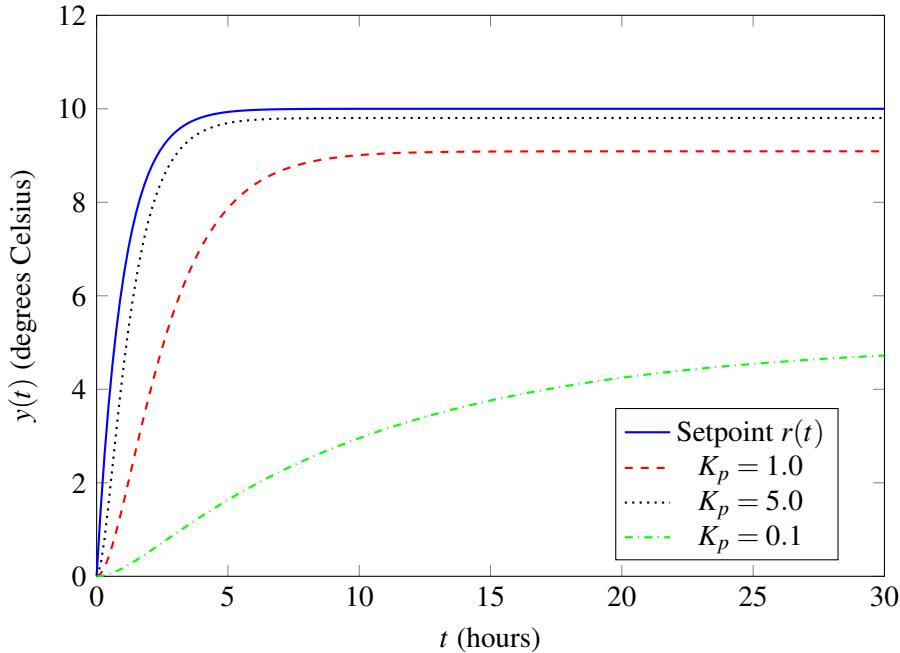


Figure 5.20: Incubator setpoint temperature (solid blue curve) with proportional closed-loop control for temperature, with proportional gains $K_p = 1$, $K_p = 5$, and $K_p = 0.1$.

even small differences $r(t) - y(t)$ may result in unrealistically large control signals. We can't make K_p arbitrarily large.

Even though no choice of K_p is perfect, we might hope for something more limited in scope, such as $\lim_{t \rightarrow \infty} (r(t) - y(t)) = 0$, especially in the case that the setpoint $r(t)$ is constant or approaches a constant. Such an $r(t)$ would correspond to the case in which the incubator specimens are to be held at constant temperature. Proportional control cannot achieve even this limited goal.

To see why suppose that

$$\lim_{t \rightarrow \infty} r(t) = r_0$$

for some constant r_0 . From the final value theorem (Theorem 5.2.5, with appropriate hypotheses checked) this yields

$$\lim_{s \rightarrow 0^+} sR(s) = r_0.$$

From (5.121) and $Y(s) = G(s)R(s)$ we find

$$\begin{aligned} \lim_{s \rightarrow 0^+} sY(s) &= \lim_{s \rightarrow 0^+} sG(s)R(s) \\ &= \left(\lim_{s \rightarrow 0^+} G(s) \right) \left(\lim_{s \rightarrow 0^+} sR(s) \right) \\ &= \left(\frac{KK_p}{KK_p + k} \right) r_0, \end{aligned} \tag{5.122}$$

and another application of Theorem 5.2.5 gives

$$\begin{aligned}\lim_{t \rightarrow \infty} y(t) &= \left(\frac{KK_p}{KK_p + k} \right) r_0 \\ &= r_0 - \underbrace{\left(\frac{k}{KK_p + k} \right) r_0}_{\delta}.\end{aligned}\tag{5.123}$$

Since $k > 0$ we see from (5.123) that $\delta > 0$ and so proportional control (5.115) will not give us what we want, namely $\lim_{t \rightarrow \infty} y(t) = r_0$, even in the simplest case in which $r(t)$ is constant (unless $r(t) = 0$). We can only make $y(t)$ close to $r(t)$ by choosing K_p to be as large as is practical, which makes δ close to 0. We saw this in Example 5.41, especially in Figure 5.20, and you may have noticed it in Reading Exercise 5.6.6.

Reading Exercise 5.6.8 In the incubator example with $k = 0.05$, $K = 0.5$, and $r_0 = 10$, how large must $K_p > 0$ be to ensure that $y(t)$ stabilizes at a value within 0.1 degree of r_0 ?

Reading Exercise 5.6.9 Redo Example 5.41 but with proportional gain $K_p = -1$. What is $y(t)$? In plain English, what is this control strategy, and why does $y(t)$ make sense in this case?

5.6.5 Proportional-Integral Control

In **proportional-integral** control (**PI** control), the control function in (5.115) is modified as

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau\tag{5.124}$$

where $e(t) = r(t) - y(t)$ as before, and K_p and K_i are specified constants, called the **proportional gain** and the **integral gain**. PI control augments simple proportional control with an additional term that becomes increasingly vigorous if the error $e(t)$ in the response persists. For example, in the case of the incubator with $a(t) = 0$ and $r(t) \rightarrow r_0$, we had an error δr_0 given by (5.123), an error that could not be eliminated by proportional control. With PI control this error is integrated over time, with the effect of gradually increasing the control function in an attempt to diminish the error.

The PI Closed-Loop Transfer Function

The block diagram of Figure 5.19 is applicable here as well. The only change between proportional and PI control is to the function $G_c(s)$, which then changes the closed-loop transfer function $G(s)$ defined in (5.119). After accounting for that change, precisely the same computation of (5.118) is applicable to again show that $Y(s) = G(s)R(s)$. Let's work out $G_c(s)$ for PI control and then $G(s)$ for this specific PI controller.

Laplace transforming (5.124) yields (recall (5.95)-(5.96) or see Exercise 5.2.18)

$$G_c(s) = K_p + \frac{K_i}{s}\tag{5.125}$$

for the transfer function for the PI controller. Use (5.125) and $G_p(s) = K/(s + k)$ in $G(s)$ from (5.119) to find that

$$G(s) = \frac{K(K_p s + K_i)}{s^2 + (KK_p + k)s + KK_i}\tag{5.126}$$

is the closed-loop transfer function from $R(s)$ to $Y(s)$.

■ **Example 5.42** Let's revisit the situation of Example 5.41 with $k = 0.05$, $K = 0.5$ and setpoint $r(t) = 10 - 10e^{-t}$. There we used proportional control with $u(t) = K_p e(t)$ and $K_p = 1$. Let us now try PI control with $K_p = 1$ and $K_i = 0.1$. The control function is thus

$$u(t) = e(t) + 0.1 \int_0^t e(\tau) d\tau.$$

The goal is to work out $y(t)$ for the resulting controlled ODE (5.110) and examine how well $y(t)$ tracks the setpoint $r(t)$.

This is most easily done with Laplace transforms. We have $Y(s) = G(s)R(s)$. With the constants in (5.126) we find

$$G(s) = \frac{0.5s + 0.05}{s^2 + 0.55s + 0.05}.$$

From $R(s) = \mathcal{L}(r(t)) = \frac{10}{s(s+1)}$ we then find that

$$Y(s) = G(s)R(s) = \frac{100s + 10}{s(s+1)(20s^2 + 11s + 1)}$$

after simplification. Inverse Laplace transforming $Y(s)$ shows that

$$y(t) \approx 10 + 9e^{-t} + 2.29e^{-0.115t} - 21.29e^{-0.435t}.$$

The function $y(t)$ is plotted in Figure 5.21 as the dotted black curve, with the desired setpoint $r(t)$ as solid blue curve and the solution (5.120) with proportional control ($K_p = 1$) from Example 5.41 as the dashed red curve, for reference. Compare the solution with PI control to that with proportional control; the PI-controlled solution approaches the correct value, although it overshoots initially. Would a different choice for K_p or K_i give better results? Adjusting the control gains K_p and K_i is known as **tuning** the controller, something we'll discuss shortly. ■

Observations on PI Control

Since the numerator for G in (5.126) is linear in s and the denominator is always quadratic, we cannot obtain $G(s) = 1$ for all s for any choice of the constants K_p and K_i , and so cannot arrange $Y(s) = R(s)$, nor $y(t) = r(t)$. However, PI control has certain advantages over proportional control. First, note that with $G(s)$ as in (5.126) we have

$$\lim_{s \rightarrow 0^+} G(s) = \lim_{s \rightarrow 0^+} \left(\frac{K(K_p s + K_i)}{s^2 + (KK_p + k)s + KK_i} \right) = 1 \quad (5.127)$$

if $K_i \neq 0$. A computation similar to that of (5.122) to (5.123) shows that

$$\begin{aligned} \lim_{s \rightarrow 0^+} sY(s) &= \lim_{s \rightarrow 0^+} sG(s)R(s) \\ &= \left(\lim_{s \rightarrow 0^+} sR(s) \right) \left(\lim_{s \rightarrow 0^+} G(s) \right) \\ &= \left(\lim_{s \rightarrow 0^+} sR(s) \right). \end{aligned} \quad (5.128)$$

If $R(s)$ satisfies the hypotheses of Theorem 5.2.5, we can conclude that if $\lim_{t \rightarrow \infty} r(t) = r_0$ then

$$r_0 = \lim_{s \rightarrow 0^+} sR(s) = \lim_{s \rightarrow 0^+} sY(s) = \lim_{t \rightarrow \infty} y(t).$$

The rightmost limit above shows that the controller will asymptotically drive $y(t)$ to the correct value r_0 , for any choice of K_p and K_i , an advantage over straight proportional control. This is illustrated in Figure 5.21 in which the PI-controlled temperature approaches 10 and the proportional-controlled temperature does not.

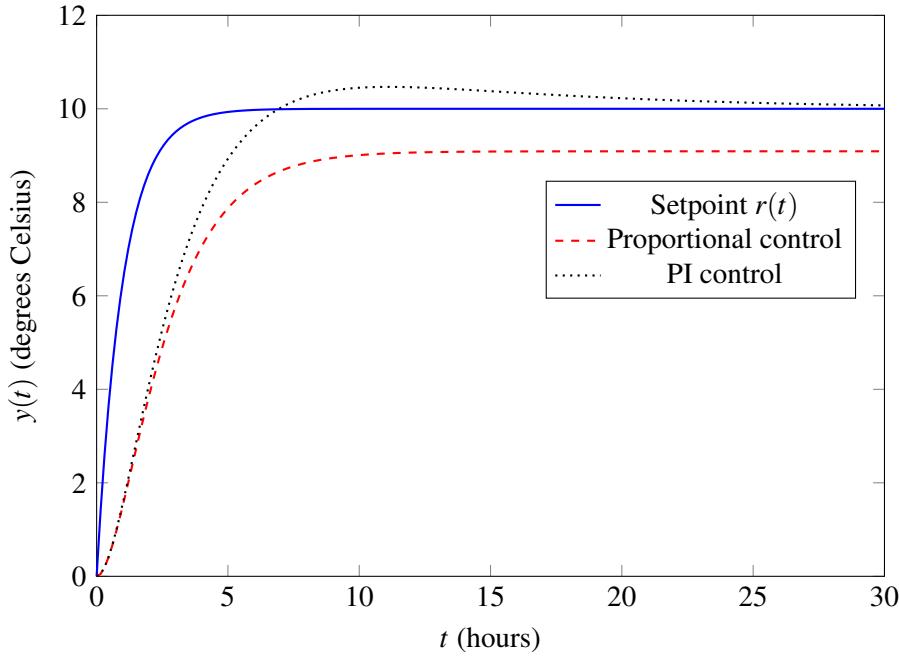


Figure 5.21: Incubator temperature with PI control ($K_p = 1, K_i = 0.1$, dotted black curve) and desired temperature setpoint (solid blue). Solution (5.120) with proportional control ($K_p = 1$) from Example 5.41 shown as dashed red curve.

Reading Exercise 5.6.10 Redo Example 5.42 with integral control gain $K_i = 1$. In particular, find $y(t)$ and plot it on the range $0 \leq t \leq 30$. How does the solution behave? Does $y(t)$ still approach 10? Can you explain the behavior of the solution in terms of the poles of $G(s)$?

5.6.6 Proportional-Integral-Derivative Control

In Proportional-Integral-Derivative Control (known as **PID control**) the control function $u(t)$ is computed as

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d e'(t) \quad (5.129)$$

for constants K_p, K_i , and K_d , known as the **proportional**, **integral**, and **derivative** gains, respectively. The $K_d e'(t)$ term incorporates the rate at which the error is changing into the control scheme. Compare (5.129) to (5.115) and (5.124). By taking any of the constants in (5.129) to be zero we can obtain many types of controllers, e.g., integration-only, proportional-derivative, and so on.

The feedback control analysis for the transfer function from the setpoint transform $R(s)$ to the process variable transform $Y(s)$ based on Figure 5.19 and (5.119) still holds. For the incubator model we still have $G_p(s) = K/(s+k)$. Based on (5.129) we find that $U(s) = G_c(s)E(s)$ (again, recall (5.95) and (5.96) or Exercise 5.2.18), where

$$G_c(s) = K_p + \frac{K_i}{s} + K_d s \quad (5.130)$$

is the transfer function for the PID controller. Based on (5.119) the closed-loop transfer function is

$$G(s) = \frac{K(K_d s^2 + K_p s + K_i)}{(1 + K K_d)s^2 + (K K_p + k)s + K K_i}. \quad (5.131)$$

With PID control there are three parameters at our disposal, K_p , K_i , and K_d . For any choice of these parameters, however,

$$\lim_{s \rightarrow 0^+} G(s) = 1 \quad (5.132)$$

provided $KK_i \neq 0$. Therefore, the same computations as we performed in (5.127) and (5.128) show that if the setpoint $r(t)$ limits to a value r_0 , the temperature will do the same.

Reading Exercise 5.6.11 Verify (5.132).

■ **Example 5.43** Let's revisit the situation of Example 5.42 (recall also Example 5.41) in which $y(t)$ satisfies $y'(t) = -ky(t) + Ku(t)$ (equation (5.110)) with $k = 0.05$, $K = 0.5$, and setpoint $r(t) = 10 - 10e^{-t}$. In Example 5.42 we used PI control with $K_p = 1$ and $K_i = 0.1$. Let us now try PID control with the same choices for K_p and K_i and with $K_d = 1$. The control function $u(t)$ is

$$u(t) = e(t) + 0.1 \int_0^t e(\tau) d\tau + e'(t).$$

The goal is to determine the time domain response $y(t)$ for the resulting controlled ODE (5.110).

As before, this is most easily done with Laplace transforms. From (5.131) and the choices for the gains K_p , K_i , and K_d we find

$$G(s) = \frac{10s^2 + 10s + 1}{(5s + 1)(6s + 1)}.$$

Using $Y(s) = G(s)R(s)$ and $R(s) = \mathcal{L}(r(t)) = \frac{10}{s(s+1)}$ we can solve for $Y(s)$ to find

$$Y(s) = \frac{100s^2 + 100s + 10}{s(s+1)(5s+1)(6s+1)}$$

after simplification. An inverse Laplace transform shows that

$$y(t) = 10 + 28e^{-t/6} - e^{-t}/2 - 75e^{-t/5}/2.$$

The function $y(t)$ plotted in Figure 5.22 as the dashed red curve, with the desired temperature $r(t)$ in solid blue and the temperature obtained from PI control from Figure 5.21 as the dotted black curve. As in Figure 5.21, the PID-controlled solution also approaches the correct value. PID control here results in a smaller overshoot of the desired temperature, but both controllers seem to yield quite similar results. As with PI control, we might wonder whether a different choice for K_p , K_i , or K_d would give better results. This again leads to the topic of tuning the control, to be discussed. ■

5.6.7 Disturbances

An important aspect of controller design is how the system responds to unexpected changes in the system or environment, which are called **disturbances**, or how the controller responds to setpoint changes, for example, turning up the temperature in an incubator. In the next example we consider the responses of proportional control, PI control, and PID control to an abrupt change in the ambient temperature or setpoint temperature for the incubator.

■ **Example 5.44** Let's consider the incubator governed by Newton's law of cooling as in previous examples but in which the ambient temperature may not be constant. The controlled ODE is

$$y'(t) = -k(y(t) - a(t)) + Ku(t) \quad (5.133)$$

with possibly nonzero initial temperature $y(t) = y_0$. The PID control function $u(t)$ will take the form (5.130) where $e(t) = r(t) - y(t)$. For simplicity let us take $K = 1$. We want the incubator

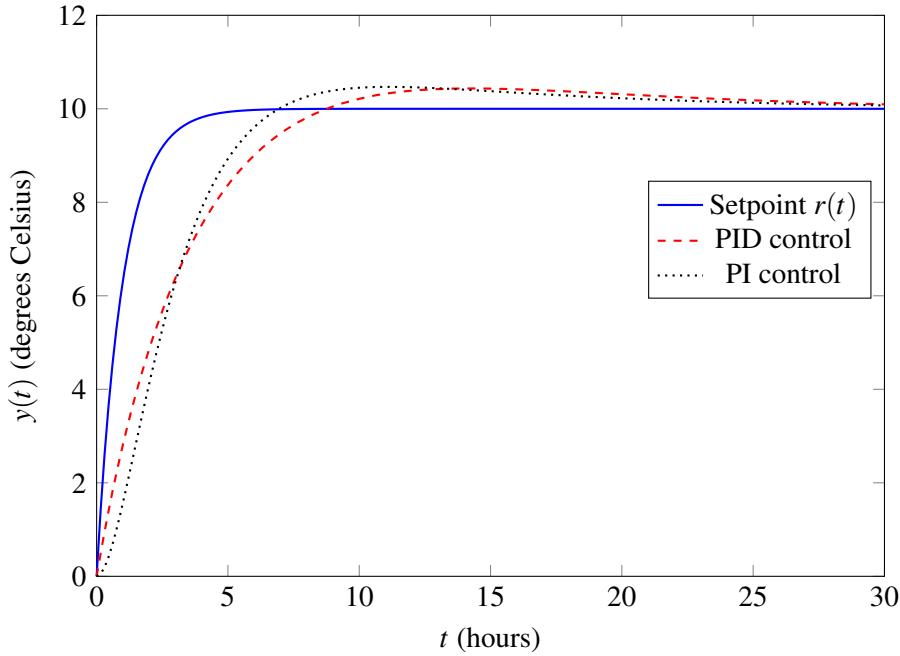


Figure 5.22: Incubator temperature with PID control (dashed red), PI control (dotted black), and setpoint temperature profile (solid blue).

temperature to be 0 degrees up to time $t = 20$ and then increase to 3 degrees, an abrupt (but here planned) change in the setpoint. This is modeled by the setpoint function $r(t) = 3H(t - 20)$. Suppose also that $y_0 = 3$, so the incubator starts at a temperature that is not consistent with the setpoint, since $r(0) = 0$. Finally, suppose that the lab ambient temperature is 0 degrees up to time $t = 40$, at which point the ambient temperature suddenly drops 5 degrees. This is modeled by taking $a(t) = -5H(t - 40)$. The controller here thus has several challenges: alter the incorrect initial condition to match the setpoint, then adapt to a setpoint change, and then adapt to a disturbance in the form of a 5 degree drop in the ambient temperature.

To determine how well the control works we must find the resulting incubator temperature $y(t)$ by solving (5.133), a task well suited to Laplace transforms. Begin by computing the Laplace transform of both sides of (5.133) to find $sY(s) - y_0 = -k(Y(s) - A(s)) + U(s)$, where $A = \mathcal{L}(a(t))$. Next make use of the fact that $y_0 = 3$ and $U(s) = G_c(s)(R(s) - Y(s))$ with $G_c(s)$ as given by (5.130), to find

$$sY(s) - y_0 = -k(Y(s) - A(s)) + G_c(s)(R(s) - Y(s)).$$

By making use of $G_p(s) = 1/(s + k)$ (recall that $K = 1$ here) this last equation can be written as

$$(1/G_p(s) + G_c(s))Y(s) = y_0 + kA(s) + G_c(R(s)).$$

Solve for $Y(s)$ as

$$Y(s) = G(s)R(s) + \frac{y_0G_p(s)}{1 + G_p(s)G_c(s)} + \frac{kA(s)G_p(s)}{1 + G_p(s)G_c(s)} \quad (5.134)$$

where $G(s) = G_p(s)G_c(s)/(1 + G_p(s)G_c(s))$ is the closed-loop transfer function in (5.119). After choosing and substituting in the chosen control gains K_p, K_i , and K_d that appear in $G_c(s)$, the time domain response $y(t)$ can be found by inverse transforming (5.134). We will take the cooling constant as $k = 0.05$.

For simple proportional control we'll use $K_p = 1$ with $K_i = K_d = 0$. For PI control we take $K_p = 1, K_i = 0.1$, and $K_d = 0$. For full PID control we will use $K_p = 1, K_i = 0.1$, and $K_d = 1$. A computer algebra system is certainly helpful for the computations here. The resulting incubator temperature $y(t)$ for each type of control is shown in Figure 5.23. ■

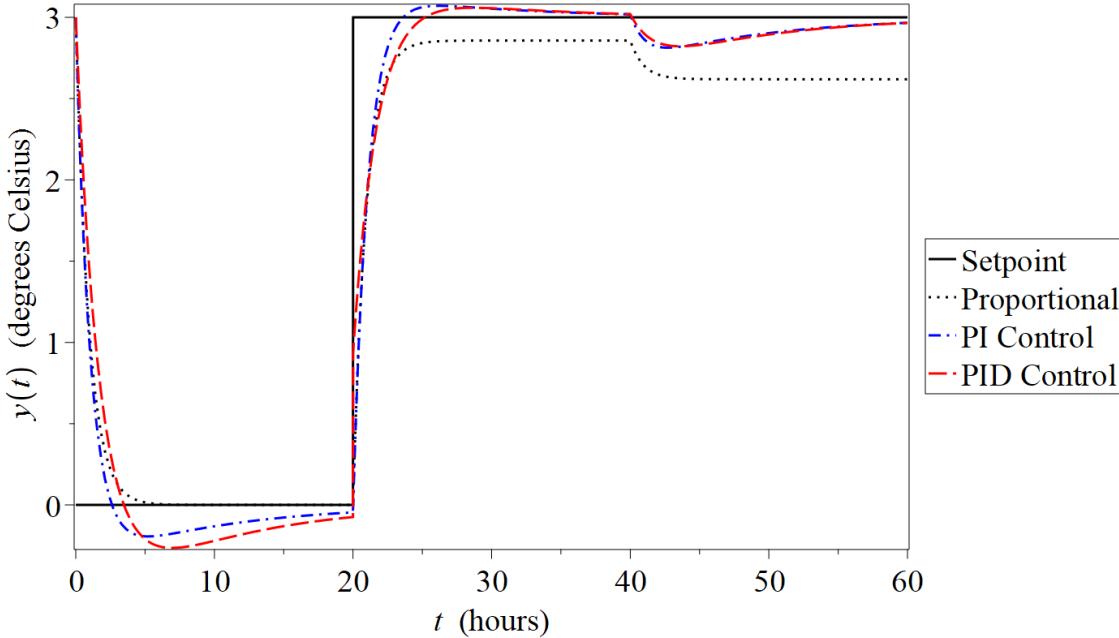


Figure 5.23: Setpoint for Example 5.44 and incubator temperature response for each of proportional, PI, and PID control.

Stability and Tuning the Controller Gains

As already mentioned, finding good choices for the parameters K_p , K_i , and K_d in PID control is known as tuning the controller and is beyond the scope of this book. However, one important consideration is that the controller be stable. Roughly speaking, an input-output system is **stable** if any $r(t)$ that is bounded for all $t \geq 0$ yields an output $y(t)$ that is also bounded for all $t \geq 0$. Depending on the system, achieving stable control may or may not be easy.

To illustrate how poor parameter choices lead to unstable control, consider our incubator model with proportional control, $u(t) = K_p e(t)$. The resulting system transfer function was computed in (5.121) and is

$$G(s) = \frac{KK_p}{KK_p + s + k}.$$

Suppose we make the poor choice of taking $K_p < 0$ (recall Reading Exercise 5.6.9). That is, when the incubator is running hotter than desired, turn the heat up, and when it's running too cold, turn up the cooling. In particular, consider $k = 0.05$ and $K = 0.5$ as we've previously used, but with $K_p = -1$. We'll use $r(t) = 10 - 10e^{-t}$. In this case the closed-loop transfer function is

$$G(s) = \frac{1/2}{9/20 - s}.$$

Using $Y(s) = G(s)R(s)$ with $R(s) = \mathcal{L}(r(t)) = \frac{10}{s(s+1)}$ yields

$$Y(s) = \frac{100}{s(s+1)(20s-9)}$$

after simplification. An inverse Laplace transform yields

$$y(t) = \frac{100}{9} - \frac{100}{29}e^{-t} - \frac{2000}{261}e^{9t/20}.$$

The solution $y(t)$ to the controlled ODE contains a term that grows exponentially, which is clearly undesirable. The heart of the problem is that the closed-loop transfer function $G(s)$ here has a pole with positive real part, at $s = 9/20$. This corresponds exactly to the exponentially growing term that involves $e^{9t/20}$.

Stable control requires that all poles for the closed-loop system transfer function $G(s)$ defined in (5.119) have negative real part. This condition puts certain constraints on the constants K_p , K_i , and K_d in a PID controller, constraints that depend on the nature of the system being controlled (through the plant transfer function $G_p(s)$). Beyond stability, however, tuning can be used to make the process variable $y(t)$ respond more rapidly to changes in the setpoint, reduce the overshoot, or satisfy other criteria. See [62] for more information.

5.6.8 Summary and Comments

Control theory plays a vital role in almost every aspect of our lives, from electrical power generation to automatic transmissions to industrial manufacturing to aircraft. Any piece of modern technology that has internal regulation, monitoring, and self-correction makes use of these techniques. The mathematical operations involved in PID control can be implemented electronically in digital or analog form (see [66]). This is why virtually all engineering curricula include coursework on this topic. The intelligent application of control theory can even stabilize physical systems that are inherently unstable and would hence be useless. With proper control these systems can be stabilized, and even offer advantages over other inherently stable designs. An example is the Lockheed F-117 aircraft: this plane has a fuselage designed for stealth, but that renders the plane aerodynamically unstable without active feedback control; see [81, 17]. By the appropriate application of control theory we get a stealthy plane that a human being can actually fly. See the project “Segway Scooters and the Inverted Pendulum” in Section 5.7 for an example of how a properly designed controller can stabilize an otherwise unstable system.

5.6.9 Exercises

Exercise 5.6.1 Show that the choice

$$u(t) = \frac{r'(t) + kr(t)}{K} \quad (5.135)$$

for an open-loop control in (5.110) with $y(0) = r(0)$ yields $y(t) = r(t)$. Then Laplace transform both sides of (5.135) and compare to (5.113) under the assumption $r(0) = 0$ (as was done there) noting that $G_c(s) = (s + k)/K$ from (5.114). Does the time domain choice for $u(t)$ in (5.135) correspond with the s -domain computation?

Exercise 5.6.2 Consider a system governed by the ODE $y'(t) = 0$. (Yes, it’s pretty boring.) However, suppose we incorporate a control function $u(t)$ as

$$y'(t) = u(t). \quad (5.136)$$

We want to control $y(t)$ so that $y(t) = r(t)$ for some setpoint $r(t)$. Assume we have initial condition $y(0) = r(0) = 0$.

- (a) Show that taking $u(t) = r'(t)$ works for open-loop control. Hint: verify that with this choice for $u(t)$ the solution to (5.136) with $y(0) = r(0)$ is $y(t) = r(t)$.
- (b) Repeat part (a) but in the s -domain. It may be helpful to refer to Figure 5.17. Specifically,
- Use (5.136) to write out the dependence of $Y(s)$ on $U(s)$. Show that the transfer function $G_p(s)$ is given by $G_p(s) = 1/s$.
 - If we take $u(t) = r'(t)$ as in part (a), what is the dependence of $U(s)$ on $R(s)$? What is the transfer function $G_c(s)$ here?
 - Verify that $G_p(s)G_c(s) = 1$, so that $Y(s) = G_p(s)U(s) = G_p(s)C_c(s)R(s) = R(s)$.

Exercise 5.6.3 Consider again the ODE (5.136) from Exercise 5.6.2, with setpoint $r(t)$ and initial condition $y(0) = r(0) = 0$. Suppose we implement proportional control in (5.136) by taking $u(t) = K_p e(t)$ with $e(t) = r(t) - y(t)$.

- (a) Use $u(t) = K_p e(t)$ to write out the controller transfer function $G_c(s)$ in $U(s) = G_c(s)E(s)$. With the plant transfer function $G_p(s) = 1/s$ (deduced in Exercise 5.6.2), use (5.119) to compute the closed-loop transfer function $G(s)$ in terms of K_p and s .
- (b) Take $r(t) = 5 - 5e^{-2t}$ and $K_p = 1$. Compute $R(s)$ and then use $Y(s) = G(s)R(s)$ to find $Y(s)$. Inverse transform $Y(s)$ to find $y(t)$. Plot $y(t)$ and $r(t)$ on the range $0 \leq t \leq 5$ and compare the plots. Does the controlled solution stabilize? To what value? Experiment with other values for K_p .
- (c) Use $\lim_{s \rightarrow 0^+} G(s) = 1$ and the final value theorem (Theorem 5.2.5) to show that if $r(t)$ approaches a limit r_0 for this particular control problem then $y(t)$ approaches the same limit. Hint: see (5.128).

Exercise 5.6.4 Consider again the ODE (5.136) from Exercise 5.6.2, with setpoint $r(t)$, and initial condition $y(0) = r(0) = 0$. Suppose we implement full PID control in (5.136), by taking

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d e'(t)$$

(this is just (5.129) again) with $e(t) = r(t) - y(t)$.

- (a) Write out the controller transfer function $G_c(s)$ in $U(s) = G_c(s)E(s)$. With the plant transfer function $G_p(s) = 1/s$ (deduced in Exercise 5.6.2), use (5.119) to compute the closed-loop transfer function $G(s)$ in terms of K_p, K_i, K_d , and s .
- (b) Take $r(t) = 5 - 5e^{-2t}$ and $K_p = K_i = K_d = 1$. Compute $R(s)$ and then use $Y(s) = G(s)R(s)$ to find $Y(s)$. Inverse transform $Y(s)$ to find $y(t)$. Plot and compare $y(t)$ and $r(t)$ on the interval $0 \leq t \leq 25$. Does the controlled solution stabilize? To what value? Experiment with other values for the control constants.
- (c) Suppose the controlled system in (5.136) is subject to an impulsive disturbance of total impulse 7 at time $t = 10$, so (5.136) becomes

$$y'(t) = 7\delta(t - 10) + u(t). \quad (5.137)$$

Solve (5.137) using $K_p = K_i = K_d = 1$ and $y(0) = 0$, then plot the solution for $0 \leq t \leq 50$. Does the controller deal with the disturbance effectively? Hint: to solve (5.137) just use Laplace transforms to obtain

$$sY(s) = 7e^{-10s} + G_c(s)(R(s) - Y(s)),$$

solve for $Y(s)$, and inverse transform.

Exercise 5.6.5 Consider an undamped spring-mass-damper system $mx''(t) + cx'(t) + kx(t) = 0$ but with an actuator that can exert a force of our choosing on the mass—the control $u(t)$ here is a force. The equation of interest is then

$$mx''(t) + cx'(t) + kx(t) = u(t).$$

Assume $x(0) = x'(0) = 0$. We want to control the mass's position so that $x(t) \approx r(t)$ for some chosen $r(t)$.

- (a) The plant here is the spring-mass-damper system with an input force $u(t)$ and output position $x(t)$. We thus assume that we can measure $x(t)$ at all times t . Write out the transfer function $G_p(s)$ that relates input $U(s)$ to output $X(s)$.
- (b) If we use PID control then

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d e'(t)$$

where $e(t) = r(t) - x(t)$, as usual. As a result, the controller transfer function in $U(s) = G_c(s)E(s)$ is exactly as in (5.130). Use value for $U(s)$ and (5.119) to compute the closed-loop transfer function $G(s)$ for the system transfer function. Show that $G(s)$ simplifies to

$$G(s) = \frac{K_d s^2 + K_p s + K_i}{ms^3 + (c + K_d)s^2 + (k + K_p)s + K_i}.$$

- (c) Take $K_p = 1, K_i = 0.5, K_d = 1$, along with $m = 1, c = 0.1, k = 4$, and $r(t) = 1 - e^{-t}$. Compute $Y(s) = G(s)R(s)$ and inverse transform to find $y(t)$. Plot $y(t)$ and $r(t)$ on the interval $0 \leq t \leq 30$. Comment on the effectiveness of the control.
- (d) For part (c), compute and plot $u(t)$ (the force exerted by the controller) for $0 \leq t \leq 30$.
- (e) Redo part (c) with $K_p = 1, K_i = 0.5, K_d = 0$, again with $m = 1, c = 0.1, k = 4$, and $r(t) = 1 - e^{-t}$ (so this is PI control, since $K_d = 0$). Compute $x(t)$ and plot $x(t)$ for $0 \leq t \leq 30$. Then compute the poles for $G(s)$ and explain the relationship of these poles to the behavior of $x(t)$.

5.7 Modeling Projects

5.7.1 Project: Drug Dosage

A hospitalized patient is to receive a drug at periodic intervals during a one-week stay. As in the morphine example of Section 5.1, in the absence of any additional doses the amount $u(t)$ (in mg) of this drug present in the patient's system diminishes according to

$$u'(t) = -ku(t) \tag{5.138}$$

for some positive constant k . We will measure time in hours, so k has units of reciprocal hours.

Modeling Exercise 7.1.1 Suppose the drug has a half-life of 5.8 hours in the patient. What is the appropriate constant k in $u'(t) = -ku(t)$?

The therapeutic range for the amount of the drug in the patient's system is between 15 and 30 mg; to go over this amount is to risk adverse effects, and lower than 15 mg has insufficient

therapeutic benefit. The drug is to be administered as a bolus at regular intervals, at a frequency and dose to maintain the required therapeutic range of drug in the patient's system. We would like to administer the drug with the least frequency possible—every 4 hours is better than every 2 hours—nurses are busy people. Moreover, the interval between doses should be practical, measured in a whole number of hours, not every 2 hours and 37 minutes. Finally, the dosage of this drug is standardized in 5 mg vials, so the dosage given, in mg, must be a multiple of 5. In what follows we will continue with the assumption that the half-life of the drug in the body is 5.8 hours.

Modeling Exercise 7.1.2 Suppose we begin with an initial bolus of 30 mg at time $t = 0$. Solve (5.138) with $u(0) = 30$ and plot the solution for $0 \leq t \leq 20$.

Modeling Exercise 7.1.3 Suppose we begin with an initial bolus of 30 mg at time $t = 0$ and then at time $t = 5$ hours a 10 mg bolus is administered. Modify the ODE (5.138) appropriately and solve. Plot the solution on the interval $0 \leq t \leq 20$.

Modeling Exercise 7.1.4 Devise a drug administration schedule that begins with an initial 30 mg dose at $t = 0$ and then has periodic bolus administration of the drug at regular intervals to meet the various requirements above: doses are in multiples of 5 milligrams, on a period measured in a whole number of hours, while keeping the amount of drug in the patient's system between 15 and 30 mg. The administration schedule should work for at least 72 hours. Demonstrate that your schedule meets the requirements.

Modeling Exercise 7.1.5 Suppose that after the initial 30 mg bolus the amount of drug in the patient's body should remain between 15 and 30 mg from time $t = 0$ to $t = 72$ hours but for $72 \leq t \leq 144$ hours the amount of drug should be decreased, to remain between 5 and 15 mg to good approximation, after which the drug will be discontinued. Devise a suitable administration schedule and demonstrate that your schedule meets these requirements.

5.7.2 Project: Machine Replacement

This project is based on the SIMIODE project “Machine Replacement—Laplace Transform” [125], which is itself based on a modeling project in [48] (pages 261–262). Although this project does not involve differential equations, it does involve derivatives, integrals, the Laplace transform, and demonstrates the application of convolution in a very natural and surprising way.

Consider a large manufacturing facility that contains a number of machines for producing goods, for example, stamping machines in a metal shop, printing machines at a publisher, or weaving machines in a textile factory. In order to meet manufacturing demands, the facility needs a certain number $N(t)$ of these machines to be in operation at time t ; we assume $N(t)$ is known or specified. Although $N(t)$ assumes integer values, we suppose that the number of machines is large enough that we may make the approximation that $N(t)$ takes on nonnegative real values.

Machine Failure

Machines break down or must be removed from service for maintenance, and we have to account for this in making sure that $N(t)$ machines always remain in operation. We will account for the failure or removal of machines from service over time by using a function $F(t)$ that quantifies what fraction of the machines in operation at a time t_0 are still in operation at time $t_0 + t$. That is, if $N(t_0)$ machines are in operation at time t_0 , then $N(t_0)F(t)$ of these are still functioning at time $t_0 + t$. We assume this fraction depends only on the length t of this time interval and not on the start of the interval at time t_0 .

Modeling Exercise 7.2.1

- (a) What is $F(0)$? Hint: all of the machines in operation at time t_0 are still in operation at time $t_0 + 0$.

- (b) Suppose that machines placed in service never fail or need to be taken out of service for maintenance. What is $F(t)$ in this case?
- (c) Suppose that any machine in operation at time t_0 stays in service until time $t_0 + 2$, at which point it must be turned off for maintenance. What is $F(t)$ in this case?
- (d) Suppose that half of the machines in operation at time t_0 are still in operation at time $t_0 + 3$, and half of those are still in service at time $t_0 + 6$, and so on (the machines have a half-life of 3 time units). What choice for $F(t)$ is consistent with this information?

Machine Replacement

If we start with $N(0)$ machines at time $t = 0$ and put no additional machines into operation, we will have $N(0)F(t)$ functioning machines at time t . However, in order to meet the target of $N(t)$ we may have to put additional machines into operation. Therefore we introduce a function $R(t)$ measuring the total number of replacement machines needed from time 0 to time t . In this case $R'(t)$ is the rate at which additional machines are being introduced into operation. As with $N(t)$, we assume $R(t)$ can be considered a continuously varying quantity, and is in this case differentiable.

Given a target number of machines $N(t)$ and failure rate function $F(t)$, our goal is to choose an appropriate replacement function $R(t)$ that ensures there will always be $N(t)$ machines in operation at time t .

Modeling Exercise 7.2.2 Let us partition the interval $[0, t]$ into subintervals of the form $[\tau_{k-1}, \tau_k]$ where

$$0 = \tau_0 < \tau_1 < \cdots < \tau_{n-1} < \tau_n = t.$$

Let $\Delta\tau_k = \tau_{k+1} - \tau_k$ for $0 \leq k \leq n - 1$. We will assume that $\Delta\tau_k$ is close to zero for each k (and will later approach zero).

- (a) From the definition of $R(t)$, in the time interval $[\tau_k, \tau_{k+1}]$ we place an additional $R(\tau_{k+1}) - R(\tau_k)$ machines into operation. Argue that if $\Delta\tau_k$ is close to zero then this additional number of machines at time t can well approximated as

$$R(\tau_{k+1}) - R(\tau_k) \approx R'(\tau_k)\Delta\tau_k.$$

Hint: did you pay attention to the definition of the derivative in calculus class?

- (b) Of the $N(0)$ machines put in operation at time 0, how many are still in operation at time t ?

Hint: reread the definition of F .

- (c) Consider the number of machines that were placed into operation in the interval τ_k to τ_{k+1} . Argue that at time t the number of these machines still in operation is well-approximated by $(R'(\tau_k)\Delta\tau_k)F(t - \tau_k)$. Hint: look back at part (a), and assume that F is continuous.

- (d) Argue that the number $N(t)$ of machines in operation at time t can be well-approximated as

$$N(t) \approx N(0)F(t) + \sum_{k=0}^{n-1} R'(\tau_k)F(t - \tau_k)\Delta\tau_k. \quad (5.139)$$

- (e) Argue further that if we refine the partition of $[0, t]$ so that $\max_k \Delta\tau_k \rightarrow 0$ then (5.139) becomes

$$N(t) = N(0)F(t) + \int_0^t R'(\tau)F(t - \tau) d\tau. \quad (5.140)$$

Equation (5.140) is the fundamental relation that allows us to compute how many replacement machines will be needed in any given time interval, given the failure function $F(t)$. Specifically, we know the function $N(t)$, the number of machines needed at time t , and we know the function F

(perhaps from experience or historical data). The goal is to solve (5.140) for the function R' , the rate at which new machines will be put into service. We can compute R also, if desired; note that $R(0) = 0$, from the definition of R . We now pause to consider how to solve (5.140). The tool that is essential here is the Laplace transform.

Integral Equations

Equation (5.140) is an example of an **integral equation** in which an unknown function (in this case R') is to be deduced from information concerning integrals of the function. Integral equations play an important role in applied mathematics, although they are not encountered at the undergraduate level as often as differential equations. We are well poised to handle (5.140), however, for it is a convolutional integral equation that can be solved using the Laplace transform.

Modeling Exercise 7.2.3 Laplace transform both sides of (5.140) and show that

$$\mathcal{L}(R')(s) = \frac{\mathcal{L}(N)(s)}{\mathcal{L}(F)(s)} - N(0). \quad (5.141)$$

Hint: recall the convolution theorem (Theorem 5.5.1).

Equation (5.141) allows us to solve for $\mathcal{L}(R')(s)$, from which we can inverse Laplace transform to find $R'(t)$. We can then integrate to find $R(t)$, if desired. Alternatively, since $\mathcal{L}(R')(s) = s\mathcal{L}(R)(s) - R(0)$ (and $R(0) = 0$) we can use (5.141) to find

$$\mathcal{L}(R)(s) = \frac{\mathcal{L}(N)(s)}{s\mathcal{L}(F)(s)} - \frac{N(0)}{s}, \quad (5.142)$$

and then inverse transform to find $R(t)$ directly.

Some Replacement Scenarios

Modeling Exercise 7.2.4 Suppose $N(t) = N_0$, a constant, and suppose that $F(t) = 1$ for all t . What is the interpretation of this choice for $F(t)$? (Look back at part (b) of Reading Exercise 7.2.1.) Solve (5.142) for $\mathcal{L}(R)$ and use this to find $R(t)$. Then comment on why this makes perfect sense.

Modeling Exercise 7.2.5 Suppose $N(t) = N_0$, a constant, and suppose that $F(t) = 1/2^{t/a} = e^{-t \ln(2)/a}$. There is some deeper probabilistic reasoning that underlies this choice of F , based on the assumption that a machine has a 50/50 probability of failing in any interval $[t, t+a]$, but we won't pursue that at this time. In essence, as in part (d) of Modeling Exercise 7.2.1, the machines have a half-life. Solve (5.142) for $\mathcal{L}(R)$ and use this to find $R(t)$; the answer depends on a . Does the answer seem sensible? In particular, consider $R'(t)$ (the rate at which machines must be replaced) and the dependence of $R'(t)$ on a .

Modeling Exercise 7.2.6 Consider a setting in which all machines are run for a certain period of time, say T time units, and are then replaced. In this case

$$F(t) = 1 - H(t - T),$$

where H is the Heaviside function. Justify this expression for F . With constant demand $N(t) = N_0$, use (5.142) to show that in this case

$$\mathcal{L}(R)(s) = \frac{N_0 e^{-sT}}{s(e^{-sT} - 1)}. \quad (5.143)$$

The Inverse Laplace Transform of (5.143)

The inverse transform of $\mathcal{L}(R)$ does not follow from anything in Table 5.2. To perform the inverse transform we will first use the geometric series identity

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots = \sum_{k=0}^{\infty} x^k,$$

which is valid for $|x| < 1$. With $x = e^{-sT}$ (note $0 < e^{-sT} < 1$ for $s, T > 0$) we find

$$\frac{1}{1 - e^{-sT}} = \sum_{k=0}^{\infty} e^{-skT} = 1 + e^{-sT} + e^{-2sT} + e^{-3sT} + \dots . \quad (5.144)$$

By using (5.144) we see that $\mathcal{L}(R)(s)$ in (5.143) can be expressed as

$$\begin{aligned} \mathcal{L}(R)(s) &= \frac{N_0}{s} (e^{-sT} + e^{-2sT} + e^{-3sT} + \dots) \\ &= \frac{N_0}{s} \sum_{k=1}^{\infty} e^{-skT} \\ &= \sum_{k=1}^{\infty} \frac{N_0 e^{-skT}}{s}. \end{aligned} \quad (5.145)$$

Consider a typical term $\frac{N_0 e^{-skT}}{s}$ in (5.145). Given that the inverse Laplace transform for N_0/s is just the constant function N_0 , from the second shifting theorem (Theorem 5.3.1) we conclude that

$$\mathcal{L}^{-1}(N_0 e^{-skT}/s) = N_0 H(t - kT).$$

If we use this in (5.145) and inverse Laplace transform term by term we obtain

$$\begin{aligned} R(t) &= N_0(H(t - T) + H(t - 2T) + H(t - 3T) + \dots) \\ &= N_0 \sum_{k=1}^{\infty} H(t - kT). \end{aligned} \quad (5.146)$$

Modeling Exercise 7.2.7

- (a) Suppose $N(t) = N_0 = 100$ and $T = 10$ in (5.146). What is $R(12)$ (how many replacements will be needed up to time 12)? What is $R(37)$? What is $R(79.9)$? What is $R(80.1)$? Plot $R(t)$ for $0 \leq t \leq 100$.
- (b) The quantity $R'(t)$ is the instantaneous rate at which machines must be replaced at time t . What is $R'(t)$ in part (a)? Hint: recall equation (5.65). Why does the answer make sense in this context?
- (c) Discuss how knowing the expression for $R(t)$ or $R'(t)$ would be of value to the managers of the plant. What planning or actions might it help them to take?

5.7.3 Project: Vibration Isolation Table Shakedown

Recall Example 4.2 in which we modeled a vibration isolation table as a spring-mass-damper system; see Figure 4.3. The goal in this project is to add active control to the vibration isolation table, in order to improve its performance. This is actually done in practice; see [7]. Vibration at any frequency is a problem, although frequencies less than 30 Hz are most problematic (see [111]) and are difficult to control.

Let's review the essentials of the vibration isolation table; refer to Figure 4.3 as needed. A tabletop of mass m is supported on a leg that can be modeled as a spring-damper system with spring constant k and damping constant c . The base of the leg that supports the tabletop rests on the ground. However, ground motion causes the base of this leg to move vertically with a displacement $d(t)$ in time. The tabletop itself experiences vertical motion that is transmitted through the leg. This vertical displacement of the tabletop is given by a function $y(t)$ that satisfies (4.9), reproduced here:

$$my''(t) + cy'(t) + ky(t) = k(d(t) + L_0) + cd'(t) - mg. \quad (5.147)$$

In this ODE m is the mass of the tabletop, and c and k are the damping coefficient and stiffness of the supporting leg, respectively. The parameter L_0 is the rest or equilibrium length of the leg/spring and $g > 0$ denotes gravitational acceleration. The function $d(t)$ will be considered a disturbance whose effect on the tabletop is to be controlled.

A Change of Variable

In the absence of any disturbance at the base of the table we have $d(t) = 0$ and (5.147) becomes

$$my''(t) + cy'(t) + ky(t) = kL_0 - mg. \quad (5.148)$$

It's helpful to define a new dependent variable $z(t) = y(t) - A$ for some offset A , or equivalently, $y(t) = z(t) + A$. Then the position of the tabletop when $d(t) = 0$ is given by $z(t) = 0$, or $y(t) = A$.

Modeling Exercise 7.3.1 Substitute $y(t) = A$ into (5.148) and show that if this is a solution we need $A = L_0 - mg/k$. Then substitute $y(t) = z(t) + A$ into (5.148) and show that $z(t)$ satisfies

$$mz''(t) + cz'(t) + kz(t) = 0. \quad (5.149)$$

Modeling Exercise 7.3.2 Argue that the transfer function for the system governed by (5.149) is $G_p(s) = 1/(ms^2 + cs + k)$. Hint: what is $G_p(s)$ in the relation $Z(s) = G_p(s)F(s)$ if $mz''(t) + cz'(t) + kz(t) = f(t)$ with $z(0) = z'(0) = 0$?

Modeling Exercise 7.3.3 With A and $z(t)$ as in Modeling Exercise 7.3.1, substitute $y(t) = z(t) + A$ into (5.147) and show that for a general driving function $d(t)$ the function $z(t)$ satisfies

$$mz''(t) + cz'(t) + kz(t) = kd(t) + cd'(t). \quad (5.150)$$

The function $kd(t) + cd'(t)$ on the right in (5.150) embodies a disturbance in the system. Our goal is to add a control function $u(t)$ to the system that, ideally, results in $z(t) = 0$.

Adding Control

We can add a control $u(t)$ to (5.150) to obtain

$$mz''(t) + cz'(t) + kz(t) = kd(t) + cd'(t) + u(t). \quad (5.151)$$

The control $u(t)$ here is a time-dependent force that would be implemented with some kind of actuator; the details do not concern us at the moment. We will use PID control in the usual form,

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d e'(t), \quad (5.152)$$

where $e(t) = r(t) - z(t)$ and $r(t)$ is the setpoint, the target for our control. Here we will use $r(t) = 0$ (we don't want the tabletop to move), so that $e(t) = -z(t)$.

This implementation of this control, like the implementations we've seen, assumes that we can measure the process variable, in this case $z(t)$, at all times, in order to feed that information into the control $u(t)$. That isn't necessarily the case here. As we'll see in the next few Reading Exercises, it may be only $z''(t)$ that we can measure. But let's proceed for the moment with controlling (5.150) via (5.152) with $e(t) = -z(t)$.

Let's assume that we begin with no disturbance, so $d(0) = 0$. If we Laplace transform both sides of (5.151) and use the plant transfer function $G_p(s) = 1/(ms^2 + cs + k)$ from Modeling Exercise 7.3.2 and the controller transfer function $G_c(s) = K_p + K_i/s + K_d s$ with $e(t) = -z(t)$ we obtain

$$\frac{Z(s)}{G_p(s)} = (k + cs)D(s) - G_c(s)Z(s), \quad (5.153)$$

where $Z(s) = \mathcal{L}(z(t))$ and $D(s) = \mathcal{L}(d(t))$.

Modeling Exercise 7.3.4 Solve (5.153) to show that

$$Z(s) = \frac{(cs + k)G_p(s)D(s)}{1 + G_c(s)G_p(s)}. \quad (5.154)$$

This equation relates the disturbance or ground motion $d(t)$ to the tabletop motion $z(t)$ in the s -domain, with the effect of the PID controller included.

Modeling Exercise 7.3.5 Let's simulate controlling a disturbance, say a periodic disturbance $d(t) = 0.0001 \sin(2\pi t)$, corresponding to a periodic vibration in the floor of amplitude 1/10 mm at 1 Hz. This may not sound like much, but this table may support a sensitive electron microscope or patient undergoing eye surgery. Let's assume the tabletop has mass 100 kg, that the leg spring constant is $k = 10^4$ newtons per meter, and that the system is critically damped, so $c = \sqrt{4mk} = 2000$ newtons per meter per second.

- (a) Start with control parameters $K_p = 10^5$, $K_i = 10^4$, and $K_d = 10^4$. Compute $D(s) = \mathcal{L}(d(t))$ and use (5.154) to compute $Z(s)$. Inverse transform to compute $z(t) = \mathcal{L}^{-1}(Z(s))$. Plot the tabletop motion $z(t)$ (displacement from equilibrium) for $0 \leq t \leq 10$.
- (b) Compare $z(t)$ in part (a) to the motion of the tabletop with no active control; you can do this by redoing part (a) with $K_p = K_i = K_d = 0$. Plot the motion of the tabletop with no control and compare to motion with PID control. Does the control help?
- (c) Redo parts (a)-(b) for frequencies 0.1 Hz and 10 Hz ($\pi/5$ to 20π radians per second).

Controlling Acceleration

The feedback signal for this situation would not likely be the tabletop position itself, but rather the tabletop acceleration $z''(t)$ (the same as $y''(t)$ here). The reason is that in vibration analysis, accelerations are comparatively easy to measure with an **accelerometer**, a small device that outputs an electrical signal in proportion to the acceleration it is experiencing in a fixed orientation. Such an accelerometer might be mounted on the tabletop to measure vertical (and other) acceleration. Moreover, for a sensitive experiment it's not the tabletop position $z(t)$ or velocity $z'(t)$ that is the problem (within reason), but rather the acceleration of the tabletop, $z''(t)$. Thus, what we'd really like to control is $z''(t)$, and keep it as close to zero as possible.

Look back at the closed-loop control picture in Figure 5.19. There $r(t)$ is the tabletop position, but $r''(t)$ is the desired setpoint for the acceleration, so we can consider $s^2R(s)$ as the s -domain input to the whole system; of course here our interest is $R(s) = 0$. The output for the system is the tabletop acceleration. We can capture this by modifying the plant transfer function to be $G_p(s) = s^2/(ms^2 + cs + k)$, which is the previous $G_p(s)$ but multiplied by s^2 , corresponding to the time domain operation of taking a second derivative. The full system output is then $s^2Y(s)$, corresponding to the tabletop acceleration, and this is also what is fed back to the controller.

Modeling Exercise 7.3.6 It may be helpful to refer to Figure 5.19. Use $E(s) = s^2(R(s) - Y(s))$, $G_c(s) = K_p + K_i/s + K_d s$, and $G_p(s) = s^2/(ms^2 + cs + k)$ in (5.154) along with $D(s) = \mathcal{L}(d(t))$ for $d(t) = 0.0001 \sin(2\pi t)$. Repeat parts (a)-(c) of Modeling Exercise 7.3.5 in this context, the control of acceleration. You might want to decrease K_p , K_i , and K_d by a factor of 10.

Experiment with different values of the control parameters K_p , K_i , and K_d . How effectively can you control the acceleration?

5.7.4 Project: Segway Scooters and The Inverted Pendulum

The Segway scooter made quite a splash when it was introduced in 2001. Its inventor, Dean Kamen, had high hopes for revolutionizing the personal transportation market, though things didn't work out quite as well as he had hoped. See [72] for an account of the history and development of this device, and [87] for a more recent account of the scooter's demise. The Segway scooter is a remarkable mix of technology and control algorithms that allow an inherently unstable system

(an upright two-wheeled device) to perform a useful function in a stable and controlled manner. In this project we'll consider a much simpler but somewhat similar problem, that of balancing an upside-down pendulum.

We begin with a review of the equation of motion for a pendulum and then add friction. We'll then shift our focus to the motion of the pendulum when it is nearly upside-down. In particular, we linearize the ODE for the upside-down pendulum's motion and add a control function $u(t)$ in the form of a torque. We then use PID control to design a controller that keeps the pendulum balanced vertically.

Review: Equation of Motion for a Damped Pendulum

In the project “The Pendulum 2” in Section 4.6 the ODE that governs a pendulum's motion in the presence of friction, the so-called **damped pendulum**, was derived. That equation, (4.132), is reproduced here:

$$\theta''(t) + c\theta'(t) + \frac{g}{L} \sin(\theta(t)) = 0. \quad (5.155)$$

Refer to Figure 4.32. In (5.155) $\theta(t)$ is the angle the pendulum makes with the vertical, $g > 0$ is gravitational acceleration, $L > 0$ is the length of the pendulum, and $c > 0$ is a frictional coefficient.

Modeling Exercise 7.4.1 Solve (5.155) numerically with $L = 2$, $c = 0.25$, and $g = 9.8$ and initial data $\theta(0) = 0.01$, $\theta'(0) = 0$ (this pendulum starts hanging almost straight down). Plot the solution on the interval $0 \leq t \leq 40$. Interpret—what does the plot say about the pendulum's motion?

The Inverted Pendulum

Our interest here is the **inverted pendulum**, when $\theta \approx \pi$. In particular, we're interested in how to use active feedback control to balance it. Let's make a change of variable in (5.155) by setting $\theta(t) = \pi - \alpha(t)$. The inverted pendulum then corresponds to $\alpha = 0$, with $\alpha > 0$ as clockwise position. Also, $\theta'(t) = -\alpha'(t)$ and $\theta''(t) = -\alpha''(t)$. Equation (5.155) becomes

$$\alpha''(t) + c\alpha'(t) - \frac{g}{L} \sin(\alpha(t)) = 0. \quad (5.156)$$

Note the change of sign in the last term. The solution to (5.156) with $\alpha(0) = 0$ and $\alpha'(0) = 0$ is $\alpha(t) = 0$, which is a perfectly balanced pendulum. There it will stay, so long as nothing disturbs it.

Modeling Exercise 7.4.2 Be careful in this problem: when $\alpha = 0$ the pendulum is oriented vertically upward, while $\alpha = \pm\pi$ is vertically straight down. Solve (5.156) numerically with $L = 2$, $c = 0.25$, and $g = 9.8$ and initial data $\alpha(0) = 0.01$ and $\alpha'(0) = 0$, so this pendulum starts almost upright. Plot the solution on the interval $0 \leq t \leq 40$. Interpret: what does the plot say about the pendulum's motion? If you start with $\alpha(0) = 10^{-10}$ and $\alpha'(0) = 0$, how long does the pendulum stay balanced, to visual approximation?

Linearizing and Adding Control

If the perfectly vertical pendulum is disturbed, it will tip over. Let us add active control to (5.156), by including some kind of actuator that can exert a torque on the pendulum at the pivot. A suitable modification of (5.156) is

$$\alpha''(t) + c\alpha'(t) - \frac{g}{L} \sin(\alpha(t)) = u(t), \quad (5.157)$$

for some control function $u(t)$ that is proportional to the torque exerted. The actual torque exerted would be $mL^2u(t)$, where m is the mass of the pendulum's bob, but we need not worry about that detail at the moment.

The main difficulty in controlling (5.157) is that this ODE is nonlinear; our PID control techniques require linearity of the ODE. We will thus linearize (5.157) as we did for the standard

pendulum in Section 4.6 by using the approximation $\sin(\alpha) \approx \alpha$ for $\alpha \approx 0$; this is appropriate, because we are trying to stabilize the pendulum in an inverted position. Doing this yields

$$\alpha''(t) + c\alpha'(t) - \frac{g}{L}\alpha(t) = u(t). \quad (5.158)$$

The goal is to choose a control $u(t)$ that stabilizes the pendulum at $\alpha = 0$. We thus take our setpoint $r(t) = 0$ for all $t \geq 0$. Assume that $\alpha(0) = \alpha'(0) = 0$, so the pendulum would stay upright if undisturbed, but disturbances may be introduced.

Modeling Exercise 7.4.3 Take $u(t) = 0$ (no control) and solve both (5.157) and (5.158) using $m = 1, L = 2, c = 0.5$, and $g = 9.8$ and initial data $\alpha(0) = 0.01$ and $\alpha'(0) = 0$. Plot both solutions on the interval $0 \leq t \leq 5$ and compare. Is the linearized ODE (5.158) a reasonable approximation of the full nonlinear ODE (5.157)? How large can α be before the solutions differ significantly?

Control

Modeling Exercise 7.4.4 Show that the transfer function in the s -domain that takes $u(t)$ to $\alpha(t)$ in (5.158) is given by

$$G_p(s) = \frac{1}{s^2 + cs - g/L}.$$

Modeling Exercise 7.4.5 We will implement PID control for (5.158) to obtain $\alpha(t) \approx r(t)$ (later with $r(t) = 0$) as

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d e'(t)$$

with $e(t) = r(t) - \alpha(t)$. In this case the controller transfer function $G_c(s)$ in the equation $U(s) = G_c(s)E(s)$ is exactly as in (5.130). Use $G_p(s)$ and $G_c(s)$ to write out the closed-loop transfer function $G(s)$ as in (5.119) explicitly in terms of K_p, K_i , and K_d and c, g , and L .

Modeling Exercise 7.4.6 Suppose the controlled system in (5.158) is subjected to an impulsive disturbance of total impulse 0.1 at time $t = 3$, and so (5.158) becomes

$$\alpha''(t) + c\alpha'(t) - g\alpha(t)/L = 0.1\delta(t - 3) + u(t). \quad (5.159)$$

Take $c = 0.1, L = 1$, and $g = 9.8$, with setpoint $r(t) = 0$ for $t \geq 0$ (so $R(s) = 0$ for all s). With this setpoint the controller will try to keep the pendulum vertically upright.

Solve (5.159) using $K_p = 20, K_i = 1, K_d = 1$ and initial conditions $\alpha(0) = \alpha'(0) = 0$, and then plot the solution for $0 \leq t \leq 15$. Does the controller deal with the disturbance at $t = 3$ effectively? Hint: to solve (5.159) use Laplace transforms to obtain $(s^2 + cs - g/L)A(s) = 0.1e^{-3s} + G_c(s)(0 - A(s))$ where $A(s) = \mathcal{L}(\alpha(t))$, or equivalently,

$$\frac{A(s)}{G_p(s)} = 0.1e^{-3s} + G_c(s)(0 - A(s)).$$

Solve for $A(s)$, and inverse transform. Experiment with other choices for K_p, K_i , and K_d .

Modeling Exercise 7.4.7 If you experiment with other choices for K_p, K_i , and K_d in Modeling Exercise 7.4.6 you may find the controller is unstable. Try the choices $K_p = 5, K_i = 1, K_d = 1$, solve for $\alpha(t)$ as in Modeling Exercise 5.159, and plot $\alpha(t)$ for $0 \leq t \leq 25$. What does this control do to the pendulum? Compute the poles of $G(s)$ and explain what these poles have to do with the observed behavior.

Controlling the Nonlinear Pendulum

We derived a controller that can keep the linearized equation (5.158) stabilized at $\alpha = 0$, but the real system of interest is governed by (5.157), a nonlinear ODE. That is the system on which we should test our controller. Unfortunately, we can't solve with the Laplace transform because it doesn't work on nonlinear equations. The controlled equation we have to confront is (5.159), which is

$$\alpha''(t) + c\alpha'(t) - g \sin(\alpha(t))/L = f(t) + K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d e'(t) \quad (5.160)$$

where $e(t) = r(t) - \alpha(t) = -\alpha(t)$ when $r(t) = 0$, and $f(t)$ is some type of disturbance, such as $f(t) = 0.1\delta(t - 3)$ as in Modeling Exercise 7.4.6. Equation (5.160) is a nonlinear **integrodifferential equation**, due to the presence of the integral on the right side.

Modeling Exercise 7.4.8 There is a simple case we can analyze without much trouble: take $K_i = 0$, so that we're using PD control; the integral in (5.160) disappears. Using $e(t) = -\alpha(t)$ we're left with

$$\alpha''(t) + c\alpha'(t) - g \sin(\alpha(t))/L = f(t) - K_p \alpha(t) - K_d \alpha'(t). \quad (5.161)$$

Equation (5.161) is a second-order nonlinear ODE that we can tackle with a standard numerical ODE solver.

Take $K_p = 20$ and $K_d = 1$ and solve (5.161) numerically with $f(t) = 0.1\delta(t - 3)$ and initial conditions $\alpha(0) = \alpha'(0) = 0$. Plot the solution for $0 \leq t \leq 15$. Does the PD controller work on the nonlinear system? Why should this be expected in this case?

Modeling Exercise 7.4.9 Solve (5.161) again, this time with $K_p = 3$ and $K_d = 1$. Where does the solution stabilize? Can you explain what went wrong? This illustrates that the control based on a linearized approximation does a good job as long as the system remains near the point at which the linearization is valid. If not, all bets are off.

6. Linear Systems of Differential Equations

6.1 Systems of Differential Equations

We begin this section by developing a two-compartment model for the metabolism of a drug. This model yields a coupled pair of linear ODEs for two unknown functions and provides motivation for the remainder of this chapter, which is devoted to the study of linear systems of differential equations.

6.1.1 Motivation: More Pharmacokinetics

The material in this section is based on the SIMIODE project [124], and is expanded on in Section 6.5.1 at the end of this chapter.

In the 1950s and 1960s the drug lysergic acid diethylamide (known as “LSD” or “acid” in popular slang) gained the attention of scientists [113], government agencies [1], and the general public [38], for its various psychoactive and hallucinogenic properties. A number of studies were conducted concerning the effect of LSD on humans. In the research study [14] five normal male volunteer subjects, ages 21 to 25, were administered 2 micrograms of LSD per kilogram of body mass intravenously over a 1.5 minute period. Blood samples were then drawn at 5, 15, 30, 60, 120, 240, and 480 minutes and these were tested for concentration levels of LSD. On page 612 of [14] the authors state that, “To obtain a crude index of performance, subjects were given one of a series of equivalent tests, consisting of simple addition problems, after each blood sample was drawn.” This information is given in Table 6.1. The goal of the study was to investigate the absorption and metabolism of the drug by body tissues, and examine and correlate this with the drug’s effect on the subjects’ mental abilities. The data makes it clear that the drug did alter the subject’s ability to perform simple arithmetic. (At the one-hour mark Subject 3 apparently couldn’t do a single problem correctly. Really?)

A Two-Compartment Model

Earlier in this text we considered a variety of one-compartment models. In particular, in Section 1.2 we looked at the intracochlear drug delivery model, in Section 5.1 we studied a model for morphine metabolism, and in Section 2.1.3 we considered the more abstract salt tank problems. In each case

Time (hr)		0.0833	0.25	0.5	1.0	2.0	4.0	8.0
Subject 1	Plasma Conc (ng/ml)	11.1	7.4	6.3	6.9	5.0	3.1	0.8
	Perform Score (%)	73	60	35	50	48	73	97
Subject 2	Plasma Conc (ng/ml)	10.6	7.6	7.0	4.8	2.8	2.5	2.0
	Perform Score (%)	72	55	74	81	79	89	106
Subject 3	Plasma Conc (ng/ml)	8.7	6.7	5.9	4.3	4.4	—	0.3
	Perform Score (%)	60	23	6	0	27	69	81
Subject 4	Plasma Conc (ng/ml)	10.9	8.2	7.9	6.6	5.3	3.8	1.2
	Perform Score (%)	60	20	3	5	3	20	62
Subject 5	Plasma Conc (ng/ml)	6.4	6.3	5.1	4.3	3.4	1.9	0.7
	Perform Score (%)	78	65	27	30	35	43	51

Table 6.1: Summary of data collected [14, 88] on five male volunteers who were administered LSD and then tested on performance (Perform Score (%)) on simple addition questions. Both performance Score and Plasma Concentrations of LSD were recorded at 5, 15, 30, 60, 120, 240, and 480 minutes after the initial infusion of LSD.

a conservation model was used, explicitly accounting for the changing amount of a substance in a tank or compartment with a “rate of change equals rate in minus rate out plus rate of creation” approach. The same approach can be used to model systems that consist of multiple compartments.

In the papers [88] and [113], the authors offer a **two-compartment** model for the behavior of LSD in the body. The situation is illustrated in Figure 6.1. In this model the authors divide the relevant regions of the body for the action and metabolism of LSD into a plasma compartment and a tissue compartment. These compartments are not dictated precisely by physiology, but represent abstractions of the main components in the body in which the concentration and action of LSD may differ. Roughly speaking, “plasma” refers to the liquid portion of the blood and interstitial fluid, while “tissue” refers to various organs such as the brain. The drug moves between these compartments and this is the process we want to model: how much of the drug is in each compartment as a function of time, and what is the rate at which the drug moves between the compartments? We can approach the problem by modeling the concentration of the drug in each compartment, or by modeling the actual amount of the drug. Although concentration may be the most physiologically relevant quantity, let’s stick to the actual amount of drug in each compartment; we’ll come back to concentration later in the modeling projects in Section 6.5.1.

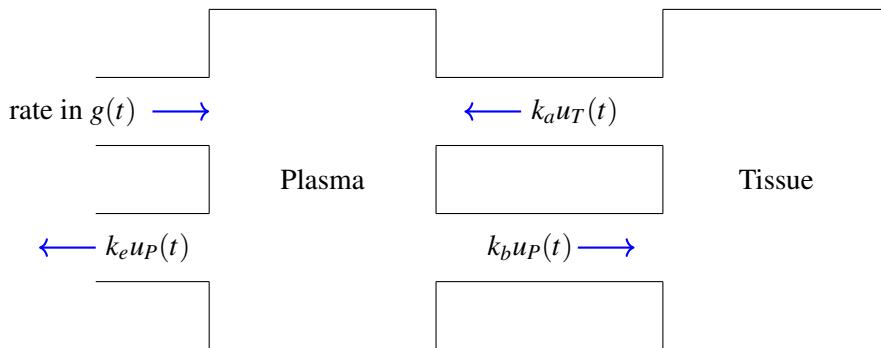


Figure 6.1: A two-compartment model with compartments corresponding to plasma and tissue.

Let $u_P(t)$ denote the amount (mass) of the drug in the plasma at time t , and let $u_T(t)$ be the amount (mass) of the drug in the tissue at time t . In [88] and [113] the authors posit a model of the

general form

$$\dot{u}_P(t) = -k_b u_P(t) - k_e u_P(t) + k_a u_T(t) + g(t) \quad (6.1)$$

$$\dot{u}_T(t) = k_b u_P(t) - k_a u_T(t) \quad (6.2)$$

for certain positive rate constants k_a , k_b , and k_e . Refer to Figure 6.1. We have adopted the convenient notation \dot{z} for the time derivative of a quantity z , which is rather common for systems of ODEs. The term $k_a u_T(t)$ on the right in (6.1) quantifies the rate at which LSD enters the plasma from the tissue through the top conduit in Figure 6.1, which is assumed to be in proportion to u_T . The term $k_b u_P(t)$ in (6.2) is similar, but it quantifies the flow of the drug from the plasma back to the tissue through the bottom conduit. The $-k_a u_T(t)$ term in (6.2) captures the conservation principle that none of the drug is lost as it flows between compartments, so anything that leaves the tissue through the top conduit enters the plasma. The $-k_b u_P(t)$ term in (6.1) has a similar interpretation. The term $g(t)$ in (6.1) indicates the rate at which the drug is administered to the patient via the plasma or bloodstream. Finally, the $-k_e u_P(t)$ term quantifies the rate at which the drug is removed from the plasma and excreted by the body. There are a number of physiological assumptions and some more detailed modeling that underlie (6.1)-(6.2) that we will explore in Section 6.5.

The more immediate concern is this: (6.1)-(6.2) constitute a pair of coupled ODEs in which there are two unknowns functions, $u_P(t)$ and $u_T(t)$. For now we will consider the constants k_a , k_b , and k_e as known (ultimately, of course, they must be measured or inferred from data). This pair of ODEs would come with two initial conditions in which the values of $u_P(0)$ and $u_T(0)$ are specified, the initial LSD concentration in the plasma and the tissue, respectively. This chapter is devoted to the analysis and solution of systems of linear, constant-coefficient differential equations like (6.1)-(6.2); see Definitions 6.1.1 and 6.1.2 below. The next chapter explores techniques for analyzing nonlinear systems.

First-Order Systems of ODEs

A first-order system of ODEs for functions $x_1(t), \dots, x_n(t)$ in **standard form** is written

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, \dots, x_n, t) \\ \dot{x}_2 &= f_2(x_1, \dots, x_n, t) \\ &\vdots \\ \dot{x}_n &= f_n(x_1, \dots, x_n, t) \end{aligned} \quad (6.3)$$

where for notational simplicity we will usually suppress the dependence of each function x_j on t . For example, if we define $x_1 = u_P$ and $x_2 = u_T$, then (6.1)-(6.2) is a system of the form (6.3), where

$$\begin{aligned} f_1(x_1, x_2, t) &= -k_b x_1 - k_e x_1 + k_a x_2 + g(t) \\ f_2(x_1, x_2, t) &= k_b x_1 - k_a x_2. \end{aligned}$$

The function f_1 depends on t through the function $g(t)$, while f_2 does not depend explicitly on t .

When convenient we may write a system like (6.3) in the vector-valued form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$ where

$$\mathbf{x}(t) = \langle x_1(t), \dots, x_n(t) \rangle,$$

and $\mathbf{f}(\mathbf{x}, t)$ is the vector-valued function

$$\mathbf{f}(\mathbf{x}, t) = \langle f_1(\mathbf{x}, t), \dots, f_n(\mathbf{x}, t) \rangle.$$

A system of the form (6.3) will usually come with initial conditions $x_j(t_0) = q_j$ for some initial time t_0 and constants q_j with $1 \leq j \leq n$; equivalently, we may write $\mathbf{x}(t_0) = \mathbf{q}$.

Some terminology is useful at this point.

Definition 6.1.1 A system of ODEs of the form (6.3) is **linear** if each function f_j is of the form

$$f_j(x_1, \dots, x_n, t) = a_{j,1}(t)x_1 + a_{j,2}(t)x_2 + \dots + a_{j,n}(t)x_n + b_j(t) \quad (6.4)$$

for functions $a_{j,1}(t), \dots, a_{j,n}(t)$ and $b_j(t)$. If each $b_j(t) = 0$ the system is **homogeneous**.

Reading Exercise 6.1.1 Verify that the system

$$\begin{aligned}\dot{x}_1 &= tx_1 - 3x_2 + t^2 \\ \dot{x}_2 &= -\sin(t)x_1 + 3tx_2 - t\end{aligned}$$

is linear. What are the functions $a_{1,1}(t), a_{1,2}(t), a_{2,1}(t), a_{2,2}(t), b_1(t)$, and $b_2(t)$ here?

Definition 6.1.2 A linear system of ODEs with f_j of the form in (6.4) is **constant-coefficient** if all of the functions $a_{j,m}(t)$ are constant (but $b_j(t)$ need not be constant). Otherwise the system is **variable-coefficient**.

For example, the system (6.1)-(6.2) is a first-order, linear, constant-coefficient system. These types of equations are the focus for this chapter.

Remark 6.1.1 It's worth noting that a solution to a system like (6.3) is an n -tuple of functions $x_1(t), \dots, x_n(t)$ and may be interpreted geometrically as parameterizing a curve, as $x_1 = x_1(t), \dots, x_n = x_n(t)$, in \mathbb{R}^n (n -dimensional space with coordinates (x_1, x_2, \dots, x_n)). This observation will be especially helpful in the next chapter.

Converting Higher-Order ODEs to First-Order Systems

Second- and higher-order scalar ODEs can usually be converted into equivalent systems of first-order ODEs, and so analyzed using the methods we introduce in this chapter and the next chapter. The best way to master this technique is to see a few worked examples, and to try some yourself.

■ **Example 6.1** Consider the spring-mass-damper system $mu''(t) + cu'(t) + ku(t) = f(t)$ with initial conditions $u(0) = u_0$ and $u'(0) = v_0$. We can convert this into an equivalent pair of first-order ODEs as follows. Define two new functions $x_1(t)$ and $x_2(t)$ as $x_1(t) = u(t)$ and $x_2(t) = u'(t)$. It is clear that the equation

$$\dot{x}_1(t) = x_2(t)$$

holds, by the definition $x_1 = u$ and $x_2 = u'$. Also note that $mu''(t) + cu'(t) + ku(t) = f(t)$ can be solved for $u''(t)$ as $u''(t) = -cu'(t)/m - ku(t)/m + f(t)$, so that

$$\begin{aligned}\dot{x}_2(t) &= u''(t) \\ &= -\frac{c}{m}u'(t) - \frac{k}{m}u(t) + f(t) \\ &= -\frac{c}{m}x_2(t) - \frac{k}{m}x_1(t) + f(t).\end{aligned}$$

In short, by setting $x_1 = u$ and $x_2 = u'$, the ODE $mu'' + cu' + ku = f$ can be written as a pair of first-order ODEs

$$\dot{x}_1 = x_2 \quad (6.5)$$

$$\dot{x}_2 = -\frac{k}{m}x_1 - \frac{c}{m}x_2 + f(t). \quad (6.6)$$

The initial conditions for u become $x_1(0) = u_0$ and $x_2(0) = v_0$.

It is important to note that the system (6.5)-(6.6) can also be converted back to $mu''(t) + cu'(t) + ku(t) = f(t)$, by differentiating (6.5) with respect to t to obtain $\ddot{x}_1 = \dot{x}_2$ and then replacing \dot{x}_2 with \ddot{x}_1 in (6.6) (and noting $x_2 = \dot{x}_1$). We find

$$\ddot{x}_1 = -\frac{k}{m}x_1 - \frac{c}{m}\dot{x}_1 + f(t).$$

If we set $u = x_1$, this last equation is equivalent to $mu'' + cu' + ku = f(t)$. Thus the original ODE and the system (6.5)-(6.6) are completely equivalent. ■

Reading Exercise 6.1.2 Use the approach of Example 6.1 to convert the nonlinear pendulum equation

$$\ddot{\theta}(t) + \frac{g}{L} \sin(\theta(t)) = 0$$

to a pair of first-order ODEs (one equation will be nonlinear), by letting $x_1 = \theta$ and $x_2 = \dot{\theta}$.

■ **Example 6.2** Systems of second-order (or higher-order) ODEs are common and can also be converted to equivalent systems of first-order equations. To illustrate, consider the damped double spring-mass system in Figure 6.2.

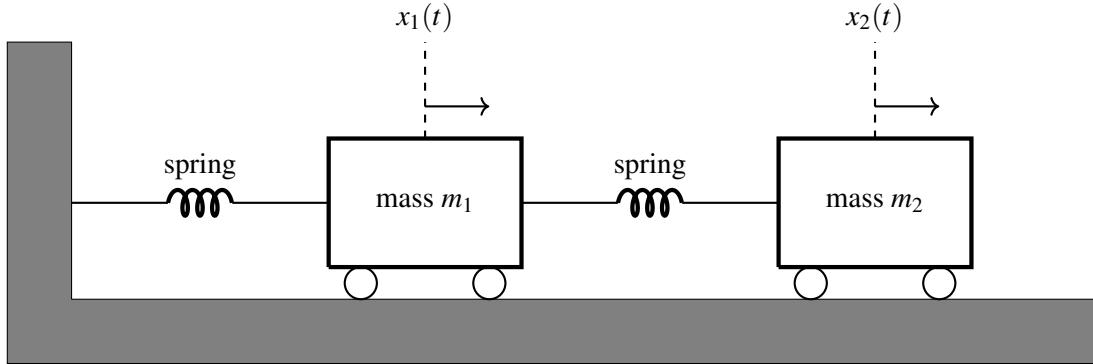


Figure 6.2: Double spring-mass system.

To model this system, let us assume that the masses have negligible widths (despite Figure 6.2). Let L_1 denote the rest length of the first spring (let us call it “spring 1”) connecting mass m_1 to the left wall. Similarly let L_2 denote the rest length of the spring connecting m_1 to m_2 (let us call it “spring 2”). We will use $x_1(t)$ to denote the position of mass m_1 with respect to the wall and $x_2(t)$ for the position of the mass m_2 . Assume the springs each obey Hooke’s law with spring constants k_1 for spring 1 and k_2 for spring 2. Although there are no explicit dampers in the system, assume that each mass experiences viscous friction in proportion to its velocity. This friction might come from, for example, the wheels in contact with the ground.

Reading Exercise 6.1.3

- Justify that the amount spring 1 is stretched or compressed is $x_1 - L_1$, and the amount spring 2 is stretched or compressed is $x_2 - x_1 - L_2$.
- Based on the above assumptions, justify that the force exerted on mass m_1 by spring 1 is $-k_1(x_1 - L_1)$, the force exerted by spring 2 is $k_2(x_2 - x_1 - L_2)$, and the frictional force on mass 1 is $-c_1\dot{x}_1$ for some nonnegative constant c_1 . A free body diagram may help.
- Based on the above assumptions, justify that the force exerted on mass m_2 by spring 2 is $-k_2(x_2 - x_1 - L_2)$, and the frictional force on mass 2 is $-c_2\dot{x}_2$ for some nonnegative constant c_2 . Again, a free body diagram may help.

Reading Exercise 6.1.4 Based on Reading Exercise 6.1.3 and Newton's second law of motion, show that $x_1(t)$ and $x_2(t)$ satisfy the coupled second-order system

$$\begin{aligned} m_1\ddot{x}_1 &= -k_1(x_1 - L_1) + k_2(x_2 - x_1 - L_2) - c_1\dot{x}_1 \\ m_2\ddot{x}_2 &= -k_2(x_2 - x_1 - L_2) - c_2\dot{x}_2. \end{aligned} \quad (6.7)$$

The system (6.7) is a nonhomogeneous second-order system. We can make a convenient change of variable, by setting $u_1(t) = x_1(t) - L_1$ and $u_2(t) = x_2(t) - L_1 - L_2$, so that $x_1(t) = u_1(t) + L_1$ and $x_2(t) = u_2(t) + L_1 + L_2$. In this case $u_1 = u_2 = 0$ corresponds to the system at equilibrium, where both springs are at their rest length. With this change of dependent variables and after dividing through by m_1 , the first ODE in the system (6.7) becomes

$$\ddot{u}_1 = -\frac{k_1 + k_2}{m_1}u_1 + \frac{k_2}{m_1}u_2 - \frac{c_1}{m_1}\dot{u}_1. \quad (6.8)$$

After dividing through by m_2 the second ODE of the system (6.7) becomes

$$\ddot{u}_2 = \frac{k_2}{m_2}u_1 - \frac{k_2}{m_2}u_2 - \frac{c_2}{m_2}\dot{u}_2. \quad (6.9)$$

Equations (6.8)-(6.9) are a coupled pair of second-order ODEs that are homogeneous. Moreover, the unnecessary details about the length of each spring are removed from the equation. If we solve for u_1 and u_2 we can then obtain the actual positions x_1 and x_2 , if desired.

The system of ODEs (6.8)-(6.9) can be converted to a system of first-order ODEs, as follows. Let $w_1(t) = u_1(t)$, $w_2(t) = \dot{u}_1(t)$, $w_3(t) = u_2(t)$, and $w_4(t) = \dot{u}_2(t)$. In this case the relation

$$\dot{w}_1 = w_2 \quad (6.10)$$

is immediate. From $\ddot{u}_1 = \dot{w}_2$ and (6.8) we obtain

$$\dot{w}_2 = -\frac{k_1 + k_2}{m_1}w_1 - \frac{c_1}{m_1}w_2 + \frac{k_2}{m_1}w_3,$$

where each term on the right in (6.8) has been replaced by its w_j equivalent. That takes care of (6.8). For (6.9), note that

$$\dot{w}_3 = w_4,$$

and since $\dot{w}_4 = \ddot{u}_2$ we can use (6.9) to write

$$\dot{w}_4 = \frac{k_2}{m_2}w_1 - \frac{k_2}{m_2}w_3 - \frac{c_2}{m_2}w_4. \quad (6.11)$$

Equations (6.10) to (6.11) for w_1, w_2, w_3 , and w_4 are a set of four first-order equations that are equivalent to (6.8)-(6.9). ■

Reading Exercise 6.1.5 Show that the displayed equations from (6.10) to (6.11) can be converted back to (6.8)-(6.9).

6.1.2 Existence and Uniqueness

There is a theorem regarding the existence and uniqueness for solutions to a first-order system of ODEs (6.3), quite analogous to Theorem 2.4.1 for scalar equations.

Theorem 6.1.1 — Existence-Uniqueness for First-Order Systems. For a first-order system of ODEs (6.3) suppose that each function f_k is continuous and has continuous partial derivatives with respect to all variables in some region

$$R = \{(x_1, \dots, x_n, t) : a_k - \delta_k < x_k < a_k + \delta_k, 1 \leq k \leq n \text{ and } t_0 - \delta_0 < t < t_0 + \delta_0\},$$

where all δ_k are positive. Then there is a unique solution to (6.3) that satisfies $x_1(t_0) = a_1, x_2(t_0) = a_2, \dots, x_n(t_0) = a_n$ for some time interval $t_0 - \varepsilon < t < t_0 + \varepsilon$ with $\varepsilon > 0$.

Briefly, Theorem 6.1.1 states that if the right hand side each equation in (6.3) is continuous and has continuous partial derivatives near the initial point of interest, then we can rest assured there is a unique solution through that initial point. Most of the systems we encounter in this text will fall under the umbrella of Theorem 6.1.1. The exceptions are, as with scalar equations, systems forced with Heaviside or Dirac delta functions. As with scalar equations, these are handled with special techniques.

6.1.3 Exercises

Exercise 6.1.1 Classify each system as linear or nonlinear. If the system is linear, classify it as constant- or variable-coefficient, and determine whether it is homogeneous.

- (a) $\dot{x}_1 = x_1 x_2, \dot{x}_2 = x_1 - x_2$
- (b) $\dot{x}_1 = x_1 + x_2, \dot{x}_2 = x_1 - 2x_2$
- (c) $\dot{x}_1 = \sin(x_1 + x_2), \dot{x}_2 = \cos(x_1 - x_2)$
- (d) $\dot{x}_1 = tx_1 - x_2, \dot{x}_2 = x_1 + \sin(t)x_2$
- (e) $\dot{x}_1 = x_1/(t^2 + 1) + x_1/x_2, \dot{x}_2 = 3$
- (f) $\dot{x}_1 = e^{x_1+2x_2}, \dot{x}_2 = x_1 - 4x_2$
- (g) $\dot{x}_1 = x_1 + x_2 + x_3, \dot{x}_2 = x_1 + tx_2 - x_3, \dot{x}_3 = x_1 - 2x_2 + e^t x_3$
- (h) $\dot{x}_1 = x_1 + x_2 x_3, \dot{x}_2 = x_1 + tx_2 - x_3, \dot{x}_3 = x_1 - 2x_2 + e^t x_3$
- (i) $\dot{x}_1 = x_1 + x_2, \dot{x}_2 = x_1 - 2x_2 + t^2$
- (j) $\dot{x}_1 = x_1 + x_2, \dot{x}_2 = \cos(t)x_1 + 5x_2$
- (k) $\dot{x}_1 = tx_1 - x_2, \dot{x}_2 = 4$

Exercise 6.1.2 Convert each second- or higher-order equation into an equivalent system of first-order ODEs, with initial conditions. Use x_1, x_2, \dots for the dependent variables of the new system, with $x_1 = u, x_2 = u'$, etc.

- (a) $3u''(t) + 5u'(t) + 4u(t) = 0$ with $u(0) = 7$ and $u'(0) = 5$
- (b) $u''(t) + u'(t) + u(t) = 0$ with $u(0) = 1$ and $u'(0) = 0$
- (c) $2u''(t) + 2\cos(u'(t)) + u(t) = 0$ with $u(0) = 3$ and $u'(0) = -1$
- (d) $u''(t) + u'(t)u(t) = 7$ with $u(1) = 2$ and $u'(1) = 4$
- (e) $u'''(t) + 2u''(t) + u'(t) + 5u(t) = 0$ with $u(0) = 1, u'(0) = 0$ and $u''(0) = -1$. Hint: take $x_1 = u, x_2 = u',$ and $x_3 = u''$. Two of the ODEs are $\dot{x}_1 = x_2$ and $\dot{x}_2 = x_3$.
- (f) $u^{(4)}(t) + u'''(t) + 4u''(t) + 5u'(t) + 3u(t) = 0$ with $u(0) = 1, u'(0) = 0, u''(0) = -1$ and $u'''(0) = 4$. Hint: take $x_1 = u, x_2 = u', x_3 = u'',$ and $x_4 = u'''$. Three of the ODEs are $\dot{x}_1 = x_2, \dot{x}_2 = x_3,$ and $\dot{x}_3 = x_4$.

Exercise 6.1.3 Convert each coupled system into an equivalent system of first-order ODEs, with initial conditions. Use x_1, x_2 , etc. for the dependent variables in the new system.

(a) $u''_1(t) + u'_1(t) - u_2(t) = \sin(t)$ and $u'_2(t) - u_2(t) + 3u_1(t) = 0$ with $u_1(0) = 1$, $u'_1(0) = 3$, and $u_2(0) = -2$. Hint: take $x_1 = u_1$, $x_2 = u'_1$, and $x_3 = u_2$.

(b) $3u''_1(t) + \sin(u'_2(t)) - u_2(t) = t$ and $u'_2(t) - 2u_2(t) + u_1(t) = 0$ with $u_1(0) = 1$, $u'_1(0) = 3$, $u_2(0) = -2$, and $u'_2(0) = 5$. Hint: follow the lead of part (a).

Exercise 6.1.4 Consider (6.1) and (6.2) in the case that $k_a = 4$, $k_b = 1$, $k_e = 3$, and $g(t) = 0$. The system becomes

$$\dot{u}_P(t) = -4u_P(t) + 4u_T(t)$$

$$\dot{u}_T(t) = u_P(t) - 4u_T(t).$$

Suppose the initial conditions are $u_P(0) = 1$ and $u_T(0) = 0$ (so a dose of 1 unit of a drug is injected into the plasma or bloodstream at time $t = 0$, with none in the tissue).

Verify that the functions

$$u_P(t) = \frac{e^{-2t}}{2} + \frac{e^{-6t}}{2}$$

$$u_T(t) = \frac{e^{-2t}}{4} - \frac{e^{-6t}}{4}$$

provide a solution with the proper initial conditions. Plot $u_P(t)$ and $u_T(t)$ both on the interval $0 \leq t \leq 5$ and comment: do these make sense for the amount of drug in the plasma and tissue?

Exercise 6.1.5 You already know how to solve linear constant-coefficient systems using the Laplace transform. Consider the ODE pair

$$\dot{x}_1 = 0x_1 + x_2$$

$$\dot{x}_2 = -x_1 + 0x_2$$

with initial conditions $x_1(0) = 0$ and $x_2(0) = 1$.

(a) Laplace transform each equation above and use the initial data to show that $sX_1(s) = X_2(s)$ and $sX_2(s) - 1 = -X_1(s)$, where $X_1 = \mathcal{L}(x_1(t))$ and $X_2 = \mathcal{L}(x_2(t))$.

(b) Solve the algebraic equations $sX_1(s) = X_2(s)$ and $sX_2(s) - 1 = -X_1(s)$ to show that

$$X_1(s) = \frac{1}{s^2 + 1} \quad \text{and} \quad X_2(s) = \frac{s}{s^2 + 1}.$$

(c) Inverse transform $X_1(s)$ and $X_2(s)$ in part (b) to show that $x_1(t) = \sin(t)$ and $x_2(t) = \cos(t)$.

Verify that these satisfy the ODEs of interest.

(d) Use the Laplace transform to solve

$$\dot{x}_1 = 2x_1 + 3x_2 + 12e^{-t}$$

$$\dot{x}_2 = x_1 + 4x_2$$

with initial data $x_1(0) = -8$ and $x_2(0) = 2$.

Exercise 6.1.6 Let's set up a salt tank problem with two tanks. Refer to Figure 6.3, with tank volumes and flow rates as indicated. We assume that the concentration of salt in each tank is uniform over the tank volume (the tanks are well-stirred) at all times and that both tanks start filled with pure water.

If salt is conserved then the rate at which the amount of salt in a tank is changing should equal the rate salt enters the tank minus the rate salt exits the tank. Let $x_1(t)$ denote the mass (kg) of salt in tank 1 and $x_2(t)$ the mass of salt in tank 2.

- Verify that the total liquid volume in each tank remains constant at all times.
- Argue that the rate at which salt enters the first tank is $1/2$ kg per second (from the upper left inlet pipe) plus $5x_2/500$ kg per second from the second tank. Also argue that the rate at which salt exits tank 1 into tank 2 is $-10x_1/1200$ kg per second. It may help to first review the reasoning used in setting up the salt tank model in Section 2.1.3.
- Using the familiar observation that the rate at which the amount of salt in the tank is changing equals the rate salt enters minus the rate salt exits, argue that

$$\dot{x}_1 = \frac{1}{2} - \frac{x_1}{120} + \frac{x_2}{100}.$$

As a sanity check, verify that both sides of this equation have dimensions of mass per time.

- Employ similar reasoning to show that

$$\dot{x}_2 = \frac{x_1}{120} - \frac{x_2}{50}.$$

It would also be wise to examine the quantity $\dot{x}_1 + \dot{x}_2$, and think about why it makes sense.

- Verify that the functions

$$\begin{aligned} x_1(t) &= 120 - \frac{60}{13}e^{-t/40} - \frac{1500}{13}e^{-t/300} \\ x_2(t) &= 50 + \frac{100}{13}e^{-t/40} - \frac{750}{13}e^{-t/300} \end{aligned}$$

satisfy the coupled ODEs for $x_1(t)$ and $x_2(t)$ from parts (c) and (d), with $x_1(0) = 0$ and $x_2(0) = 0$. Plot $x_1(t)$ and $x_2(t)$ on the interval $0 \leq t \leq 2000$. What limiting values do they assume? What is the limiting concentration of salt in each tank, and how does this compare to the concentration of the incoming salt fluid in the upper left inlet pipe?

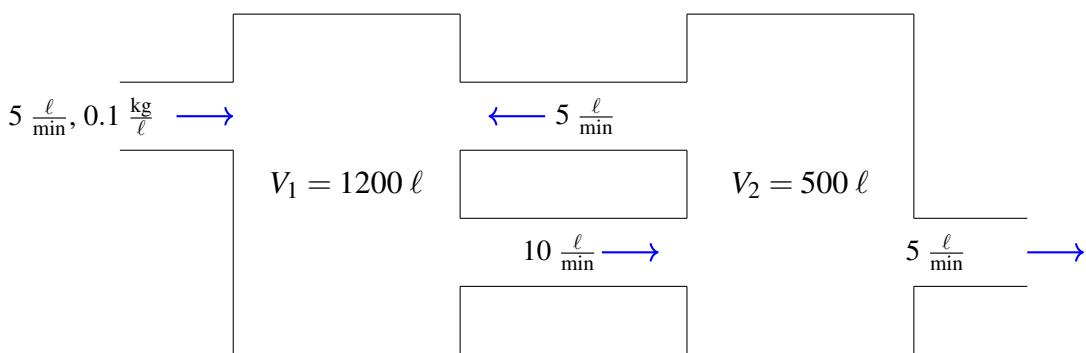


Figure 6.3: Illustration of salt tank problem with two tanks for Exercise 6.1.6, volumes and flow rates as indicated.

6.2 Linear Constant-Coefficient Homogeneous Systems of ODEs

In this section we will formulate linear systems of differential equations using matrix-vector notation, then consider how to solve them analytically using techniques based on eigenvalues and eigenvectors. We'll look at a few applications along the way. The essential background material on vectors and matrices is presented in Appendix B.

6.2.1 Matrix-Vector Formulation

Consider a linear system of ODEs of the form (6.4) in which the functions $a_{j,m}(t)$ are constant, so $a_{j,m}(t) = a_{j,m}$. This is a constant-coefficient linear system. Let $\mathbf{x}(t)$ denote the vector-valued function $\mathbf{x}(t) = \langle x_1(t), \dots, x_n(t) \rangle$ and $\mathbf{b}(t)$ denote the vector-valued function $\mathbf{b}(t) = \langle b_1(t), \dots, b_n(t) \rangle$. A constant-coefficient linear system of ODEs can be expressed conveniently in matrix-vector form as

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}(t) \quad (6.12)$$

where $\mathbf{x}(t)$ and $\mathbf{b}(t)$ in (6.12) are column vectors, $\dot{\mathbf{x}}(t)$ indicates component by component differentiation, and \mathbf{A} is the matrix with row j column m entry $a_{j,m}$ for $1 \leq j, m \leq n$. We will usually suppress the dependence of functions on t and write $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{b}$ for (6.12).

As with scalar linear equations, such a system may be homogeneous or nonhomogeneous.

Definition 6.2.1 The linear system of ODEs (6.12) is said to be **homogeneous** if $\mathbf{b}(t) = 0$ for all t . Otherwise the system is **nonhomogeneous**.

Example 6.3 The two-compartment model (6.1)-(6.2) can be written in the form of (6.12) with $\mathbf{x}(t) = \langle u_P(t), u_T(t) \rangle$, $\mathbf{b}(t) = \langle g(t), 0 \rangle$, and

$$\mathbf{A} = \begin{bmatrix} -k_b - k_e & k_a \\ k_b & -k_a \end{bmatrix}.$$

■

6.2.2 Solving the Homogeneous Case

Let's consider the solution to (6.12) in the homogeneous case $\mathbf{b}(t) = 0$. The system of interest is thus

$$\dot{\mathbf{x}} = \mathbf{Ax}. \quad (6.13)$$

Eigenvalues and Eigenvectors

To solve a system of the form (6.13) we will take inspiration from the scalar case $x'(t) = Ax(t)$ in which A is a constant scalar. Solutions to this ODE are easily verified to be of the form $x(t) = e^{At}c$ for any constant c ; we put the constant c on the right in $e^{At}c$ so the scalar case corresponds to the system case on the right in (6.14) below. Since a solution to (6.13) is a vector-valued function of t , by analogy to the scalar case we might try an ansatz for (6.13) of the form

$$\mathbf{x}(t) = e^{\lambda t} \mathbf{v} \quad (6.14)$$

for some constant vector $\mathbf{v} = \langle v_1, \dots, v_n \rangle$ that plays the role of c in the scalar equation.

Remark 6.2.1 If we truly emulate the scalar ODE $x' = Ax$ it would make even more sense to try an ansatz of the form $\mathbf{x}(t) = e^{At} \mathbf{v}$ in (6.13). This works, after we figure out what e^{At} might mean—that is, how to exponentiate a matrix. This is the topic of Section 6.4.

Based on (6.14) the components $x_j(t)$ of $\mathbf{x}(t)$ are $x_j(t) = e^{\lambda t} v_j$ and term by term differentiation of $\mathbf{x}(t)$ with respect to t shows that $\dot{x}_j = \lambda e^{\lambda t} v_j$, or

$$\dot{\mathbf{x}}(t) = \lambda e^{\lambda t} \mathbf{v}.$$

Substitute the above expression for $\dot{\mathbf{x}}$ into (6.13) along with (6.14) to replace $\mathbf{x}(t)$ and then pull the common $e^{\lambda t}$ out in front on both sides of the resulting equation (which is algebraically permitted since $e^{\lambda t}$ is a scalar) to arrive at

$$e^{\lambda t} \lambda \mathbf{v} = e^{\lambda t} \mathbf{A} \mathbf{v}.$$

Since $e^{\lambda t}$ is never 0 we can divide both sides of the above equation by this quantity, then reverse the left and right sides of the equation to obtain

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}.$$

That is, if $\mathbf{x}(t)$ in the form (6.14) satisfies $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$, then \mathbf{v} must be an **eigenvector** for \mathbf{A} with **eigenvalue** λ . You can easily check that the converse is true: if \mathbf{v} is an eigenvector for \mathbf{A} with eigenvalue λ then $\mathbf{x}(t) = e^{\lambda t} \mathbf{v}$ satisfies $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$. (For a discussion of basic matrix algebra, including eigenvalues and eigenvectors, see Appendix B.)

■ **Example 6.4** Consider the linear system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \underbrace{\begin{bmatrix} -5 & 3 \\ 1 & -3 \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

The 2×2 matrix \mathbf{A} has eigenvalues $\lambda_1 = -2$ and $\lambda_2 = -6$ with corresponding eigenvectors $\mathbf{v}_1 = \langle 1, 1 \rangle$ and $\mathbf{v}_2 = \langle -3, 1 \rangle$. This means that each of the vector-valued functions

$$\mathbf{w}_1(t) = e^{-2t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{w}_2(t) = e^{-6t} \begin{bmatrix} -3 \\ 1 \end{bmatrix}$$

satisfies the linear system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$. ■

Reading Exercise 6.2.1 The constants for the two-compartment LSD model of (6.1)-(6.2) as embodied by the matrix of Example 6.3 have been estimated as $k_a = 4.64$, $k_b = 3.19$, and $k_e = 0.411$ (all units reciprocal hours); we'll consider the problem of estimating these constants from data in Section 6.5.1. Construct the matrix \mathbf{A} in Example 6.3 and find its eigenvalues and eigenvectors. Use these to construct solutions $\mathbf{w}_1(t) = e^{\lambda_1 t} \mathbf{v}_1$ and $\mathbf{w}_2(t) = e^{\lambda_2 t} \mathbf{v}_2$ to this pair of ODEs. Why should the eigenvalues be negative (or at least have negative real part) in this physical context?

Incorporating the Initial Data

Let's start with an example of how we can find the solution with the desired initial data. We will construct a solution to the system of Example 6.4 with specified values for $x_1(0)$ and $x_2(0)$.

■ **Example 6.5** Let us find a solution to the linear system of Example 6.4 with initial conditions $x_1(0) = -1$ and $x_2(0) = 3$. The essential idea is to use linearity to construct a solution as a superposition of the functions $\mathbf{w}_1(t)$ and $\mathbf{w}_2(t)$ from that example. To see how and why this works, first note that by construction, $\dot{\mathbf{w}}_1 = \mathbf{A}\mathbf{w}_1$ and $\dot{\mathbf{w}}_2 = \mathbf{A}\mathbf{w}_2$. Define a linear combination

$$\mathbf{x}(t) = c_1 \mathbf{w}_1(t) + c_2 \mathbf{w}_2(t), \tag{6.15}$$

where c_1 and c_2 are arbitrary scalars. We can show that the function $\mathbf{x}(t)$ satisfies $\dot{\mathbf{x}} = \mathbf{Ax}$ for any choice of c_1 and c_2 . This can be done by using the linearity of differentiation and matrix-vector multiplication, so we see that

$$\begin{aligned}\dot{\mathbf{x}} &= c_1 \dot{\mathbf{w}}_1 + c_2 \dot{\mathbf{w}}_2 \\ &= c_1 \mathbf{Aw}_1 + c_2 \mathbf{Aw}_2 \\ &= \mathbf{A}(c_1 \mathbf{w}_1 + c_2 \mathbf{w}_2) \\ &= \mathbf{Ax}\end{aligned}\tag{6.16}$$

for any choice of c_1 and c_2 . The function $\mathbf{x}(t)$ in (6.15) is a **general solution** to the ODE system in Example 6.4, for we can adjust c_1 and c_2 to obtain a solution with any initial conditions.

For the present example with $x_1(0) = -1$ and $x_2(0) = 3$, use (6.15) to compute that

$$\begin{aligned}\mathbf{x}(0) &= c_1 \mathbf{w}_1(0) + c_2 \mathbf{w}_2(0) \\ &= c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 \\ &= \begin{bmatrix} c_1 - 3c_2 \\ c_1 + c_2 \end{bmatrix}.\end{aligned}$$

To satisfy the initial data we therefore need $c_1 - 3c_2 = -1$ and $c_1 + c_2 = 3$, two linear algebraic equations for c_1 and c_2 with solution $c_1 = 2$ and $c_2 = 1$. The solution with the required initial data is

$$\mathbf{x}(t) = 2e^{-2t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + e^{-6t} \begin{bmatrix} -3 \\ 1 \end{bmatrix},$$

or $x_1(t) = 2e^{-2t} - 3e^{-6t}$ and $x_2(t) = 2e^{-2t} + e^{-6t}$ if written out in component form. ■

The General Procedure for Solving $\dot{\mathbf{x}} = \mathbf{Ax}$

Examples 6.4 and 6.5 illustrate the simplest and most common case encountered when solving a homogeneous system of n linear, constant-coefficient ODEs $\dot{\mathbf{x}} = \mathbf{Ax}$ with initial data $\mathbf{x}(0) = \mathbf{x}_0$. In the procedure that follows we assume that the $n \times n$ matrix \mathbf{A} has n **linearly independent** eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$, as detailed in Section B.5 of Appendix B (see Remark B.5.1). This is always the case if the eigenvalues for \mathbf{A} are distinct (and may still be true if there are repeated eigenvalues). In the case in which \mathbf{A} is 2×2 it means that we can find two eigenvectors that are not scalar multiples of each other. The procedure to solve $\dot{\mathbf{x}} = \mathbf{Ax}$ is as follows:

1. Compute the eigenvalues λ_j and corresponding eigenvectors \mathbf{v}_j for \mathbf{A} . The vector-valued functions $\mathbf{w}_j(t) = e^{\lambda_j t} \mathbf{v}_j$ satisfy $\dot{\mathbf{w}}_j = \mathbf{Aw}_j$ for each j with $1 \leq j \leq n$.
2. Since $\dot{\mathbf{w}}_j = \mathbf{Aw}_j$, the same linearity argument that led to (6.16), shows that any superposition of the \mathbf{w}_j satisfies the ODE $\dot{\mathbf{x}} = \mathbf{Ax}$. Accordingly, define a vector-valued function as a superposition

$$\begin{aligned}\mathbf{x}(t) &= c_1 \mathbf{w}_1(t) + \cdots + c_n \mathbf{w}_n(t) \\ &= c_1 e^{\lambda_1 t} \mathbf{v}_1 + \cdots + c_n e^{\lambda_n t} \mathbf{v}_n,\end{aligned}\tag{6.17}$$

where the c_j are arbitrary constants. The function $\mathbf{x}(t)$ in (6.17) is a general solution to the linear system $\dot{\mathbf{x}} = \mathbf{Ax}$ because the c_j can be adjusted to obtain a solution with any desired initial data, as shown in the next step.

3. To find a solution with $\mathbf{x}(0) = \mathbf{x}_0$, substitute $t = 0$ into (6.17) to see that we need

$$c_1 \mathbf{v}_1 + \cdots + c_n \mathbf{v}_n = \mathbf{x}_0.$$

This equation is equivalent to the matrix-vector equation

$$\mathbf{P}\mathbf{c} = \mathbf{x}_0, \quad (6.18)$$

where \mathbf{P} is the $n \times n$ matrix whose j th column is \mathbf{v}_j and \mathbf{c} is the vector whose j th component is c_j . Since the eigenvectors \mathbf{v}_j for $1 \leq j \leq n$ are assumed to be linearly independent, the matrix \mathbf{P} is invertible (see Section B.5, Remark B.5.1). We can therefore solve for $\mathbf{c} = \mathbf{P}^{-1}\mathbf{x}_0$ and use these c_j in (6.17) to obtain the solution $\mathbf{x}(t)$.

Examples

■ **Example 6.6** Exercise 6.1.6 presented a two-compartment salt tank problem with tank volumes and flow rates as indicated in Figure 6.3. For the present example let us change the concentration of salt in the fluid entering Tank 1 to 0 kg per liter (pure water) so we have a homogeneous system. Then the resulting system that governs the amount of salt in each tank is now

$$\begin{aligned}\dot{x}_1 &= -\frac{x_1}{120} + \frac{x_2}{100} \\ \dot{x}_2 &= \frac{x_1}{120} - \frac{x_2}{50},\end{aligned}$$

where $x_1(t)$ is the amount of salt in Tank 1 and $x_2(t)$ is the amount of salt in Tank 2. For initial data we will take $x_1(0) = 1, x_2(0) = 7$, both in units of kilograms.

This linear constant-coefficient homogeneous system can be formulated as $\dot{\mathbf{x}} = \mathbf{Ax}$, where

$$\mathbf{A} = \begin{bmatrix} -1/120 & 1/100 \\ 1/120 & -1/50 \end{bmatrix}$$

and $\mathbf{x}(0) = \langle 1, 7 \rangle$. To solve this system first compute the eigenvalues and eigenvectors for \mathbf{A} to obtain $\lambda_1 = -1/300$ and $\lambda_2 = -1/40$ with corresponding eigenvectors $\mathbf{v}_1 = \langle 2, 1 \rangle$ and $\mathbf{v}_2 = \langle -3, 5 \rangle$ (or any multiples thereof). A general solution is then of the form (6.17),

$$\mathbf{x}(t) = c_1 e^{-t/300} \begin{bmatrix} 2 \\ 1 \end{bmatrix} + c_2 e^{-t/40} \begin{bmatrix} -3 \\ 5 \end{bmatrix}.$$

The last step is to construct a solution with the initial data by solving $\mathbf{x}(0) = c_1 \langle 2, 1 \rangle + c_2 \langle -3, 5 \rangle = \langle 1, 7 \rangle$; in the matrix-vector form of (6.18) this is equivalent to

$$\underbrace{\begin{bmatrix} 2 & -3 \\ 1 & 5 \end{bmatrix}}_{\mathbf{P}} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 7 \end{bmatrix}.$$

The matrix \mathbf{P} is invertible (equivalently, \mathbf{v}_1 and \mathbf{v}_2 are linearly independent) and we find a unique solution $c_1 = 2$ and $c_2 = 1$. The solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ is then

$$\mathbf{x}(t) = 2e^{-t/300} \begin{bmatrix} 2 \\ 1 \end{bmatrix} + e^{-t/40} \begin{bmatrix} -3 \\ 5 \end{bmatrix}.$$

In component form, $x_1(t) = 4e^{-t/300} - 3e^{-t/40}$ and $x_2(t) = 2e^{-t/300} + 5e^{-t/40}$. ■

Reading Exercise 6.2.2

- Use the functions $\mathbf{w}_1(t)$ and $\mathbf{w}_2(t)$ from Reading Exercise 6.2.1 to write out a general solution to the two-compartment LSD model (6.1)-(6.2).
- In the study [14] the subjects were given an IV dose of LSD equal to $2 \mu\text{g}$ per kilogram of body mass, so a 70 kg subject received a dose of $140 \mu\text{g}$. The drug was introduced into the bloodstream or plasma, so for such a subject $u_P(0) = 140 \mu\text{g}$ and $u_T(0) = 0 \mu\text{g}$. Use your general solution from part (a) to obtain this initial data and plot both $u_P(t)$ and $u_T(t)$ for $0 \leq t \leq 10$ hours. Comment—does the amount of drug in the plasma and tissue make sense?

6.2.3 Complex Eigenvalues

This solution procedure works when the eigenvalues and eigenvectors are complex. Let's look at an example.

■ **Example 6.7** Consider the underdamped spring-mass-damper system governed by $u''(t) + 2u'(t) + 5u(t) = 0$ with initial data $u(0) = 4$ and $u'(0) = 0$. Although we already know one way to solve this problem, let's approach it by converting it to a first-order system and using the approach based on eigenvalues. This second-order ODE can be formulated as a first-order system by setting $x_1 = u$ and $x_2 = u'$ to obtain

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -5 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (6.19)$$

with $x_1(0) = 4$ and $x_2(0) = 0$. The eigenvalues for the matrix in (6.19) are $\lambda_1 = -1 + 2i$ and $\lambda_2 = -1 - 2i$, with eigenvectors $\mathbf{v}_1 = \langle -1 - 2i, 5 \rangle$ and $\mathbf{v}_2 = \langle -1 + 2i, 5 \rangle$ (the eigenvectors can be multiplied by any nonzero scalar). The eigenvalues are complex conjugates, and the eigenvectors are too, component by component. From (6.17) we conclude that

$$\mathbf{x}(t) = c_1 e^{(-1+2i)t} \begin{bmatrix} -1 - 2i \\ 5 \end{bmatrix} + c_2 e^{(-1-2i)t} \begin{bmatrix} -1 + 2i \\ 5 \end{bmatrix} \quad (6.20)$$

is a general solution for this system. We can obtain the desired initial data by setting $\mathbf{x}(0) = \langle 4, 0 \rangle$, which leads to

$$\underbrace{\begin{bmatrix} -1 - 2i & -1 + 2i \\ 5 & 5 \end{bmatrix}}_{\mathbf{P}} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

(this is (6.18) in this example) with solution $c_1 = i$ and $c_2 = -i$; note that c_1 and c_2 are complex conjugates. The solution with $x_1(0) = 4$ and $x_2(0) = 0$ is then

$$\mathbf{x}(t) = ie^{(-1+2i)t} \begin{bmatrix} -1 - 2i \\ 5 \end{bmatrix} - ie^{(-1-2i)t} \begin{bmatrix} -1 + 2i \\ 5 \end{bmatrix}. \quad (6.21)$$

This solution is perfectly valid, but full of complex numbers, even though $\mathbf{x}(t)$ should be real-valued. The situation is reminiscent of that encountered when using the complex-valued general solution for solving underdamped spring-mass systems as in Section 4.2.4.

To obtain a real-valued solution, apply Euler's formula to write $e^{(-1+2i)t} = e^{-t}(\cos(2t) + i\sin(2t))$ and $e^{(-1-2i)t} = e^{-t}(\cos(2t) - i\sin(2t))$. Substitute this in (6.21) and collect like terms. After a bit of algebra we obtain

$$\mathbf{x}(t) = e^{-t} \sin(2t) \begin{bmatrix} 2 \\ -10 \end{bmatrix} + e^{-t} \cos(2t) \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

or $x_1(t) = 2e^{-t} \sin(2t) + 4e^{-t} \cos(2t)$ and $x_2(t) = -10e^{-t} \sin(2t)$ in component form. The solution is real-valued, as it should be.

The eigenvalues $-1 \pm 2i$ manifest themselves in this real-valued solution: the real part of the eigenvalues (-1) dictates the decay rate of the solution (the e^{-t} piece) and the imaginary parts (± 2) indicate the radial frequency of oscillation in the $\sin(2t)$ and $\cos(2t)$ pieces. ■

Real-Valued General Solutions

If the eigenvalues for \mathbf{A} are complex, then the general solution (6.17) will be complex-valued. However it is possible to write out a real-valued general solution. The procedure is similar to that used for second-order equations in Section 4.2.4. In what follows we focus on the case in which the system has two ODEs with two unknown functions. We then indicate how to handle higher-dimensional systems with an example.

First note that since \mathbf{A} has real entries, its eigenvalues and eigenvectors come in complex-conjugate pairs. If $\lambda = a + bi$ is an eigenvalue with complex eigenvector $\mathbf{v} = \mathbf{v}_r + i\mathbf{v}_i$ (split the eigenvector into a real part \mathbf{v}_r and imaginary part $i\mathbf{v}_i$), then the complex-valued general solution (6.17) can be expressed as

$$\mathbf{x}(t) = c_1 \mathbf{w}_1(t) + c_2 \mathbf{w}_2(t), \quad (6.22)$$

where

$$\mathbf{w}_1(t) = e^{(a+bi)t} (\mathbf{v}_r + i\mathbf{v}_i) \quad \text{and} \quad \mathbf{w}_2(t) = e^{(a-bi)t} (\mathbf{v}_r - i\mathbf{v}_i).$$

From the basic properties of complex conjugation in Appendix A, the vector-valued functions $\mathbf{w}_1(t)$ and $\mathbf{w}_2(t)$ must also be conjugate to each other, so that $\mathbf{w}_2(t) = \overline{\mathbf{w}_1(t)}$ (where $\overline{\mathbf{w}_1(t)}$ denotes the complex conjugate of $\mathbf{w}_1(t)$, component by component). Break $\mathbf{w}_1(t)$ into its real and imaginary parts component by component to form $\mathbf{y}_r(t) = \operatorname{Re}(\mathbf{w}_1(t))$ and $\mathbf{y}_i(t) = \operatorname{Im}(\mathbf{w}_1(t))$. Then

$$\mathbf{w}_1(t) = \mathbf{y}_r(t) + i\mathbf{y}_i(t)$$

and $\mathbf{w}_2(t) = \mathbf{y}_r(t) - i\mathbf{y}_i(t)$. The general solution (6.22) can then be expressed in the form

$$\begin{aligned} \mathbf{x}(t) &= c_1 \mathbf{w}_1(t) + c_2 \mathbf{w}_2(t) \\ &= c_1(\mathbf{y}_r(t) + i\mathbf{y}_i(t)) + c_2(\mathbf{y}_r(t) - i\mathbf{y}_i(t)) \\ &= \underbrace{(c_1 + c_2)}_{d_1} \mathbf{y}_r(t) + \underbrace{i(c_1 - c_2)}_{d_2} \mathbf{y}_i(t). \end{aligned} \quad (6.23)$$

Compare (6.23) to (4.32). Since c_1 and c_2 are arbitrary constants, so are d_1 and d_2 in (6.23), and (6.23) provides an alternate general solution to $\dot{\mathbf{x}} = \mathbf{Ax}$, but one that is real-valued since $\mathbf{y}_r(t)$ and $\mathbf{y}_i(t)$ are real-valued.

The essential observation is that the linear combination of the complex-valued functions $\mathbf{w}_1(t)$ and $\mathbf{w}_2(t)$ in (6.22) can be replaced by a linear combination of the real-valued functions $\operatorname{Re}(\mathbf{w}_1(t))$ and $\operatorname{Im}(\mathbf{w}_1(t))$. Let's summarize this in the following theorem.

Theorem 6.2.1 Suppose \mathbf{A} is a real-valued 2×2 matrix with complex eigenvalues λ_1 and λ_2 and corresponding eigenvectors \mathbf{v}_1 and \mathbf{v}_2 . A real-valued general solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ is

$$\mathbf{x}(t) = d_1 \mathbf{y}_r(t) + d_2 \mathbf{y}_i(t),$$

where $\mathbf{y}_r(t)$ and $\mathbf{y}_i(t)$ are the real and imaginary parts of $e^{\lambda_1 t} \mathbf{v}_1$. We can also take $\mathbf{y}_r(t)$ and $\mathbf{y}_i(t)$ as the real and imaginary parts of $e^{\lambda_2 t} \mathbf{v}_2$.

Compare the general solution of Theorem 6.2.1 to the real-valued general solution (4.33).

■ **Example 6.8** To illustrate Theorem 6.2.1, for the underdamped spring-mass system of Example 6.7 let us use eigenvalue $\lambda_1 = -1 + 2i$ and corresponding eigenvector $\mathbf{v}_1 = \langle -1 - 2i, 5 \rangle$. We compute the real and imaginary parts of $e^{\lambda_1 t} \mathbf{v}_1$, which yields

$$e^{(-1+2i)t} \begin{bmatrix} -1 - 2i \\ 5 \end{bmatrix} = e^{-2t} \begin{bmatrix} -\cos(2t) + 2\sin(2t) \\ 5\cos(2t) \end{bmatrix} + ie^{-2t} \begin{bmatrix} -2\cos(2t) - \sin(2t) \\ 5\sin(2t) \end{bmatrix}.$$

The two complex-valued pieces in (6.20) can be replaced by the real and imaginary parts on the right above. A real-valued general solution is then

$$\mathbf{x}(t) = \underbrace{d_1 e^{-2t} \begin{bmatrix} -\cos(2t) + 2\sin(2t) \\ 5\cos(2t) \end{bmatrix}}_{\text{Re}(e^{\lambda_1 t} \mathbf{v}_1)} + \underbrace{d_2 e^{-2t} \begin{bmatrix} -2\cos(2t) - \sin(2t) \\ 5\sin(2t) \end{bmatrix}}_{\text{Im}(e^{\lambda_1 t} \mathbf{v}_1)}.$$

■

Reading Exercise 6.2.3 Redo the computation of Example 6.8 using the real and imaginary parts of $e^{\lambda_2 t} \mathbf{v}_2$ where $\lambda_2 = -1 - 2i$ and $\mathbf{v}_2 = \langle -1 + 2i, 5 \rangle$ are the conjugate eigenvalue and eigenvector to λ_1 and \mathbf{v}_1 in that example. Convince yourself that the resulting real-valued general solution is equivalent to the real-valued general solution obtained in that example.

The next example illustrates how to construct a real-valued solution for a higher-dimensional system.

■ **Example 6.9** Consider the double spring-mass-damper system in Example 6.2, in particular, the displayed equations from (6.10) to (6.11). We will focus on the undamped case $c_1 = c_2 = 0$, and choose specific values $k_1 = 2, k_2 = 1, m_1 = 2$, and $m_2 = 1$. This system of four ODEs with dependent variables $w_1(t), w_2(t), w_3(t), w_4(t)$ can then be written as $\dot{\mathbf{w}} = \mathbf{Aw}$, where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -3/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 \end{bmatrix}. \quad (6.24)$$

This matrix has eigenvalues $\lambda_1 = i/\sqrt{2}, \lambda_2 = -i/\sqrt{2}, \lambda_3 = i\sqrt{2}$, and $\lambda_4 = -i\sqrt{2}$, with corresponding eigenvectors

$$\mathbf{v}_1 = \begin{bmatrix} -i/\sqrt{2} \\ 1/2 \\ -i\sqrt{2} \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} i/\sqrt{2} \\ 1/2 \\ i\sqrt{2} \\ 1 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} i/\sqrt{2} \\ -1 \\ -i/\sqrt{2} \\ 1 \end{bmatrix}, \quad \mathbf{v}_4 = \begin{bmatrix} -i/\sqrt{2} \\ -1 \\ i/\sqrt{2} \\ 1 \end{bmatrix}.$$

The eigenvalues and eigenvectors come in complex-conjugate pairs. A complex-valued general solution can be constructed using (6.17) in the form $\mathbf{w}(t) = \sum_{k=1}^4 c_k e^{\lambda_k t} \mathbf{v}_k$.

As an alternative to the complex-exponential general solution, we can employ the same reasoning that led to Theorem 6.2.1 in the case that \mathbf{A} was 2×2 . Specifically, replace the terms $e^{\lambda_1 t} \mathbf{v}_1$ and $e^{\lambda_2 t} \mathbf{v}_2$ in the general solution (that involves conjugate eigenvalues λ_1 and λ_2) with a linear combination of $\text{Re}(e^{\lambda_1 t} \mathbf{v}_1)$ and $\text{Im}(e^{\lambda_1 t} \mathbf{v}_1)$. Also replace the terms $e^{\lambda_3 t} \mathbf{v}_3$ and $e^{\lambda_4 t} \mathbf{v}_4$ (that involve conjugate eigenvalues λ_3 and λ_4) in the general solution with a linear combination of $\text{Re}(e^{\lambda_3 t} \mathbf{v}_3)$ and $\text{Im}(e^{\lambda_3 t} \mathbf{v}_3)$. For notational convenience define $\alpha = \sqrt{2}$. A bit of algebra shows that

$$\begin{aligned} \text{Re}(e^{\lambda_1 t} \mathbf{v}_1) &= \begin{bmatrix} \sin(t/\alpha)/\alpha \\ \cos(t/\alpha)/2 \\ \alpha \sin(t/\alpha) \\ \cos(t/\alpha) \end{bmatrix}, \quad \text{Im}(e^{\lambda_1 t} \mathbf{v}_1) = \begin{bmatrix} -\cos(t/\alpha)/\alpha \\ \sin(t/\alpha)/2 \\ -\alpha \cos(t/\alpha) \\ \sin(t/\alpha) \end{bmatrix} \\ \text{Re}(e^{\lambda_3 t} \mathbf{v}_3) &= \begin{bmatrix} -\sin(\alpha t)/\alpha \\ -\cos(\alpha t) \\ \sin(\alpha t)/\alpha \\ \cos(\alpha t) \end{bmatrix}, \quad \text{Im}(e^{\lambda_3 t} \mathbf{v}_3) = \begin{bmatrix} \cos(\alpha t)/\alpha \\ -\sin(\alpha t) \\ -\cos(\alpha t)/\alpha \\ \sin(\alpha t) \end{bmatrix}. \end{aligned}$$

These can be assembled into a real-valued general solution

$$\mathbf{w}(t) = c_1 \begin{bmatrix} \sin(t/\alpha)/\alpha \\ \cos(t/\alpha)/2 \\ \alpha \sin(t/\alpha) \\ \cos(t/\alpha) \end{bmatrix} + c_2 \begin{bmatrix} -\cos(t/\alpha)/\alpha \\ \sin(t/\alpha)/2 \\ -\alpha \cos(t/\alpha) \\ \sin(t/\alpha) \end{bmatrix} + c_3 \begin{bmatrix} -\sin(\alpha t)/\alpha \\ -\cos(\alpha t) \\ \sin(\alpha t)/\alpha \\ \cos(\alpha t) \end{bmatrix} + c_4 \begin{bmatrix} \cos(\alpha t)/\alpha \\ -\sin(\alpha t) \\ -\cos(\alpha t)/\alpha \\ \sin(\alpha t) \end{bmatrix}. \quad (6.25)$$

The solution is oscillatory and never decays, because there's no damping here. The solution components $w_1(t)$ and $w_3(t)$ (the mass positions) with initial data $\mathbf{w}(0) = \langle 1, 0, 0, 0 \rangle$ are shown in Figure 6.4, $w_1(t)$ as the solid red curve, $w_3(t)$ as the dashed black curve. There are two frequencies present in the solution, $1/\alpha \approx 0.707$ and $\alpha \approx 1.41$ radians per unit time. This is typical of a double spring-mass system—there are two natural frequencies. More generally, a system with n masses typically has n natural frequencies. When the physical parameters are chosen carefully, this kind of system can be designed to rapidly damp out or resist vibration; see the project “Tuned Mass Dampers” in Section 6.5.3. ■

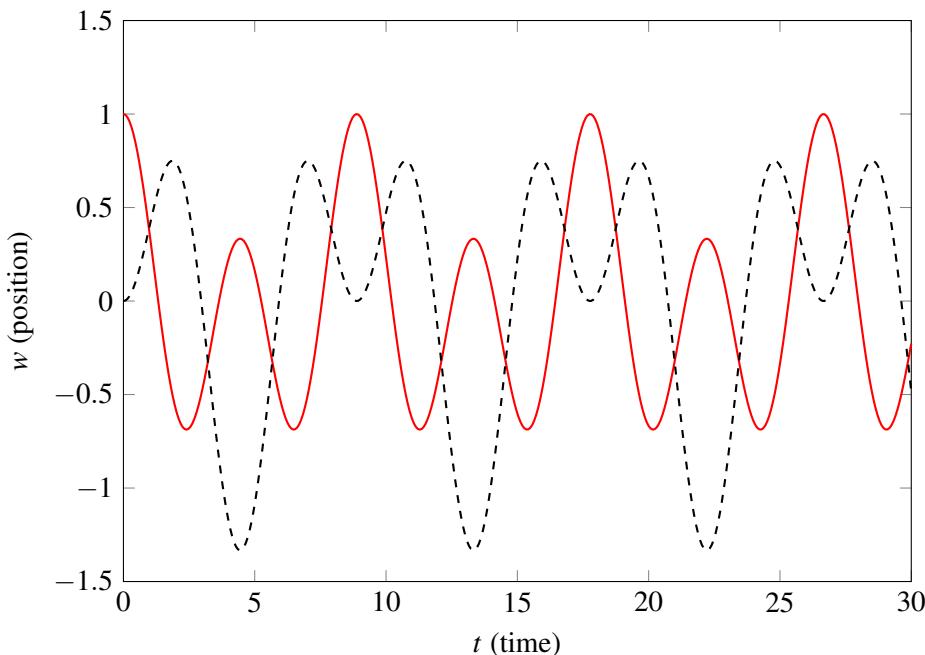


Figure 6.4: Motion of the masses in double spring-mass system (6.10)-(6.11), position $w_1(t)$ of mass 1 as red solid curve, position $w_3(t)$ of mass 2 as dashed black curve.

Reading Exercise 6.2.4 Show that with the general solution $\mathbf{w}(t)$ in (6.25), the initial condition $\mathbf{w}(0) = \langle 1, 0, 0, 0 \rangle$ leads to the equations $(-c_2 + c_4)/\alpha = 1$, $c_1/2 - c_3 = 0$, $-\alpha c_2 - c_4/\alpha = 0$, and $c_1 + c_3 = 0$, where $\alpha = \sqrt{2}$. Solve these equations and write out $\mathbf{w}(t)$ explicitly.

6.2.4 Defective Matrices

The above analysis is predicated on the $n \times n$ matrix \mathbf{A} having n linearly independent eigenvectors. That condition ensures that the system (6.18) can be solved to obtain any desired initial conditions, because the matrix \mathbf{P} will be invertible. But an $n \times n$ matrix may not possess n linearly independent eigenvectors; such matrices are said to be **defective**.

A Specific Example

Let's begin by considering a system of ODEs governed by a 2×2 defective matrix.

■ **Example 6.10** Consider the critically damped mass-spring system governed by

$$u''(t) + 4u'(t) + 4u(t) = 0, \quad (6.26)$$

which was analyzed in Example 4.10. If we let $x_1 = u$ and $x_2 = u'$ then (6.26) is equivalent to the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ -4 & -4 \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

The matrix \mathbf{A} has eigenvalues $\lambda_1 = \lambda_2 = -2$; that is to say it has a double eigenvalue $\lambda = -2$, but the only eigenvector is $\mathbf{v} = \langle 1, -2 \rangle$ or nonzero multiples thereof. As such, we can produce a solution

$$\mathbf{x}(t) = c_1 e^{-2t} \begin{bmatrix} 1 \\ -2 \end{bmatrix},$$

where c_1 is an arbitrary constant, but this is not a general solution. The initial data $\mathbf{x}(0) = c_1 \langle 1, -2 \rangle = \mathbf{x}_0$ can be obtained only when \mathbf{x}_0 is a scalar multiple of $\langle 1, -2 \rangle$. ■

In Example 6.10 the eigenvalue technique for producing two independent solutions fails. To address this problem recall how we attacked the critically damped case for the harmonic oscillator in Section 4.2, in particular, the procedure that led to (4.39). In that case the characteristic equation had a double root at λ and we were able to produce a solution $x(t) = ce^{\lambda t}$ to the ODE, but not a second independent solution. The remedy was to consider the ansatz $x(t) = (c_1 + c_2 t)e^{\lambda t}$, and this yielded another solution $x(t) = te^{\lambda t}$. If this approach worked once, we should try it again.

A General Solution for the Defective 2×2 Case

Let's focus on the 2×2 case. Let \mathbf{A} be a 2×2 matrix with double eigenvalue λ and only one eigenvector \mathbf{v} (or nonzero multiples thereof). We know how to produce a solution to this system, of the form

$$\mathbf{w}_1(t) = e^{\lambda t} \mathbf{v}. \quad (6.27)$$

The goal now is to produce a second solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ that is not simply a multiple of $\mathbf{w}_1(t)$. To produce such a solution we will try an ansatz of the form

$$\mathbf{w}_2(t) = e^{\alpha t} (\mathbf{v}_1 + t\mathbf{v}_2) \quad (6.28)$$

in $\dot{\mathbf{x}} = \mathbf{Ax}$, where the scalar α and vectors \mathbf{v}_1 and \mathbf{v}_2 are to be determined.

To figure out what is needed from α , \mathbf{v}_1 , and \mathbf{v}_2 , substitute $\mathbf{x}(t) = \mathbf{w}_2(t)$ into $\dot{\mathbf{x}} = \mathbf{Ax}$ and use the product and chain rules to compute

$$\dot{\mathbf{w}}_2 = e^{\alpha t} (\alpha \mathbf{v}_1 + \alpha t \mathbf{v}_2 + \mathbf{v}_2).$$

Substitute this expression for $\dot{\mathbf{x}}$ into $\dot{\mathbf{x}} = \mathbf{Ax}$ along with $\mathbf{x} = \mathbf{w}_2$ and make use of (6.28). After dividing through by $e^{\alpha t}$ we find

$$\alpha \mathbf{v}_1 + t \alpha \mathbf{v}_2 + \mathbf{v}_2 = \mathbf{Av}_1 + t \mathbf{Av}_2. \quad (6.29)$$

The goal is to make both sides of (6.29) identical as functions of t . If we choose α and \mathbf{v}_2 so that $\mathbf{Av}_2 = \alpha \mathbf{v}_2$, then the corresponding terms in (6.29) with the t coefficients will cancel. Of course

$\mathbf{A}\mathbf{v}_2 = \alpha\mathbf{v}_2$ means that \mathbf{v}_2 should be an eigenvector for \mathbf{A} with eigenvalue α . Since \mathbf{A} has only the eigenvalue λ and corresponding eigenvector \mathbf{v} , we must take $\alpha = \lambda$ and $\mathbf{v}_2 = \mathbf{v}$ in (6.29).

With these choices (6.29) becomes $\lambda\mathbf{v}_1 + t\lambda\mathbf{v} + \mathbf{v}_2 = \mathbf{Av}_1 + t\mathbf{Av}$, where $t\lambda\mathbf{v} = t\mathbf{Av}$. Cancelling these equivalent terms produces

$$\lambda\mathbf{v}_1 + \mathbf{v} = \mathbf{Av}_1$$

or, after a bit of rearrangement,

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_1 = \mathbf{v} \quad (6.30)$$

where \mathbf{I} is the identity matrix. It would be nice to assert that we can take $\mathbf{v}_1 = (\mathbf{A} - \lambda\mathbf{I})^{-1}\mathbf{v}$, but because λ is an eigenvalue for \mathbf{A} the matrix $(\mathbf{A} - \lambda\mathbf{I})$ cannot be invertible. There is no obvious reason to believe that (6.30) has a solution for \mathbf{v}_1 .

But it does. It can be shown that if λ is a double eigenvalue for a defective matrix \mathbf{A} and \mathbf{v} is an eigenvector, then (6.30) has a solution for \mathbf{v}_1 , and in fact there are infinitely many solutions. For a proof of this fact see [19]. We can use \mathbf{v}_1 in the ansatz (6.28) along with $\alpha = \lambda$ and $\mathbf{v}_2 = \mathbf{v}$ to produce a second solution to $\dot{\mathbf{x}} = \mathbf{Ax}$, specifically

$$\mathbf{w}_2(t) = e^{\lambda t}(\mathbf{v}_1 + t\mathbf{v}).$$

The solution $\mathbf{w}_2(t)$ can be used in conjunction $\mathbf{w}_1(t)$ (as defined by (6.27)) to construct a general solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ as a linear combination $\mathbf{x}(t) = c_1\mathbf{w}_1(t) + c_2\mathbf{w}_2(t)$. Let us state this as a theorem.

Theorem 6.2.2 Let \mathbf{A} be a 2×2 matrix with defective double eigenvalue λ and eigenvector \mathbf{v} . A general solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ is given by

$$\mathbf{x}(t) = c_1 e^{\lambda t} \mathbf{v} + c_2 e^{\lambda t} (\mathbf{v}_1 + t\mathbf{v}) \quad (6.31)$$

where \mathbf{v}_1 satisfies $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_1 = \mathbf{v}$ (such a vector \mathbf{v}_1 must exist).

It should be emphasized that for the formula (6.31) to be a general solution, it is necessary that λ is a double eigenvalue with eigenvector \mathbf{v} for the 2×2 defective matrix \mathbf{A} .

■ **Example 6.11** Let's return to the critically damped spring-mass system of Example 6.10 in which \mathbf{A} is defective with double eigenvalue $\lambda = -2$ and eigenvector $\mathbf{v} = \langle 1, -2 \rangle$. To find \mathbf{v}_1 in the general solution (6.31), note that (6.30) becomes $(\mathbf{A} + 2\mathbf{I})\mathbf{v}_1 = \mathbf{v}$ or

$$\begin{bmatrix} 2 & 1 \\ -4 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix},$$

where $\mathbf{v}_1 = \langle x, y \rangle$. There are infinitely many solutions to this system; all that is required is that $2x + y = 1$, so for example, $\mathbf{v}_1 = \langle -1, 3 \rangle$ works. Thus according to (6.31),

$$\mathbf{x}(t) = c_1 e^{-2t} \begin{bmatrix} 1 \\ -2 \end{bmatrix} + c_2 e^{-2t} \left(\begin{bmatrix} -1 \\ 3 \end{bmatrix} + t \begin{bmatrix} 1 \\ -2 \end{bmatrix} \right)$$

is a general solution to $\dot{\mathbf{x}} = \mathbf{Ax}$. To see that this is a general solution, note that the initial condition $\mathbf{x}(0) = \langle a, b \rangle$ leads to $\mathbf{x}(0) = \langle c_1 - c_2, -2c_1 + 3c_2 \rangle = \langle a, b \rangle$ or $c_1 - c_2 = a$ and $-2c_1 + 3c_2 = b$. For any initial data a and b these equations are always uniquely solvable for c_1 and c_2 .

For example, with initial data $\mathbf{x}_0 = \langle 1, -5 \rangle$ we find that $c_1 - c_2 = 1$ and $-2c_1 + 3c_2 = -5$ has solution $c_1 = -2, c_2 = -3$. The solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ with $\mathbf{x}(0) = \langle 1, -5 \rangle$ is then

$$\mathbf{x}(t) = -2e^{-2t} \begin{bmatrix} 1 \\ -2 \end{bmatrix} - 3e^{-2t} \left(\begin{bmatrix} -1 \\ 3 \end{bmatrix} + t \begin{bmatrix} 1 \\ -2 \end{bmatrix} \right)$$

or $x_1(t) = e^{-2t} - 3te^{-2t}, x_2(t) = -5e^{-2t} + 6te^{-2t}$ in component form. ■

Higher-Order Defective Cases

The difficulty for the 2×2 case analyzed above can occur in larger systems of ODEs as well. An $n \times n$ matrix \mathbf{A} may have distinct eigenvalues $\lambda_1, \dots, \lambda_m$ where $m < n$, and the number of linearly independent eigenvectors for \mathbf{A} may be less than n . In this case we cannot construct a general solution in the form (6.17). The procedure that led to Theorem 6.2.2 can be adapted to handle this case, but it requires a more detailed analysis of \mathbf{A} and some more sophisticated matrix algebra. We will not pursue this here, but will note that the method of Laplace transforms and the technique of matrix exponentiation in the next section can each handle these cases.

6.2.5 Exercises

Exercise 6.2.1 For each system of ODE in (a)-(g), do the following:

- Formulate the system as $\dot{\mathbf{x}} = \mathbf{Ax}$, by explicitly writing out the matrix \mathbf{A} .
 - Find the eigenvalues and eigenvectors of \mathbf{A} and use them to write out a general solution using (6.17).
 - If any eigenvalues for \mathbf{A} are complex, write out a real-valued general solution.
 - Use either form of the general solution to obtain the given initial data.
- (a) $\dot{x}_1 = 7x_1 - 4x_2, \dot{x}_2 = 20x_1 - 11x_2$ with $x_1(0) = 3$ and $x_2(0) = 8$.
 - (b) $\dot{x}_1 = -x_2, \dot{x}_2 = 6x_1 - 5x_2$ with $x_1(0) = 2$ and $x_2(0) = 5$.
 - (c) $\dot{x}_1 = x_1 - x_2, \dot{x}_2 = 5x_1 - 3x_2$ with $x_1(0) = 0$ and $x_2(0) = 2$.
 - (d) $\dot{x}_1 = -2x_1 - 3x_2, \dot{x}_2 = 3x_1 - 2x_2$ with $x_1(0) = 2$ and $x_2(0) = -2$.
 - (e) $\dot{x}_1 = -6x_1 + 9x_2 - 4x_3, \dot{x}_2 = -6x_1 + 11x_2 - 6x_3, \dot{x}_3 = -10x_1 + 21x_2 - 12x_3$ with $x_1(0) = -1, x_2(0) = 0$, and $x_3(0) = 2$.
 - (f) $\dot{x}_1 = -7x_1 + 2x_2 + 6x_3, \dot{x}_2 = -6x_1 - x_2 + 4x_3, \dot{x}_3 = -9x_1 + 2x_2 + 8x_3$ with $x_1(0) = -2, x_2(0) = 2$, and $x_3(0) = -4$.
 - (g) $\dot{x}_1 = -4x_1 - x_2 + 2x_3 - x_4, \dot{x}_2 = x_1 - x_3 + x_4, \dot{x}_3 = -x_3, \dot{x}_4 = x_1 - x_2 - 2x_4$ with $x_1(0) = 2, x_2(0) = 1, x_3(0) = 4$, and $x_4(0) = 1$.

Exercise 6.2.2 The systems in (a)-(d) involve defective matrices. For each system:

- Formulate the system as $\dot{\mathbf{x}} = \mathbf{Ax}$, by explicitly writing out the matrix \mathbf{A} .
 - Find the eigenvalues and eigenvectors of \mathbf{A} and use (6.31) (with (6.30)) to find a general solution.
 - Use the general solution to obtain the given initial data.
- (a) $\dot{x}_1 = 3x_1 - x_2, \dot{x}_2 = 4x_1 - x_2$ with $x_1(0) = 1$ and $x_2(0) = 3$.
 - (b) $\dot{x}_1 = 7x_1 - 3x_2, \dot{x}_2 = 12x_1 - 5x_2$ with $x_1(0) = 1$ and $x_2(0) = 1$.
 - (c) $\dot{x}_1 = -10x_1 - 8x_2, \dot{x}_2 = 8x_1 + 6x_2$ with $x_1(0) = 2$ and $x_2(0) = 0$.
 - (d) $\dot{x}_1 = 6x_1 + 5x_2 + 4x_3, \dot{x}_2 = -2x_1 - x_2 - x_3, \dot{x}_3 = -6x_1 - 6x_2 - 5x_3$ with $x_1(0) = 1, x_2(0) = 2$, and $x_3(0) = -1$. (This is a 3×3 system, but the technique of Section 6.2.4 can easily be adapted.)

Exercise 6.2.3 Consider a spring-mass system governed by $mx''(t) + cx'(t) + kx(t) = 0$.

- (a) Let $m = 1, c = 3$, and $k = 2$. Write out the characteristic equation for this ODE, find the roots, and explicitly write out a general solution.
- (b) Again with $m = 1, c = 3$, and $k = 2$, convert this second-order ODE into a pair of coupled first-order ODEs; use $x_1 = x$ and $x_2 = x'$. Formulate this system as $\dot{\mathbf{x}} = \mathbf{Ax}$ by writing out

the matrix \mathbf{A} explicitly.

- (c) Find a general solution to the system in part (b) by using eigenvector techniques. Verify that $x_1(t)$ has the same form as $x(t)$ in part (a). How are the eigenvalues of \mathbf{A} related to the roots of the characteristic equation?
- (d) Formulate the general equation $mx''(t) + cx'(t) + kx(t) = 0$ (with no specific choices for m , c , and k) as a first-order system $\dot{\mathbf{x}} = \mathbf{Ax}$. Show that the eigenvalues λ_1 and λ_2 of \mathbf{A} are exactly the roots r_1 and r_2 of the characteristic equation for the second-order ODE. Verify that the eigenvectors of \mathbf{A} are

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ \lambda_1 \end{bmatrix} \quad \text{and} \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ \lambda_2 \end{bmatrix},$$

(or multiples thereof, e.g., $\mathbf{v}_1 = \langle 1/\lambda_1, 1 \rangle$). Verify that the general solution for $x_1(t)$ is equivalent to that for $x(t)$. (Assume the system is not critically damped.)

Exercise 6.2.4 Consider the system of linear ODEs $\dot{x}_1 = x_2$, $\dot{x}_2 = -x_2$ and $\dot{x}_3 = x_1$.

- (a) Formulate this system as $\dot{\mathbf{x}} = \mathbf{Ax}$.
- (b) Find the eigenvalues and eigenvectors for \mathbf{A} and show that it is defective.
- (c) Find the general solution to the system and show that solutions may be constant, grow without bound, or decay to zero, depending on the initial conditions. (This is a 3×3 system, but the technique of Section 6.2.4 can easily be adapted.)

Exercise 6.2.5 Consider the double spring-mass-damper system of (6.10)-(6.11).

- (a) Write out the 4×4 matrix \mathbf{A} that governs this linear system $\dot{\mathbf{w}} = \mathbf{Aw}$.
- (b) With $k_1 = 2, k_2 = 4, m_1 = 1, m_2 = 1, c_1 = 1/2$, and $c_2 = 1/2$, compute the eigenvalues of \mathbf{A} . (You'll want to use a computer, and report numerical approximations.) What do they say about the motion of the system?
- (c) Repeat part (b) using $c_1 = c_2 = 10$ (leave k_1, k_2, m_1, m_2 the same as part (b)) and compute the eigenvalues of \mathbf{A} . What do they say about the motion of the system? What has changed from part (b)? Why does this make sense?

Exercise 6.2.6 Consider a two-compartment salt tank problem in the arrangement of Figure 6.5 (the inlet pipe on the upper left carries fresh water), with volumes and flow rates as indicated. Let $x_1(t)$ denote the amount of salt in Tank 1 and $x_2(t)$ the amount of salt in Tank 2. Suppose Tank 1 starts with 10 kg of salt and Tank 2 with 5 kg of salt at time $t = 0$.

- (a) Formulate this as a homogeneous linear system of ODEs for x_1 and x_2 , and then cast it in the form $\dot{\mathbf{x}} = \mathbf{Ax}$. Write out \mathbf{A} explicitly. What are the initial conditions?
- (b) Solve the system using eigenvalue techniques.
- (c) Solve the system using the method of Laplace transforms (see Exercise 6.1.5).

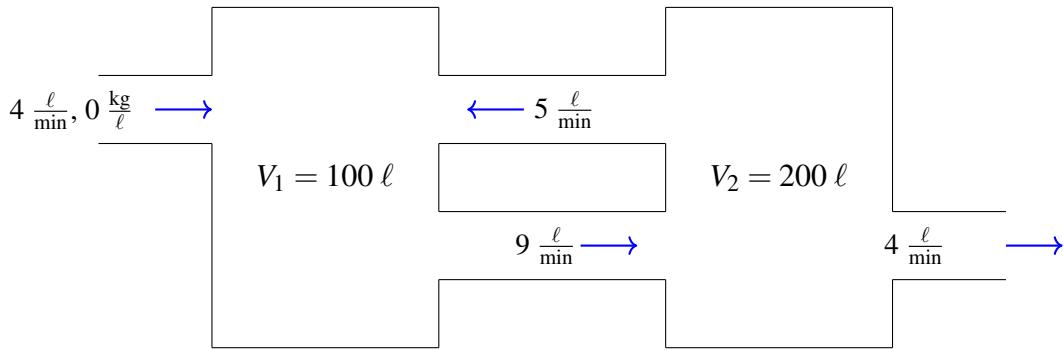


Figure 6.5: Illustration for Exercise 6.2.6, a salt tank problem with two tanks.

6.3 Linear Constant-Coefficient Nonhomogeneous Systems of ODEs

In this section we will use two different approaches to solve nonhomogeneous systems.

6.3.1 Solving Linear Systems of ODEs with Laplace Transforms

A nonhomogeneous system of linear constant-coefficient ODEs, as in Definition 6.2.1, can be handled using either the Laplace transform or by the method of undetermined coefficients. Both approaches are a very natural extension of the corresponding ideas for scalar equations. Let's first consider the method of Laplace transforms.

Laplace Transform Solution to Systems of ODEs

If you did Exercise 6.1.5, you've already seen this technique in action. Consider a system of n ODEs of the form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{f}, \quad (6.32)$$

where \mathbf{A} is an $n \times n$ matrix and $\mathbf{f} = \langle f_1(t), \dots, f_n(t) \rangle$. This system of ODEs can be Laplace transformed and turned into a system of algebraic equations for the Laplace transforms $X_1(s), \dots, X_n(s)$ of the solution components. Specifically, after Laplace transforming (and invoking the linearity of the Laplace transform) (6.32) becomes

$$s\mathbf{X}(s) - \mathbf{x}_0 = \mathbf{A}\mathbf{X}(s) + \mathbf{F}(s),$$

where $\mathbf{X}(s) = \langle X_1(s), \dots, X_n(s) \rangle$, $\mathbf{F}(s) = \langle F_1(s), \dots, F_n(s) \rangle$, and \mathbf{x}_0 is the vector of initial conditions. This system can be written as

$$(s\mathbf{I} - \mathbf{A})\mathbf{X}(s) = \mathbf{F}(s) + \mathbf{x}_0, \quad (6.33)$$

where \mathbf{I} denotes the $n \times n$ identity matrix. The matrix equation (6.33) embodies a system of n linear equations in the n unknowns $X_1(s), \dots, X_n(s)$. This algebraic system can then be solved for $\mathbf{X}(s)$ and then each component $X_k(s)$ inverse transformed to find $x_k(t)$. The approach of using Laplace transforms also has the advantage of allowing us to deal with discontinuous and impulsive driving functions.

Here is an example that illustrates the technique.

Example 6.12 Consider a salt tank problem with two tanks, volumes and flow rates as indicated in Figure 6.6. (See also Exercise 6.1.6). Let $x_1(t)$ denote the amount of salt in Tank 1 at time t and let $x_2(t)$ denote the amount of salt in Tank 2. As usual, we assume the tanks remain well-stirred. We also assume both tanks start filled with pure water.

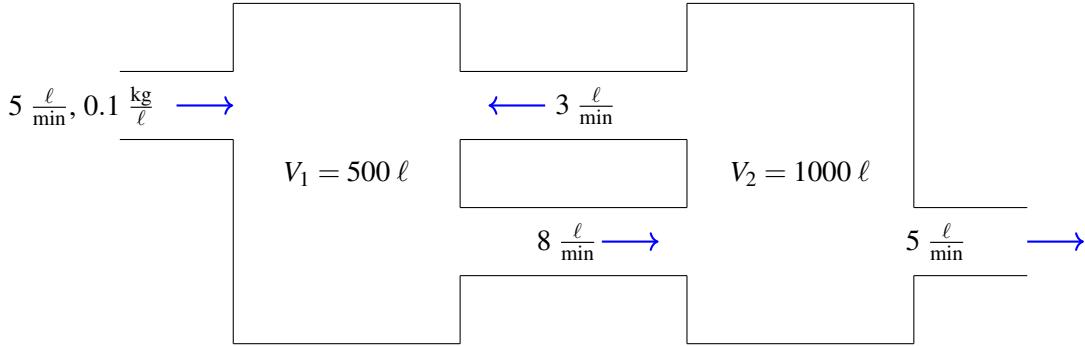


Figure 6.6: Illustration for Example 6.12, a salt tank problem with two tanks.

Salt enters the first tank through the upper left pipe at a rate of $(0.1 \frac{\text{kg}}{\ell}) \times (5 \frac{\ell}{\text{min}}) = 0.5 \frac{\text{kg}}{\text{min}}$. Salt also enters Tank 1 from Tank 2 at a rate of $(3 \frac{\ell}{\text{min}}) \times (\frac{x_2}{1000} \frac{\text{kg}}{\ell}) = \frac{3x_2}{1000} \frac{\text{kg}}{\text{min}}$. Salt exits Tank 1 into Tank 2 at a rate of $(8 \frac{\ell}{\text{min}}) \times (\frac{x_1}{500} \frac{\text{kg}}{\ell}) = \frac{2x_1}{125} \frac{\text{kg}}{\text{min}}$. Since salt is conserved we conclude that

$$\dot{x}_1 = -\frac{2}{125}x_1 + \frac{3}{1000}x_2 + \frac{1}{2}. \quad (6.34)$$

Applying similar conservation reasoning to Tank 2 yields

$$\dot{x}_2 = \frac{2}{125}x_1 - \frac{1}{125}x_2. \quad (6.35)$$

Laplace transforming the system (6.34)-(6.35) and using $x_1(0) = x_2(0) = 0$ (the tanks start filled with pure water) yields an algebraic system of equations

$$\begin{aligned} sX_1(s) &= -\frac{2}{125}X_1(s) + \frac{3}{1000}X_2(s) + \frac{1}{2s} \\ sX_2(s) &= \frac{2}{125}X_1(s) - \frac{1}{125}X_2(s), \end{aligned} \quad (6.36)$$

where $X_1(s) = \mathcal{L}(x_1(t))$ and $X_2(s) = \mathcal{L}(x_2(t))$. The system (6.36) can be solved for $X_1(s)$ and $X_2(s)$. We find

$$\begin{aligned} X_1(s) &= \frac{50(125s+1)}{s(12500s^2+300s+1)} \\ X_2(s) &= \frac{100}{s(12500s^2+300s+1)}. \end{aligned}$$

The denominator for X_1 and X_2 factors as $12500s(s+1/50)(s+1/250)$, so we expect to see terms $e^{-t/50}$, $e^{-t/250}$, and constants in the inverse transform. Inverse transforming shows that

$$\begin{aligned} x_1(t) &= 50 - \frac{75}{4}e^{-t/50} - \frac{125}{4}e^{-t/250} \\ x_2(t) &= 100 + 25e^{-t/50} - 125e^{-t/250}. \end{aligned}$$

The solution components $x_1(t)$ and $x_2(t)$ are plotted in Figure 6.7. The concentration of salt in each tank (obtained by dividing the amount of salt in each tank by that tank's volume) approaches the constant concentration of the incoming salt solution. ■

The Laplace transform offers a practical way to solve a nonhomogeneous system analytically, although a computer algebra system is helpful for the computations.

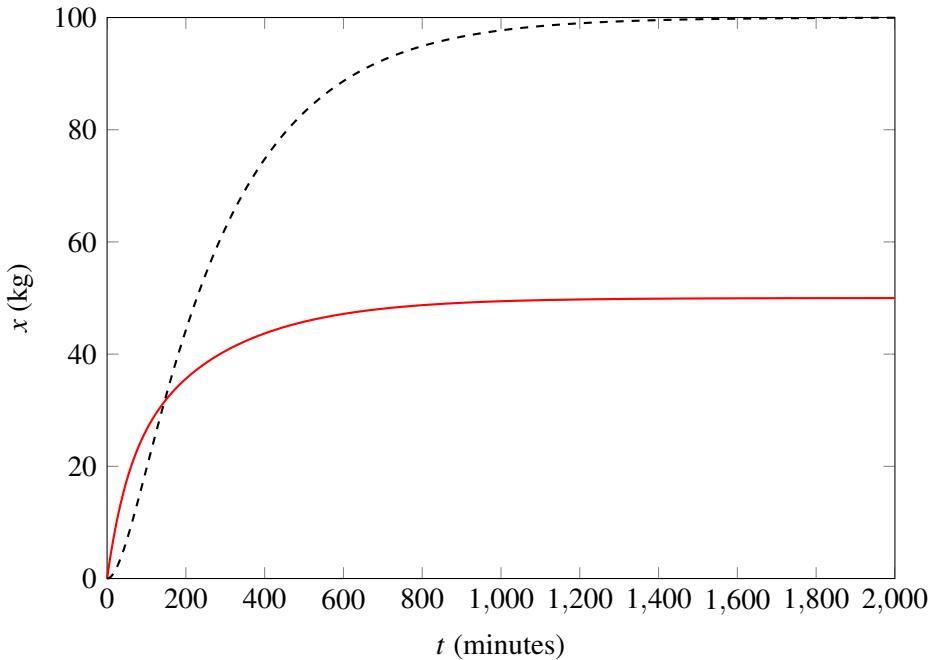


Figure 6.7: Concentration of salt in Tanks 1 and 2 in Example 6.12, $x_1(t)$ as solid red curve, $x_2(t)$ as dashed black curve.

Remark 6.3.1 When using the Laplace transform to solve a system, (6.33) gives the transform of the solution as

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{F}(s) + (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}_0, \quad (6.37)$$

where $(s\mathbf{I} - \mathbf{A})^{-1}$ is the inverse (symbolic) of the matrix $(s\mathbf{I} - \mathbf{A})$. If we solve a scalar ODE $\dot{x} = ax + f(t)$ with $x(0) = x_0$ using the Laplace transform we find that

$$X(s) = F(s)/(s - a) + x_0/(s - a). \quad (6.38)$$

Note the parallel structure of the scalar equation (6.38) with that of (6.37). Also observe that $1/(s - a)$ in (6.38) corresponds to e^{at} in the time domain. Does $(s\mathbf{I} - \mathbf{A})^{-1}$ correspond to some kind of exponential function involving \mathbf{A} ? What does it even mean to exponentiate a matrix? Stay tuned for more on this question in Section 6.4.

Reading Exercise 6.3.1 Solve the system

$$\begin{aligned} \dot{x}_1 &= 3x_1 - x_2 + 3e^t \\ \dot{x}_2 &= -x_1 + 3x_2 \end{aligned}$$

with initial conditions $x_1(0) = -1/2$ and $x_2(0) = 1/2$.

6.3.2 Undetermined Coefficients for Systems of ODEs

The method of undetermined coefficients from Section 4.3 can also be used to solve systems of nonhomogeneous, linear, constant-coefficient ODEs. If we embrace matrix notation, the process is essentially identical to the cases we've already examined in that section.

Producing a General Solution

Consider a system of n ODEs of the form (6.32). Let $\mathbf{x}_h(t)$ denote a general solution to the homogeneous system $\dot{\mathbf{x}} = \mathbf{Ax}$, formed according to (6.17) or other methods, as appropriate. This general solution comes with n arbitrary constants c_1, \dots, c_n that can be used to obtain a solution with any desired initial conditions. Let $\mathbf{x}_p(t)$ denote any particular solution to (6.32). We will use the method of undetermined coefficients to produce $\mathbf{x}_p(t)$ and then form the vector-valued function

$$\mathbf{x}(t) = \mathbf{x}_p(t) + \mathbf{x}_h(t). \quad (6.39)$$

The function $\mathbf{x}(t)$ is a general solution to $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f}$. To see this first note that $\mathbf{x}(t)$ satisfies $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f}$ because

$$\begin{aligned}\dot{\mathbf{x}} &= \dot{\mathbf{x}}_p + \dot{\mathbf{x}}_h \\ &= \mathbf{Ax}_p + \mathbf{f} + \mathbf{Ax}_h \\ &= \mathbf{Ax} + \mathbf{f}.\end{aligned}$$

Moreover, \mathbf{x} contains n arbitrary constants that can be used to obtain any desired initial conditions.

Finding a general solution to $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f}$ thus comes down to finding a general solution to the homogeneous problem $\dot{\mathbf{x}} = \mathbf{Ax}$ using eigenvalue techniques, and then producing a particular solution \mathbf{x}_p . Next we consider how to find such a particular solution.

Producing Particular Solutions with Undetermined Coefficients

Let's look at some typical examples of using undetermined coefficients for systems.

■ **Example 6.13** Consider the system (6.32) where

$$\mathbf{A} = \begin{bmatrix} 3 & -2 \\ 12 & -7 \end{bmatrix} \quad \text{and} \quad \mathbf{f}(t) = \begin{bmatrix} e^{-4t} \\ 3e^{-4t} \end{bmatrix}.$$

We seek a solution with initial data $x_1(0) = 1$ and $x_2(0) = 2$.

First, one can verify that \mathbf{A} has eigenvalues $\lambda_1 = -1, \lambda_2 = -3$, with eigenvectors $\mathbf{v}_1 = \langle 1, 2 \rangle$ and $\mathbf{v}_2 = \langle 1, 3 \rangle$, respectively. A general solution to the homogeneous system $\dot{\mathbf{x}} = \mathbf{Ax}$ is thus

$$\mathbf{x}_h(t) = c_1 e^{-t} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + c_2 e^{-3t} \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

We will now find a particular solution $\mathbf{x}_p(t)$ using the method of undetermined coefficients. The undetermined coefficients in the scalar equations of Section 4.3 are replaced by undetermined vectors. To illustrate, let's express $\mathbf{f}(t)$ as $\mathbf{f}(t) = e^{-4t} \mathbf{w}$ where $\mathbf{w} = \langle 1, 3 \rangle$. Based on the form of $\mathbf{f}(t)$ we use an ansatz in $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f}$ of the form

$$\mathbf{x}_p(t) = e^{-4t} \mathbf{v}$$

for some undetermined constant vector \mathbf{v} . To determine \mathbf{v} we substitute the ansatz $\mathbf{x}_p(t)$ into the nonhomogeneous ODE. From $\dot{\mathbf{x}}_p = -4e^{-4t} \mathbf{v}$ and the form of \mathbf{x}_p , we obtain

$$-4e^{-4t} \mathbf{v} = e^{-4t} \mathbf{Av} + e^{-4t} \mathbf{w}.$$

Divide both sides by e^{-4t} and rearrange to obtain

$$(\mathbf{A} + 4\mathbf{I})\mathbf{v} = -\mathbf{w}.$$

The matrix \mathbf{A} has eigenvalues -1 and -3 , so $\mathbf{A} + 4\mathbf{I}$ must be invertible or else -4 would be an eigenvalue for \mathbf{A} . We can then compute that

$$\mathbf{v} = -(\mathbf{A} + 4\mathbf{I})^{-1}\mathbf{w} = \begin{bmatrix} -1 \\ -3 \end{bmatrix}.$$

This yields a particular solution

$$\mathbf{x}_p(t) = e^{-4t} \begin{bmatrix} -1 \\ -3 \end{bmatrix}.$$

By using (6.39) a general solution to the nonhomogeneous system is

$$\mathbf{x}(t) = e^{-4t} \begin{bmatrix} -1 \\ -3 \end{bmatrix} + c_1 e^{-t} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + c_2 e^{-3t} \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

Requiring $\mathbf{x}(0) = \langle 1, 2 \rangle$ leads to the equations $-1 + c_1 + c_2 = 1$ and $-3 + 2c_1 + 3c_2 = 2$ with solution $c_1 = 1$ and $c_2 = 1$. The solution to the nonhomogeneous system is therefore

$$\mathbf{x}(t) = e^{-4t} \begin{bmatrix} -1 \\ -3 \end{bmatrix} + e^{-t} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + e^{-3t} \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

■

Reading Exercise 6.3.2 Consider the system

$$\begin{aligned} \dot{x}_1 &= 2x_1 + 3x_2 + 5e^t \\ \dot{x}_2 &= -6x_1 - 7x_2 + 10e^t. \end{aligned}$$

- (a) Formulate the system as $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f}(t)$ (write out \mathbf{A} and $\mathbf{f}(t)$ explicitly).
- (b) Find a particular solution $\mathbf{x}_p(t)$ of the form $\mathbf{x}_p(t) = e^t \mathbf{v}$ by emulating Example 6.13.
- (c) Compute the eigenvalues and eigenvectors of \mathbf{A} and find a general solution $\mathbf{x}_h(t)$ to $\dot{\mathbf{x}} = \mathbf{Ax}$.
- (d) Form a general solution $\mathbf{x}(t) = \mathbf{x}_p(t) + \mathbf{x}_h(t)$ for $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f}(t)$. Adjust the arbitrary constants in this general solution to obtain a particular solution with the required initial data $x_1(0) = 1$ and $x_2(0) = 3$.

■ **Example 6.14** Consider the double spring-mass system of Example 6.2. In Example 6.9 the undamped case with $k_1 = 2, k_2 = 1, m_1 = 2$, and $m_2 = 1$ was formulated as the homogeneous system $\dot{\mathbf{w}} = \mathbf{Aw}$ where \mathbf{A} was defined in (6.24). Suppose now that the second mass is acted on by an external force $f(t)$. In this case (6.11) becomes $\dot{w}_4 = w_1 - w_3 + f(t)$ (using $k_1 = 2, k_2 = 1, m_1 = 2$, and $m_2 = 1$) and the matrix formulation of the resulting ODE system becomes

$$\dot{\mathbf{w}} = \mathbf{Aw} + \mathbf{f}, \tag{6.40}$$

where \mathbf{f} is the vector

$$\mathbf{f} = \langle 0, 0, 0, f(t) \rangle.$$

We computed a general solution $\mathbf{w}_h(t)$ to the homogeneous equation $\dot{\mathbf{w}} = \mathbf{Aw}$ in Example 6.9. As a further illustration of the method of undetermined coefficients let's focus on finding a particular solution $\mathbf{w}_p(t)$ to $\dot{\mathbf{w}} = \mathbf{Aw} + \mathbf{f}$ in the case when $f(t) = \sin(\omega t)$. In this case

$$\mathbf{f} = \sin(\omega t) \mathbf{e}_4$$

where $\mathbf{e}_4 = \langle 0, 0, 0, 1 \rangle$ (the fourth standard basis vector in \mathbb{R}^4). Based on our experience with the scalar case, we should try something of the form

$$\mathbf{w}_p(t) = \cos(\omega t)\mathbf{a} + \sin(\omega t)\mathbf{b} \quad (6.41)$$

for some undetermined constant vectors \mathbf{a} and \mathbf{b} . Differentiating shows that

$$\dot{\mathbf{w}}_p = -\omega \sin(\omega t)\mathbf{a} + \omega \cos(\omega t)\mathbf{b}$$

and using this in $\dot{\mathbf{w}} = \mathbf{Aw} + \mathbf{f}$ yields,

$$-\omega \sin(\omega t)\mathbf{a} + \omega \cos(\omega t)\mathbf{b} = \cos(\omega t)\mathbf{A}\mathbf{a} + \sin(\omega t)\mathbf{A}\mathbf{b} + \sin(\omega t)\mathbf{e}_4. \quad (6.42)$$

As in the scalar case, match the $\sin(\omega t)$ and $\cos(\omega t)$ terms on the left and right in (6.42) to find that the vectors \mathbf{a} and \mathbf{b} must satisfy

$$\mathbf{Ab} + \omega\mathbf{a} = -\mathbf{e}_4 \quad (6.43)$$

$$\mathbf{Aa} - \omega\mathbf{b} = \mathbf{0}. \quad (6.44)$$

To find \mathbf{a} and \mathbf{b} , solve (6.44) for $\mathbf{b} = (\mathbf{Aa})/\omega$ and use this expression for \mathbf{b} in (6.43) to obtain

$$\frac{1}{\omega}\mathbf{A}^2\mathbf{a} + \omega\mathbf{a} = -\mathbf{e}_4,$$

which can be written as

$$(\mathbf{A}^2 + \omega^2\mathbf{I})\mathbf{a} = -\omega\mathbf{e}_4 \quad (6.45)$$

after multiplying through by ω . We can solve (6.45) to find the vector \mathbf{a} and then use (6.44) to compute $\mathbf{b} = (\mathbf{Aa})/\omega$. This assumes that the matrix $(\mathbf{A}^2 + \omega^2\mathbf{I})$ on the left in (6.45) is invertible, so that a solution for \mathbf{a} is guaranteed to exist.

As an illustration, suppose $\omega = 1$. In this case (6.45) becomes

$$\begin{bmatrix} -1/2 & 0 & 1/2 & 0 \\ 0 & -1/2 & 0 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \mathbf{a} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \end{bmatrix}$$

with solution $\mathbf{a} = \langle 0, -1, 0, -1 \rangle$. From $\mathbf{b} = (\mathbf{Aa})/\omega$ we find $\mathbf{b} = \langle -1, 0, -1, 0 \rangle$. A particular solution to (6.40) is given by $\mathbf{w}_p(t)$ in (6.41) and is

$$\mathbf{w}_p(t) = -\cos(t) \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} - \sin(t) \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix},$$

or $w_1(t) = -\sin(t)$, $w_2(t) = -\cos(t)$, $w_3(t) = -\sin(t)$ and $w_4(t) = -\cos(t)$ in component form. A general solution to $\dot{\mathbf{w}} = \mathbf{Aw} + \mathbf{f}$ in this case would be $\mathbf{w}(t) = \mathbf{w}_p(t) + \mathbf{w}_h(t)$ where $\mathbf{w}_h(t)$ is as in Example 6.9. ■

6.3.3 The Significance of Eigenvalues

The behavior of solutions to the linear system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is largely dictated by the eigenvalues of \mathbf{A} . From the general solution (6.17) we see that if all of the eigenvalues have negative real part (this includes negative real numbers, of course) then all solutions will decay to the zero vector as $t \rightarrow \infty$. If any eigenvalue λ_k has positive real part, the corresponding term $c_k e^{\lambda_k t} \mathbf{v}_k$ in the solution will grow without bound as $t \rightarrow \infty$, except in the unlikely case that $c_k = 0$. These conclusions hold even for defective matrices. In the case that the eigenvalues are all merely nonnegative, the situation is more delicate; see Exercise 6.2.4.

For the nonhomogeneous case in which the forcing $\mathbf{f}(t)$ is periodic and \mathbf{A} has eigenvalues with negative real part, the solution consists of a transient portion that decays on a time scale dictated by the real part of the eigenvalues, and a long-term periodic response dictated by the behavior of $\mathbf{f}(t)$. It should not be surprising that the linear physical systems we've encountered—damped spring-mass systems, compartment models, and RC circuits—give rise to matrices with eigenvalues that have negative real parts, since the action of these systems is expected to decay to zero (or periodic motion, if driven periodically). However, positive eigenvalues can arise when we approximate nonlinear systems, as we will do in the next chapter.

6.3.4 Exercises

Exercise 6.3.1 For each system of ODEs, Laplace transform the ODEs with the given initial conditions and solve the resulting algebraic equations to find the Laplace transform of each solution component. Then inverse transform to find each solution component $x_1(t), x_2(t), \dots$.

- $\dot{x}_1 = 7x_1 - 4x_2, \dot{x}_2 = 20x_1 - 11x_2$ with $x_1(0) = 3$ and $x_2(0) = 8$. Compare to the result of part (a) of Exercise 6.2.1.
- $\dot{x}_1 = 7x_1 - 4x_2 + 3e^{-2t}, \dot{x}_2 = 20x_1 - 11x_2 + 7e^{-2t}$ with $x_1(0) = 2$ and $x_2(0) = 3$.
- $\dot{x}_1 = -6x_1 + 2x_2 - 3e^{-3t}, \dot{x}_2 = -15x_1 + 5x_2 - 9e^{-3t}$ with $x_1(0) = 1$ and $x_2(0) = 2$.
- $\dot{x}_1 = 3x_1 - 2x_2 + 1 - 5t, \dot{x}_2 = 10x_1 - 6x_2 - 1 - 16t$ with $x_1(0) = 1$ and $x_2(0) = 2$.
- $\dot{x}_1 = 3x_1 - 2x_2 - \sin(t) - \cos(t), \dot{x}_2 = 10x_1 - 6x_2 - 3\sin(t)$ with $x_1(0) = 1$ and $x_2(0) = 3$.
- $\dot{x}_1 = 3x_1 - 2x_2 + 5\cos(t) - 3\sin(t), \dot{x}_2 = 10x_1 - 6x_2 + 12\cos(t) - 12\sin(t)$ with $x_1(0) = 0$ and $x_2(0) = 3$.
- $\dot{x}_1 = 2x_1 - 4x_2 - 3x_3 - 5, \dot{x}_2 = -x_1 - x_2 + x_3 + 2, \dot{x}_3 = 6x_1 - 6x_2 - 7x_3 - 13$ with $x_1(0) = 1$, $x_2(0) = 1$ and $x_3(0) = -1$.
- $\dot{x}_1 = 2x_1 - 4x_2 - 3x_3 - \sin(t) + \cos(t), \dot{x}_2 = -x_1 - x_2 + x_3, \dot{x}_3 = 6x_1 - 6x_2 - 7x_3 - \sin(t) + \cos(t)$ with $x_1(0) = 1$, $x_2(0) = 2$ and $x_3(0) = 1$.

Exercise 6.3.2 For each system of ODEs:

- Formulate the system as $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{f}(t)$ by explicitly writing out the matrix \mathbf{A} and function $\mathbf{f}(t)$.
 - Use the method of undetermined coefficients to find a particular solution $\mathbf{x}_p(t)$ to the system.
 - Find the eigenvalues and eigenvectors of \mathbf{A} and use them to write out a general solution to the homogeneous system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ using (6.17).
 - Form the general solution $\mathbf{x}(t)$ to the nonhomogeneous system and obtain the given initial data.
- $\dot{x}_1 = 7x_1 - 4x_2 + 3e^{-2t}, \dot{x}_2 = 20x_1 - 11x_2 + 7e^{-2t}$ with $x_1(0) = 2$ and $x_2(0) = 3$.
 - $\dot{x}_1 = -6x_1 + 2x_2 - 3e^{-3t}, \dot{x}_2 = -15x_1 + 5x_2 - 9e^{-3t}$ with $x_1(0) = 1$ and $x_2(0) = 2$.
 - $\dot{x}_1 = 3x_1 - 2x_2 + 2, \dot{x}_2 = 10x_1 - 6x_2 - 2$ with $x_1(0) = 1$ and $x_2(0) = 2$.

- (d) $\dot{x}_1 = 3x_1 - 2x_2 - \sin(t) - \cos(t)$, $\dot{x}_2 = 10x_1 - 6x_2 - 3\sin(t)$ with $x_1(0) = 1$ and $x_2(0) = 3$.

Hints for (d):

- Write $\mathbf{f}(t) = \cos(t)\mathbf{w}_1 + \sin(t)\mathbf{w}_2$ for suitable vectors \mathbf{w}_1 and \mathbf{w}_2 , and try an ansatz $\mathbf{x}_p(t) = \cos(t)\mathbf{v}_1 + \sin(t)\mathbf{v}_2$.
- Show that $-\mathbf{v}_1 = \mathbf{Av}_2 + \mathbf{w}_2$ and $\mathbf{v}_2 = \mathbf{Av}_1 + \mathbf{w}_1$.
- Use the last step to show that $(\mathbf{A}^2 + \mathbf{I})\mathbf{v}_1 = -(\mathbf{Aw}_1 + \mathbf{w}_2)$ and use this last equation to find \mathbf{v}_1 . Then determine \mathbf{v}_2 .

- (e) $\dot{x}_1 = 3x_1 - 2x_2 + 5\cos(t) - 3\sin(t)$, $\dot{x}_2 = 10x_1 - 6x_2 + 12\cos(t) - 12\sin(t)$ with $x_1(0) = 0$ and $x_2(0) = 3$. Hint: see the last exercise.

- (f) $\dot{x}_1 = -6x_1 + 9x_2 - 4x_3 - 4$, $\dot{x}_2 = -6x_1 + 11x_2 - 6x_3 - 4$, $\dot{x}_3 = -10x_1 + 21x_2 - 12x_3 - 8$ with $x_1(0) = -1$, $x_2(0) = 1$ and $x_3(0) = 1$.

- (g) $\dot{x}_1 = -6x_1 + 9x_2 - 4x_3 - 4e^t$, $\dot{x}_2 = -6x_1 + 11x_2 - 6x_3 - 4e^t$, $\dot{x}_3 = -10x_1 + 21x_2 - 12x_3 - 8e^t$ with $x_1(0) = 2$, $x_2(0) = 1$ and $x_3(0) = -2$.

Exercise 6.3.3 (This is a variation on Exercise 6.2.6.) Consider a two-compartment salt tank problem in the arrangement of Figure 6.8. Let $x_1(t)$ denote the amount of salt in tank 1 and $x_2(t)$ the amount of salt in tank 2. Suppose both tanks start filled with pure water.

- Formulate this salt tank problem as a nonhomogeneous linear system of ODEs in the form $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f}$. Explicitly write out \mathbf{A} and \mathbf{f} . What are the initial conditions?
- Solve the system using the method of undetermined coefficients.
- Solve the system using the method of Laplace transforms.

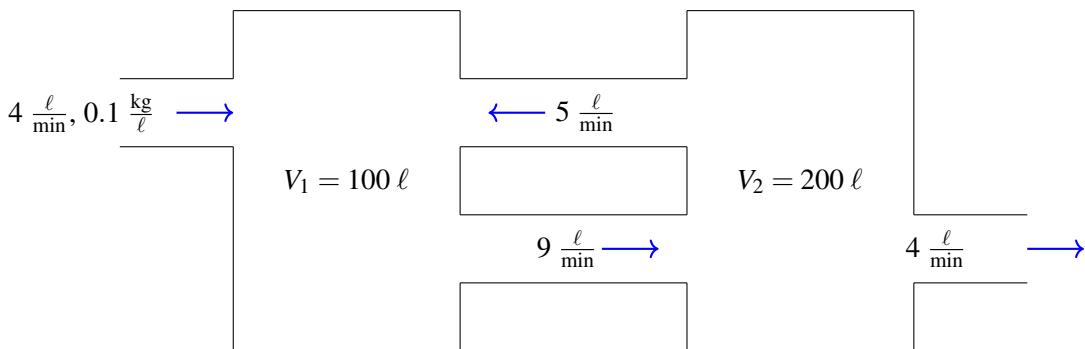


Figure 6.8: A two-tank salt tank problem for Exercise 6.3.3.

Exercise 6.3.4 Consider the system of Example 6.14.

- Find a particular solution $\mathbf{w}_p(t)$ of the form (6.41) in the case that $\omega = 2$ (so the system is driven at $\omega = 2$ radians per second) by solving (6.45) for \mathbf{a} and the using (6.44) to find \mathbf{b} .
- What difficulty arises when you use this procedure to find a particular solution of the form (6.41) when $\omega = \sqrt{2}$?
- There is one other (positive) value for ω in which a particular solution of the form (6.41) fails to exist. Find that value. Hint: look back at the homogeneous system solution in Example 6.9, and in particular at the two natural frequencies at which this system vibrates.

Exercise 6.3.5 Consider the double loop RC/RL circuit in Figure 6.9. Despite the fact that this problem involves a circuit, the analysis of this system is very similar to that of the mechanical system of Example 6.14, so it may be useful to refer back to that example. Here $q(t)$ denotes the charge on the capacitor.

- (a) Apply Kirchhoff's voltage law to the loop containing $V(t)$ and C and conclude that

$$V(t) - R_1 I_1 - q/C = 0.$$

- (b) Apply Kirchhoff's voltage law to the loop containing C, L , and R_2 and conclude that

$$-L\dot{I}_2 - R_2 I_2 + q/C = 0.$$

- (c) Apply Kirchhoff's current law to the node N and conclude that

$$I = I_1 - I_2.$$

- (d) Use (a)-(c) along with $\dot{q} = I$ to show that $q(t)$ and $I_2(t)$ satisfy the coupled ODEs

$$\begin{aligned}\dot{q} &= \frac{V(t)}{R_1} - \frac{1}{R_1 C} q - I_2 \\ I_2 &= \frac{1}{LC} q - \frac{R_2}{L} I_2.\end{aligned}\tag{6.46}$$

- (e) With $\mathbf{x} = \langle q, I_2 \rangle$, formulate (6.46) in the form $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f}(t)$; show \mathbf{A} and $\mathbf{f}(t)$ explicitly.
(f) Suppose $V(t) = \sin(1000t)$, $R_1 = 1$ ohm, $R_2 = 10$ ohms, $C = 10^{-4}$ farad, and $L = 10^{-3}$ henries. Solve the system using either Laplace transforms or undetermined coefficients, with initial data $q(0) = 0$ and $I_2(0) = 0$. Plot $q(t)$ and $i_2(t)$ on the interval $0 \leq t \leq 0.05$ seconds.

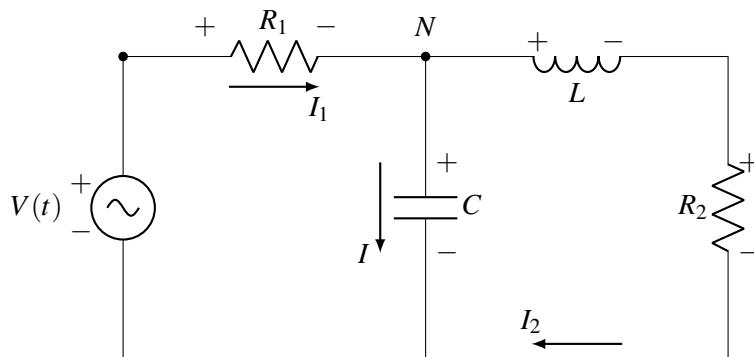


Figure 6.9: Double loop RC-RL circuit for Exercise 6.3.5.

6.4 The Matrix Exponential

The matrix exponential is a useful tool for analyzing systems of linear, constant-coefficient, ODEs. In this section we offer a quick introduction to the technique, with examples and exercises. We also include an introduction to Putzer's algorithm for computing the matrix exponential. This simple algorithm allows one to compute the exponential of any matrix whose eigenvalues are known. A computer algebra system can be especially helpful for this material, since the computations can be

a bit cumbersome.

6.4.1 Inspiration

The scalar ODE $\dot{x}(t) = ax(t)$ with initial data $x(0) = x_0$ has solution $x(t) = x_0 e^{at}$, easily obtained via separation of variables or an integrating factor. We can also solve it using Laplace transforms. Laplace transforming this ODE and solving for $X(s)$ shows that $X(s) = x_0/(s - a)$, which corresponds to $x(t) = x_0 e^{at}$ in the time domain.

A constant-coefficient system of linear first-order equations in matrix-vector notation can be expressed as $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$, with an initial condition $\mathbf{x}(0) = \mathbf{x}_0$. The notation is quite similar to the scalar case. Is it possible that we might make sense of solutions in the form $\mathbf{x}(t) = \mathbf{x}_0 e^{\mathbf{A}t}$, paralleling the scalar case? This idea was raised in Remark 6.3.1, in which we saw that the Laplace transform of the vector $\mathbf{X}(s) = \mathcal{L}(\mathbf{x}(t))$ is $\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}_0$, also very similar to the scalar equation $X(s) = x_0/(s - a)$.

In brief, this approach does work and allows us to elegantly express solutions to linear systems of constant-coefficient ODEs. First, note that we write $t\mathbf{A}$, since it is conventional to write a scalar-matrix product with the scalar on the left. Then, in a nutshell, we will see that:

- The matrix exponential $e^{t\mathbf{A}}$ can be defined as an $n \times n$ matrix with entries that are functions of t .
- The n -dimensional vector-valued function $\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}_0$ satisfies $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ with initial condition $\mathbf{x}(0) = \mathbf{x}_0$. The vector \mathbf{x}_0 appears on the right in $e^{t\mathbf{A}}\mathbf{x}_0$ since this is the product of the $n \times n$ matrix $e^{t\mathbf{A}}$ with the n -dimensional vector \mathbf{x}_0 .

6.4.2 Definition of the Matrix Exponential

Let's begin by defining $e^{\mathbf{B}}$ where \mathbf{B} is an $n \times n$ matrix of scalars, real or complex. We'll draw inspiration from the Taylor series for e^t , which is

$$e^t = 1 + t + \frac{t^2}{2} + \frac{t^3}{6} + \dots = \sum_{k=0}^{\infty} \frac{t^k}{k!}. \quad (6.47)$$

The series (6.47) converges for any real or complex number t .

We define $e^{\mathbf{B}}$ for an $n \times n$ matrix similarly as

$$e^{\mathbf{B}} = \mathbf{I} + \mathbf{B} + \frac{\mathbf{B}^2}{2} + \frac{\mathbf{B}^3}{6} + \dots = \sum_{k=0}^{\infty} \frac{\mathbf{B}^k}{k!}. \quad (6.48)$$

Here \mathbf{I} is the $n \times n$ identity matrix. A few remarks are in order.

1. If \mathbf{B} is an $n \times n$ matrix, then so are $\mathbf{B}^2, \mathbf{B}^3$, etc. More generally, \mathbf{B}^k is an $n \times n$ matrix for any exponent k .
2. As a consequence of point (1) above and the fact that \mathbf{I} is $n \times n$, each summand in (6.48) is an $n \times n$ matrix, so the sum of any finite number of terms in (6.48)

$$\mathbf{S}_m = \sum_{k=0}^m \frac{\mathbf{B}^k}{k!} \quad (6.49)$$

is an $n \times n$ matrix.

3. It's not obvious that the sum on the right in (6.48) (the limit of (6.49) as $m \rightarrow \infty$) converges, but it does.

Recall from calculus that for any real number t the infinite sum on the right in the scalar Taylor series (6.47) converges to some real number, and this number is e^t . In fact, the sum on the right in (6.47) can be taken as the very definition of e^t . Perhaps we can similarly show that the sum on the right in (6.48) converges. This would provide a method for actually defining $e^{\mathbf{B}}$, as that matrix to which the series converges.

An example is illuminating.

■ **Example 6.15** Let

$$\mathbf{B} = \begin{bmatrix} 3 & -2 \\ 4 & -3 \end{bmatrix}.$$

If we take $m = 0$ in (6.49) we just get $\mathbf{S}_0 = \mathbf{I}$, the 2×2 identity matrix, which doesn't even depend on \mathbf{B} . Using $m = 1$ in (6.49) produces

$$\mathbf{S}_1 = \sum_{k=0}^1 \frac{\mathbf{B}^k}{k!} = \mathbf{I} + \mathbf{B} = \begin{bmatrix} 4 & -2 \\ 4 & -2 \end{bmatrix}.$$

Taking $m = 2$ in (6.49) produces

$$\mathbf{S}_2 = \sum_{k=0}^2 \frac{\mathbf{B}^k}{k!} = \mathbf{I} + \mathbf{B} + \frac{\mathbf{B}^2}{2} = \begin{bmatrix} 9/2 & -2 \\ 4 & -3/2 \end{bmatrix}.$$

When $m = 5$ we find

$$\begin{aligned} \mathbf{S}_5 &= \sum_{k=0}^5 \frac{\mathbf{B}^k}{k!} \\ &= \mathbf{I} + \mathbf{B} + \frac{\mathbf{B}^2}{2} + \cdots + \frac{\mathbf{B}^5}{120} \\ &= \begin{bmatrix} 76/15 & -2 \\ 4 & -3/2 \end{bmatrix} \\ &\approx \begin{bmatrix} 5.067 & -2.350 \\ 4.700 & -1.983 \end{bmatrix}. \end{aligned}$$

When $m = 10$ we find that to three significant figures

$$\mathbf{S}_{10} = \sum_{k=0}^{10} \frac{\mathbf{B}^k}{k!} \approx \begin{bmatrix} 5.069 & -2.350 \\ 4.700 & -1.983 \end{bmatrix}.$$

It seems that as m increases, the sum converges to something. As in the scalar case (6.47), this convergence is facilitated by the rapid growth of $k!$ in the denominator of each summand. ■

Reading Exercise 6.4.1 Let

$$\mathbf{B} = \begin{bmatrix} 5 & -2 \\ 6 & -2 \end{bmatrix}.$$

Compute the sum \mathbf{S}_m in (6.49) for $m = 1, 2, 5$ and 10 ; a computer algebra system is helpful, and you may find it helpful to evaluate the sum numerically. How do \mathbf{S}_5 and \mathbf{S}_{10} compare?

If we use $(\mathbf{S}_m)_{jk}$ to denote the row j , column k entry in the matrix \mathbf{S}_m defined by (6.49), then it is a fact that these entries converge as $m \rightarrow \infty$. Specifically, for each j and k with $1 \leq j, k \leq n$,

$$\lim_{m \rightarrow \infty} (\mathbf{S}_m)_{jk} = E_{jk} \tag{6.50}$$

for some limit E_{jk} . For a proof see [30]. This behavior was suggested in Example 6.15 and Reading Exercise 6.4.1, and allows us to define the matrix exponential.

Definition 6.4.1 For a matrix \mathbf{B} we define $e^{\mathbf{B}}$ as that $n \times n$ matrix with row j , column k components E_{jk} as in (6.50).

6.4.3 Properties of the Matrix Exponential

Here are a few properties of the matrix exponential. They are identical to or closely related to those for exponentials of real or complex numbers.

1. $e^{\mathbf{0}} = \mathbf{I}$, where $\mathbf{0}$ denotes the $n \times n$ square matrix with all components equal to 0 and \mathbf{I} is the $n \times n$ identity matrix.
2. If $\mathbf{BC} = \mathbf{CB}$ (that is, \mathbf{B} and \mathbf{C} commute) then $e^{\mathbf{B}+\mathbf{C}} = e^{\mathbf{B}}e^{\mathbf{C}}$. This is usually false if $\mathbf{BC} \neq \mathbf{CB}$.
3. For any square matrix \mathbf{B} we have $e^{\mathbf{B}}e^{-\mathbf{B}} = \mathbf{I}$. In particular, $e^{\mathbf{B}}$ is always invertible with inverse $e^{-\mathbf{B}}$.
4. For any matrix \mathbf{B} we have $\mathbf{B}e^{\mathbf{B}} = e^{\mathbf{B}}\mathbf{B}$.

Proving Property (1) is an easy exercise using (6.48). For proofs of Properties (2) and (4) see [30].

Reading Exercise 6.4.2 Demonstrate that Property (3) above is true. Hint: use (1) and (2) with $\mathbf{C} = -\mathbf{B}$, noting that \mathbf{B} and $-\mathbf{B}$ commute under matrix multiplication.

6.4.4 Solving ODEs with the Matrix Exponential

The Matrix $e^{t\mathbf{A}}$

If \mathbf{A} is an $n \times n$ matrix then $e^{t\mathbf{A}}$ is defined by taking $\mathbf{B} = t\mathbf{A}$ in (6.48). When written out explicitly we find that

$$e^{t\mathbf{A}} = \mathbf{I} + t\mathbf{A} + t^2 \frac{\mathbf{A}^2}{2} + t^3 \frac{\mathbf{A}^3}{6} + \cdots = \sum_{k=0}^{\infty} t^k \frac{\mathbf{A}^k}{k!}. \quad (6.51)$$

Each entry of the $n \times n$ matrix $e^{t\mathbf{A}}$ is a function of t .

■ **Example 6.16** If

$$\mathbf{A} = \begin{bmatrix} -4 & 2 \\ -6 & 3 \end{bmatrix}$$

then it can be shown (as you will below in Reading Exercise 6.4.5) that $e^{t\mathbf{A}}$ is

$$e^{t\mathbf{A}} = \begin{bmatrix} -3 + 4e^{-t} & 2 - 2e^{-t} \\ 6e^{-t} - 6 & 4 - 3e^{-t} \end{bmatrix}.$$

Computing $e^{t\mathbf{A}}$ by using the Taylor series (6.51) isn't very efficient. We'll see how to perform this computation more insightfully a bit later in this section. ■

In order to use $e^{t\mathbf{A}}$ to solve differential equations, we need to compute $d(e^{t\mathbf{A}})/dt$, in which we differentiate each of the n^2 components of $e^{t\mathbf{A}}$ with respect to t . This can be done using the series expansion (6.51) and term-by-term differentiation (if permitted), and a bit of algebraic manipulation.

We see that

$$\begin{aligned}
 \frac{d(e^{t\mathbf{A}})}{dt} &= \sum_{k=0}^{\infty} \frac{d(t^k)}{dt} \frac{\mathbf{A}^k}{k!} \\
 &= \sum_{k=1}^{\infty} k t^{k-1} \frac{\mathbf{A}^k}{k!} && (\text{the } k=0 \text{ term drops out}) \\
 &= \sum_{k=1}^{\infty} t^{k-1} \frac{\mathbf{A}^k}{(k-1)!} && (k/k! = 1/(k-1)!) \\
 &= \sum_{j=0}^{\infty} t^j \frac{\mathbf{A}^{j+1}}{j!} && (\text{let } j=k-1 \text{ above}) \\
 &= \mathbf{A} \sum_{j=0}^{\infty} t^j \frac{\mathbf{A}^j}{j!} && (\text{factor an } \mathbf{A} \text{ out of the sum}) \\
 &= \mathbf{A} e^{t\mathbf{A}}. && (6.52)
 \end{aligned}$$

We could also have factored the common \mathbf{A} term in the infinite sum out to the right side of the sum, in which case we find that $d(e^{t\mathbf{A}})/dt = e^{t\mathbf{A}}\mathbf{A}$ is also correct. The validity of the computations that lead to (6.52) (in particular, that the series for $e^{t\mathbf{A}}$ can be differentiated term by term) can also be found in [30].

In summary, the derivative of $e^{t\mathbf{A}}$ with respect to t is the $n \times n$ matrix

$$\frac{d}{dt}(e^{t\mathbf{A}}) = \mathbf{A} e^{t\mathbf{A}} = e^{t\mathbf{A}}\mathbf{A}. \quad (6.53)$$

Both $\mathbf{A}e^{t\mathbf{A}}$ and $e^{t\mathbf{A}}\mathbf{A}$ in (6.53) are the product of the $n \times n$ matrix \mathbf{A} with the $n \times n$ matrix $e^{t\mathbf{A}}$, and so both are $n \times n$ matrices, as they should be. Also note the similarity of (6.53) to the scalar computation $d(e^{at})/dt = ae^{at} = e^{at}a$ that follows from the chain rule.

Reading Exercise 6.4.3

- (a) For the matrix \mathbf{A} in Example 6.16, compute $\frac{d}{dt}(e^{t\mathbf{A}})$ directly, by differentiating each component with respect to t .
- (b) For the matrix \mathbf{A} in Example 6.16, compute $\frac{d}{dt}(e^{t\mathbf{A}})$ by using (6.53), then compute each expression $\mathbf{A}e^{t\mathbf{A}}$ and $e^{t\mathbf{A}}\mathbf{A}$. Verify that you obtain the same answer as in part (a).

The Fundamental Matrix Solution

The matrix exponential allows us to solve $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ with $\mathbf{x}(0) = \mathbf{x}_0$ for any square matrix \mathbf{A} and initial data \mathbf{x}_0 . To see how, define $\mathbf{X}(t) = e^{t\mathbf{A}}$, so $\mathbf{X}(t)$ is a matrix-valued function of t : for each input t , $\mathbf{X}(t)$ outputs a matrix. The function $\mathbf{X}(t)$ is called a **fundamental matrix solution** to $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$; according to (6.53), $\mathbf{X}(t)$ satisfies

$$\dot{\mathbf{X}} = \mathbf{A}\mathbf{X}.$$

If we consider the $n \times n$ matrix-valued function $\mathbf{X}(t)$ to consist of n vector-valued functions $\mathbf{x}_1(t), \dots, \mathbf{x}_n(t)$ (column vectors) in the form

$$\mathbf{X}(t) = [\mathbf{x}_1(t) \quad \mathbf{x}_2(t) \quad \cdots \quad \mathbf{x}_n(t)].$$

then from the definition of matrix multiplication (specifically, that \mathbf{AX} is computed by multiplying \mathbf{X} , column by column, by \mathbf{A}), we have

$$[\dot{\mathbf{x}}_1(t) \quad \dot{\mathbf{x}}_2(t) \quad \cdots \quad \dot{\mathbf{x}}_n(t)] = [\mathbf{A}\mathbf{x}_1(t) \quad \mathbf{A}\mathbf{x}_2(t) \quad \cdots \quad \mathbf{A}\mathbf{x}_n(t)].$$

We conclude that each individual column $\mathbf{x}_k(t)$ of $\mathbf{X}(t)$ is a solution to $\dot{\mathbf{x}} = \mathbf{Ax}$. Moreover, any linear combination of the columns $\mathbf{x}_k(t)$ satisfies this ODE, since

$$\begin{aligned} \frac{d}{dt} \left(\sum_{k=1}^n c_k \mathbf{x}_k(t) \right) &= \sum_{k=1}^n c_k \dot{\mathbf{x}}_k \\ &= \sum_{k=1}^n c_k \mathbf{Ax}_k \\ &= \mathbf{A} \left(\sum_{k=1}^n c_k \mathbf{x}_k(t) \right). \end{aligned} \quad (6.54)$$

We will show how to use $\mathbf{X}(t)$ in conjunction with the initial data vector \mathbf{x}_0 to produce a solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ with $\mathbf{x}(0) = \mathbf{x}_0$.

Obtaining the Initial Data

Notice that $\mathbf{X}(0) = e^0 = \mathbf{I}$, or equivalently, that $\mathbf{x}_k(0) = \mathbf{e}_k$ where \mathbf{e}_k is the k th standard basis vector in \mathbb{R}^n . To construct a solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ with initial data $\mathbf{x}_0 = \langle c_1, c_2, \dots, c_n \rangle$, form the vector-valued function

$$\mathbf{x}(t) = c_1 \mathbf{x}_1(t) + \cdots + c_n \mathbf{x}_n(t). \quad (6.55)$$

From (6.54) we see that (6.55) $\dot{\mathbf{x}} = \mathbf{Ax}$ and that $\mathbf{x}(0) = c_1 \mathbf{e}_1 + \cdots + c_n \mathbf{e}_n = \mathbf{x}_0$; therefore $\mathbf{x}(t)$ provides the solution we sought. From the definition of matrix-vector multiplication, the quantity on the right in (6.55) can be written more succinctly as $\mathbf{X}\mathbf{x}_0$ or $e^{t\mathbf{A}}\mathbf{x}_0$.

We have demonstrated the following theorem:

Theorem 6.4.1 The solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ with initial data $\mathbf{x}(0) = \mathbf{x}_0$ is

$$\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}_0.$$

■ **Example 6.17** In Example 6.16 we presented the matrix

$$\mathbf{A} = \begin{bmatrix} -4 & 2 \\ -6 & 3 \end{bmatrix}$$

and remarked that it can be shown that $e^{t\mathbf{A}}$ is

$$e^{t\mathbf{A}} = \begin{bmatrix} -3 + 4e^{-t} & 2 - 2e^{-t} \\ 6e^{-t} - 6 & 4 - 3e^{-t} \end{bmatrix}.$$

We can use $e^{t\mathbf{A}}$ to solve the ODE system $\dot{\mathbf{x}} = \mathbf{Ax}$ with any initial data, say $\mathbf{x}(0) = \langle 1, 3 \rangle$. According to Theorem 6.4.1, the solution is

$$\mathbf{x}(t) = e^{t\mathbf{A}} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} -3 + 4e^{-t} & 2 - 2e^{-t} \\ 6e^{-t} - 6 & 4 - 3e^{-t} \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 - 2e^{-t} \\ 6 - 3e^{-t} \end{bmatrix}.$$

■

■ **Example 6.18** In Section 6.2 we considered a critically damped spring-mass system governed by $u''(t) + 4u'(t) + 4u(t) = 0$. With $x_1 = u$ and $x_2 = u'$ this is equivalent to the system $\dot{\mathbf{x}} = \mathbf{Ax}$ with

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -4 & -4 \end{bmatrix}.$$

The matrix \mathbf{A} is defective, with double eigenvalue $\lambda = -2$ and eigenvector $\mathbf{v} = \langle -1, 2 \rangle$. As a result, we needed special procedures to solve the system.

But with the matrix exponential, the solution takes precisely the form of Theorem 6.4.1. We can compute

$$e^{t\mathbf{A}} = \begin{bmatrix} (2t+1)e^{2t} & te^{-2t} \\ -4te^{-2t} & (-2t+1)e^{-2t} \end{bmatrix}.$$

We will discuss how $e^{t\mathbf{A}}$ above was computed when we consider Putzer's algorithm in Section 6.4.7. But with $e^{t\mathbf{A}}$ in hand, the solution to $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ with, for example, $\mathbf{x}(0) = \langle 1, -5 \rangle$ is

$$\mathbf{x}(t) = e^{t\mathbf{A}} \begin{bmatrix} 1 \\ -5 \end{bmatrix} = \begin{bmatrix} (2t+1)e^{2t} & te^{-2t} \\ -4te^{-2t} & (-2t+1)e^{-2t} \end{bmatrix} \begin{bmatrix} 1 \\ -5 \end{bmatrix} = \begin{bmatrix} (-3t+1)e^{-2t} \\ (6t-5)e^{-2t} \end{bmatrix}. \quad \blacksquare$$

6.4.5 Computing The Matrix Exponential: The Diagonal Case

There is one case in which computing the matrix exponential $e^{\mathbf{B}}$ using (6.48) is particularly easy: when the matrix is diagonal. Suppose that \mathbf{D} is an $n \times n$ diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \dots, \lambda_n$, so

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & \lambda_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix}. \quad (6.56)$$

You can easily check that any power \mathbf{D}^k can be computed as

$$\mathbf{D}^k = \begin{bmatrix} \lambda_1^k & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2^k & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & \lambda_{n-1}^k & 0 \\ 0 & 0 & \cdots & 0 & \lambda_n^k \end{bmatrix}.$$

When we substitute the diagonal matrix \mathbf{D} into the Taylor series (6.48) and sum the matrices component by component we obtain

$$e^{\mathbf{D}} = \sum_{k=0}^{\infty} \frac{\mathbf{D}^k}{k!} = \begin{bmatrix} \sum_{k=0}^{\infty} \frac{\lambda_1^k}{k!} & 0 & \cdots & 0 & 0 \\ 0 & \sum_{k=0}^{\infty} \frac{\lambda_2^k}{k!} & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & \sum_{k=0}^{\infty} \frac{\lambda_{n-1}^k}{k!} & 0 \\ 0 & 0 & \cdots & 0 & \sum_{k=0}^{\infty} \frac{\lambda_n^k}{k!} \end{bmatrix}.$$

But the sums on the diagonals are just the Taylor series for e^{λ_m} , with $1 \leq m \leq n$. Therefore we have that

$$e^{\mathbf{D}} = \begin{bmatrix} e^{\lambda_1} & 0 & \cdots & 0 & 0 \\ 0 & e^{\lambda_2} & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & e^{\lambda_{n-1}} & 0 \\ 0 & 0 & \cdots & 0 & e^{\lambda_n} \end{bmatrix}. \quad (6.57)$$

To compute $e^{t\mathbf{D}}$ note that $t\mathbf{D}$ is a diagonal matrix with m th diagonal entry $\lambda_m t$, so from (6.57), $e^{t\mathbf{D}}$ is the diagonal matrix with m th diagonal entry $e^{\lambda_m t}$.

Reading Exercise 6.4.4 Formulate the system

$$\begin{aligned}x'_1(t) &= 3x_1(t) \\x'_2(t) &= x_2(t)\end{aligned}$$

as $\mathbf{x}'(t) = \mathbf{D}\mathbf{x}(t)$. Write out \mathbf{D} explicitly. Use the matrix exponential (6.57) with λ_m replaced by $\lambda_m t$ to solve the system with initial conditions $x_1(0) = 2$ and $x_2(0) = 5$.

6.4.6 Computing The Matrix Exponential: The Diagonalizable Case

Let's look at an efficient method for computing $e^{\mathbf{B}}$ in a very common case; we will then apply the method with $\mathbf{B} = t\mathbf{A}$ to solve systems of ODEs. We will assume that the matrix \mathbf{B} is diagonalizable as detailed in Section B.5 of Appendix B. Specifically, suppose \mathbf{B} has eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$, allowing repeated or complex eigenvalues. We suppose it is possible to choose a set $S = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ of eigenvectors (where \mathbf{v}_k corresponds to eigenvalue λ_k) so that the matrix

$$\mathbf{P} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_n]$$

with k th column \mathbf{v}_k is invertible. Equivalently, the set S is linearly independent. Let \mathbf{D} be the $n \times n$ matrix with k th diagonal element λ_k , in the form of (6.56). Under these assumptions the matrix \mathbf{B} can be written in the form

$$\mathbf{B} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}, \quad (6.58)$$

or **diagonalized**, as described in Appendix B. Moreover, as shown in Appendix B,

$$\mathbf{B}^k = \mathbf{P}\mathbf{D}^k\mathbf{P}^{-1}. \quad (6.59)$$

By using (6.59) we find that

$$\begin{aligned}\sum_{k=0}^m \mathbf{B}^k &= \sum_{k=0}^m \mathbf{P}\mathbf{D}^k\mathbf{P}^{-1} \\&= \mathbf{P} \left(\sum_{k=0}^m \mathbf{D}^k \right) \mathbf{P}^{-1} \\&= \mathbf{P} \begin{bmatrix} \sum_{k=0}^m \frac{\lambda_1^k}{k!} & 0 & \cdots & 0 & 0 \\ 0 & \sum_{k=0}^m \frac{\lambda_2^k}{k!} & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & \sum_{k=0}^m \frac{\lambda_{n-1}^k}{k!} & 0 \\ 0 & 0 & \cdots & 0 & \sum_{k=0}^m \frac{\lambda_n^k}{k!} \end{bmatrix} \mathbf{P}^{-1}.\end{aligned}$$

As m approaches infinity the i th diagonal sum approaches e^{λ_i} . We conclude that

$$e^{\mathbf{B}} = \mathbf{P}e^{\mathbf{D}}\mathbf{P}^{-1}, \quad (6.60)$$

where $e^{\mathbf{D}}$ is the diagonal matrix with i th diagonal entry e^{λ_i} .

In the case in which we want to compute $e^{t\mathbf{A}}$, substitute $\mathbf{B} = t\mathbf{A}$ in (6.60) and note that $t\mathbf{A}$ is just t times \mathbf{A} and so has the same eigenvectors as \mathbf{A} , but with eigenvalues $\lambda_k t$. We can conclude that

$$e^{t\mathbf{A}} = \mathbf{P}e^{t\mathbf{D}}\mathbf{P}^{-1}. \quad (6.61)$$

■ **Example 6.19** Let's use the matrix exponential to solve the linear system

$$\begin{aligned}x'_1(t) &= 7x_1(t) - 6x_2(t) \\x'_2(t) &= 12x_1(t) - 10x_2(t)\end{aligned}$$

with initial conditions $x_1(0) = 1$ and $x_2(0) = 2$. In matrix terms we have $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ with

$$\mathbf{A} = \begin{bmatrix} 7 & -6 \\ 12 & -10 \end{bmatrix}.$$

The eigenvalues for \mathbf{A} are $\lambda_1 = -2$ and $\lambda_2 = -1$ with corresponding eigenvectors $\mathbf{v}_1 = \langle 2, 3 \rangle$ and $\mathbf{v}_2 = \langle 3, 4 \rangle$. Thus in (6.58) we take

$$\mathbf{D} = \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix} \text{ and } \mathbf{P} = \begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix}.$$

A straightforward matrix multiplication shows that

$$\begin{aligned}e^{t\mathbf{A}} &= \mathbf{P}e^{t\mathbf{D}}\mathbf{P}^{-1} \\&= \begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} e^{-2t} & 0 \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} -4 & 3 \\ 3 & -2 \end{bmatrix} \\&= \begin{bmatrix} 9e^{-t} - 8e^{-2t} & -6e^{-t} + 6e^{-2t} \\ 12e^{-t} - 12e^{-2t} & -8e^{-t} + 9e^{-2t} \end{bmatrix}.\end{aligned}$$

The solution to the system with the given initial conditions is then

$$\mathbf{x}(t) = e^{t\mathbf{A}} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 9e^{-t} - 8e^{-2t} & -6e^{-t} + 6e^{-2t} \\ 12e^{-t} - 12e^{-2t} & -8e^{-t} + 9e^{-2t} \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -3e^{-t} + 4e^{-2t} \\ -4e^{-t} + 6e^{-2t} \end{bmatrix}.$$

■

Reading Exercise 6.4.5 Verify that $e^{t\mathbf{A}}$ in Example 6.16 is correct by diagonalizing \mathbf{A} and using (6.61).

Reading Exercise 6.4.6 What goes wrong with (6.61) when you try to compute $e^{t\mathbf{A}}$ for the matrix of Example 6.18?

6.4.7 Computing The Matrix Exponential: Putzer's Algorithm

In Reading Exercise 6.4.6 you identified what can go wrong with using diagonalization to compute the matrix exponential: not all matrices are diagonalizable. Nonetheless, the series (6.48) converges for any matrix, and so all matrices can be exponentiated. How do we compute the matrix exponential when a matrix cannot be diagonalized? Putzer's algorithm provides a procedure for exponentiating any matrix, diagonalizable or not. It's really geared toward computing $e^{t\mathbf{A}}$ (t already included), so we'll examine it in that form. We present the algorithm and examples below. The reader interested in a rigorous proof that Putzer's algorithm actually produces $e^{t\mathbf{A}}$ can consult [99].

We begin by supposing that the $n \times n$ matrix \mathbf{A} has eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$; these eigenvalues need not be distinct, but they all must be computed or somehow known. Putzer's algorithm expresses $e^{t\mathbf{A}}$ in the form

$$e^{t\mathbf{A}} = \sum_{j=0}^{n-1} r_{j+1}(t) \mathbf{P}_j, \quad (6.62)$$

where the $r_{j+1}(t)$ are polynomials in t and the \mathbf{P}_j are certain matrices computed as follows. First, set $\mathbf{P}_0 = \mathbf{I}$ and \mathbf{P}_j as the product

$$\mathbf{P}_j = \prod_{k=1}^j (\mathbf{A} - \lambda_k \mathbf{I})$$

for $1 \leq j \leq n-1$. The $r_j(t)$ are the scalar components of the vector $\mathbf{r}(t) = \langle r_1(t), \dots, r_n(t) \rangle$ that satisfies the system of ODEs

$$\dot{\mathbf{r}}(t) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 & 0 \\ 1 & \lambda_2 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \lambda_{n-1} & 0 \\ 0 & 0 & \cdots & 1 & \lambda_n \end{bmatrix} \mathbf{r}(t) \quad (6.63)$$

with initial condition $\mathbf{r}(t) = \langle 1, 0, 0, \dots, 0 \rangle$.

Examples

■ **Example 6.20** Let \mathbf{A} be the matrix in Example 6.19. We'll compute $e^{t\mathbf{A}}$ using Putzer's algorithm. Here $n = 2$ and the eigenvalues of this matrix are $\lambda_1 = -1$ and $\lambda_2 = -2$. We find that

$$\begin{aligned} \mathbf{P}_0 &= \mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ \mathbf{P}_1 &= \prod_{k=1}^1 (\mathbf{A} - \lambda_k \mathbf{I}) = (\mathbf{A} + \mathbf{I}) = \begin{bmatrix} 8 & -6 \\ 12 & -9 \end{bmatrix}. \end{aligned}$$

Equation (6.63) becomes

$$\begin{bmatrix} \dot{r}_1(t) \\ \dot{r}_2(t) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} r_1(t) \\ r_2(t) \end{bmatrix},$$

with $r_1(0) = 1$ and $r_2(0) = 0$. The first equation, $\dot{r}_1(t) = -r_1(t)$ with $r_1(0) = 1$, is decoupled from the second and has solution $r_1(t) = e^{-t}$. The second equation for $r_2(t)$ then becomes $\dot{r}_2(t) = e^{-t} - 2r_2(t)$ with $r_2(0) = 0$. The ODE for $r_2(t)$ is a scalar constant-coefficient linear ODE, easy to solve with the integrating factor approach. We find $r_2(t) = e^{-t} - e^{-2t}$. From (6.62) we have

$$\begin{aligned} e^{t\mathbf{A}} &= \sum_{j=0}^1 r_{j+1}(t) \mathbf{P}_j \\ &= r_1(t) \mathbf{P}_0 + r_2(t) \mathbf{P}_1 \\ &= e^{-t} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + (e^{-t} - e^{-2t}) \begin{bmatrix} 8 & -6 \\ 12 & -9 \end{bmatrix} \\ &= \begin{bmatrix} 9e^{-t} - 8e^{-2t} & -6e^{-t} + 6e^{-2t} \\ 12e^{-t} - 12e^{-2t} & -8e^{-t} + 9e^{-2t} \end{bmatrix}, \end{aligned}$$

just as in Example 6.19. ■

■ **Example 6.21** Let's compute $e^{t\mathbf{A}}$ using \mathbf{A} from Example 6.18. Recall that this matrix is not diagonalizable, but Putzer's algorithm will work here. The matrix is

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -4 & -4 \end{bmatrix}$$

with eigenvalues $\lambda_1 = \lambda_2 = -2$ (recall that \mathbf{A} is defective, since there is only a single eigenvector). Then we find

$$\mathbf{P}_0 = \mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mathbf{P}_1 = \prod_{k=1}^1 (\mathbf{A} - \lambda_k \mathbf{I}) = (\mathbf{A} + 2\mathbf{I}) = \begin{bmatrix} 2 & 1 \\ -4 & -2 \end{bmatrix}.$$

Equation (6.63) becomes

$$\begin{bmatrix} \dot{r}_1(t) \\ \dot{r}_2(t) \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} r_1(t) \\ r_2(t) \end{bmatrix}$$

with $r_1(0) = 1$ and $r_2(0) = 0$. The first equation, $\dot{r}_1(t) = -2r_1(t)$ with $r_1(0) = 1$ is decoupled from the second, with solution $r_1(t) = e^{-2t}$. The second equation for $r_2(t)$ then becomes $\dot{r}_2(t) = e^{-2t} - 2r_2(t)$ with $r_2(0) = 0$. The ODE for $r_2(t)$ is a scalar constant-coefficient linear ODE, and is easy to solve with the integrating factor approach. We find $r_2(t) = te^{-2t}$. From (6.62) we have

$$\begin{aligned} e^{t\mathbf{A}} &= \sum_{j=0}^1 r_{j+1}(t) \mathbf{P}_j \\ &= r_1(t) \mathbf{P}_0 + r_2(t) \mathbf{P}_1 \\ &= e^{-2t} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + te^{-2t} \begin{bmatrix} 2 & 1 \\ -4 & -2 \end{bmatrix} \\ &= \begin{bmatrix} (2t+1)e^{-2t} & te^{-2t} \\ -4te^{-2t} & (-2t+1)e^{-2t} \end{bmatrix}. \end{aligned}$$

■

■ **Example 6.22** Let

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 1 \\ -3 & -1 & 2 \\ 0 & 0 & 2 \end{bmatrix}.$$

This matrix is not diagonalizable. The eigenvalues are $\lambda_1 = -1$, $\lambda_2 = 2$, and $\lambda_3 = 2$, although the order doesn't matter. There is an eigenvector $\langle 0, 1, 0 \rangle$ for λ_1 , but because the double eigenvalue $\lambda_2 = \lambda_3 = 2$ has only the single eigenvector $\langle -1, 1, 0 \rangle$ (or multiples thereof), the diagonalization approach to computing $e^{t\mathbf{A}}$ won't work. We will therefore use Putzer's algorithm. With $n = 3$ we find

$$\mathbf{P}_0 = \mathbf{I}$$

$$\mathbf{P}_1 = (\mathbf{A} + \mathbf{I}) = \begin{bmatrix} 3 & 0 & 1 \\ -3 & 0 & 2 \\ 0 & 0 & 3 \end{bmatrix}$$

$$\mathbf{P}_2 = (\mathbf{A} + \mathbf{I})(\mathbf{A} - 2\mathbf{I}) = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & -3 \\ 0 & 0 & 0 \end{bmatrix}.$$

Equation (6.63) becomes

$$\begin{bmatrix} \dot{r}_1(t) \\ \dot{r}_2(t) \\ \dot{r}_3(t) \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} r_1(t) \\ r_2(t) \\ r_3(t) \end{bmatrix}$$

with $r_1(0) = 1, r_2(0) = 0$ and $r_3(0) = 0$. As in the previous examples, the first equation, $\dot{r}_1(t) = -r_1(t)$ with $r_1(0) = 1$ is decoupled from the second, with solution $r_1(t) = e^{-t}$. The second equation for $r_2(t)$ then becomes $\dot{r}_2(t) = e^{-t} + 2r_2(t)$ with $r_2(0) = 0$. This is a scalar constant-coefficient linear ODE, and easy to solve with the integrating factor approach. We find $r_2(t) = e^{2t}/3 - e^{-t}/3$. With $r_2(t)$ in hand the third equation becomes $\dot{r}_3(t) = e^{2t}/3 - e^{-t}/3 + 2r_3(t)$ with $r_3(0) = 0$. Again, this is a scalar constant-coefficient linear ODE, easy to solve with the integrating factor approach. We find $r_3(t) = (3t - 1)e^{2t}/9 + e^{-t}/9$. From (6.62) we have

$$\begin{aligned} e^{t\mathbf{A}} &= \sum_{j=0}^2 r_{j+1}(t)\mathbf{P}_j \\ &= r_1(t)\mathbf{P}_0 + r_2(t)\mathbf{P}_1 + r_3(t)\mathbf{P}_2 \\ &= e^{-t} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \left(\frac{e^{2t} - e^{-t}}{3}\right) \begin{bmatrix} 3 & 0 & 1 \\ -3 & 0 & 2 \\ 0 & 0 & 3 \end{bmatrix} + \left(\frac{(3t - 1)e^{2t} + e^{-t}}{9}\right) \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & -3 \\ 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} e^{2t} & 0 & te^{2t} \\ e^{-t} - e^{2t} & e^{-t} & e^{2t} - e^{-t} + te^{-2t} \\ 0 & 0 & e^{2t} \end{bmatrix}. \end{aligned}$$

■

Notice how in each example we can find the $r_j(t)$ one at a time, first $r_1(t)$ in isolation from the other $r_j(t)$, then $r_2(t)$ (using knowledge of r_1), then $r_3(t)$ from knowledge of $r_1(t)$ and $r_2(t)$, and so on. In each case we can compute $r_j(t)$ as the solution to a scalar ODE by using an integrating factor approach.

6.4.8 Final Remarks

The solution $x(t) = x_0 e^{at}$ to the scalar ODE $x'(t) = ax(t)$ can be thought of as a prescription for how the initial data value x_0 is evolved or pushed forward in time. If a is positive then x_0 is multiplied by a factor e^{at} that grows in time, while if a is negative then x_0 diminishes toward zero. The matrix exponential embodies the same idea. The initial data vector \mathbf{x}_0 in \mathbb{R}^n is evolved in time under the action of the time-dependent matrix $e^{t\mathbf{A}}$. For any t we can think of $e^{t\mathbf{A}}$ as an operator that acts on n -dimensional vectors like \mathbf{x}_0 and maps them to new n -dimensional vectors (the solution $\mathbf{x}(t)$ at a particular time). This framework of an operator that evolves the solution to a differential equation forward in time is very useful in more sophisticated and general situations, for example, partial differential equations or evolution equations, and plays a large role in many more advanced areas of mathematics and physics.

6.4.9 Exercises

In each of Exercises 6.4.1 to 6.4.5 the intermediate results of Putzer's algorithm will depend on the order of the eigenvalues.

Exercise 6.4.1 Let

$$\mathbf{A} = \begin{bmatrix} 2 & -6 \\ 2 & -5 \end{bmatrix}.$$

Use both diagonalization and Putzer's algorithm to compute $e^{t\mathbf{A}}$. Use $e^{t\mathbf{A}}$ to solve $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ with $x_1(0) = 1$ and $x_2(0) = 2$.

Exercise 6.4.2 Let

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}.$$

Use both diagonalization and Putzer's algorithm to compute $e^{t\mathbf{A}}$. Use $e^{t\mathbf{A}}$ to solve $\dot{\mathbf{x}}(t) = \mathbf{Ax}(t)$ with $x_1(0) = 4$ and $x_2(0) = 2$.

Exercise 6.4.3 Let

$$\mathbf{A} = \begin{bmatrix} -7 & 3 \\ -18 & 8 \end{bmatrix}.$$

Use both diagonalization and Putzer's algorithm to compute $e^{t\mathbf{A}}$. Use $e^{t\mathbf{A}}$ to solve $\dot{\mathbf{x}}(t) = \mathbf{Ax}(t)$ with $x_1(0) = 0$ and $x_2(0) = -2$.

Exercise 6.4.4 Let

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -1 & 3 \end{bmatrix}.$$

Use Putzer's algorithm to compute $e^{t\mathbf{A}}$. Use $e^{t\mathbf{A}}$ to solve $\dot{\mathbf{x}}(t) = \mathbf{Ax}(t)$ with $x_1(0) = 1$ and $x_2(0) = 2$.

Exercise 6.4.5 Let

$$\mathbf{A} = \begin{bmatrix} -1 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & -2 & -1 \end{bmatrix}.$$

Use Putzer's algorithm to compute $e^{t\mathbf{A}}$. Use $e^{t\mathbf{A}}$ to solve $\dot{\mathbf{x}}(t) = \mathbf{Ax}(t)$ with $x_1(0) = 1$, $x_2(0) = 0$ and $x_3(0) = -1$.

Exercise 6.4.6 Formulate the system

$$\begin{aligned} x'_1(t) &= x_1(t) - x_2(t) - x_3(t) \\ x'_2(t) &= x_1(t) + 3x_2(t) + x_3(t) \\ x'_3(t) &= -3x_1(t) + x_2(t) - x_3(t) \end{aligned}$$

as $\dot{\mathbf{x}}(t) = \mathbf{Ax}(t)$ and use the matrix exponential $e^{t\mathbf{A}}$ to solve the system with $x_1(0) = 1$, $x_2(0) = 0$ and $x_3(0) = -1$. Hint: the eigenvalues and eigenvectors here are integers.

Exercise 6.4.7 You've seen how to solve a scalar nonhomogeneous linear differentiable equation of the form $\dot{x}(t) = ax(t) + b(t)$, where a is a constant and $b(t)$ a function. This is usually done using the integrating factor technique. Generalize the integrating factor technique to the case of n constant-coefficient linear nonhomogeneous ODEs, of the form

$$\dot{\mathbf{x}}(t) = \mathbf{Ax}(t) + \mathbf{b}(t), \tag{6.64}$$

where $\mathbf{b}(t) = \langle b_1(t), \dots, b_n(t) \rangle$. In particular, show that

$$\mathbf{x}(t) = e^{t\mathbf{A}} \int_0^t e^{-z\mathbf{A}} \mathbf{b}(z) dz + e^{t\mathbf{A}} \mathbf{x}_0 \quad (6.65)$$

provides a solution to (6.64) with $\mathbf{x}(0) = \mathbf{x}_0$. (Here z is a dummy variable of integration.) The integral in (6.65) is to be evaluated component by component.

Then use (6.65) to solve the system

$$\begin{aligned} \dot{x}_1(t) &= x_1(t) - 2x_2(t) + 1 \\ \dot{x}_2(t) &= 4x_1(t) - 5x_2(t) + t \end{aligned}$$

with $x_1(0) = 1$ and $x_2(0) = 2$.

6.5 Modeling Projects

6.5.1 Project: LSD Compartment Model

Begin by rereading the two-compartment model for LSD metabolism in Section 6.1. The data in Table 6.1 is also provided on the book web site [8].

The model (6.1)-(6.2) we presented in Section 6.1 was based on letting $u_P(t)$ and $u_T(t)$ denote the actual amount of drug present in the body's plasma or tissue compartments, respectively, but the authors of [88] and [113] formulate their model based on the concentration of LSD in the relevant compartment. If V_P denotes the volume of the plasma compartment and V_T the volume of the tissue compartment, then $u_P(t) = V_P c_P(t)$ and $u_T(t) = V_T c_T(t)$, where $c_P(t)$ and $c_T(t)$ are the concentrations in the plasma and tissue, respectively. Then (6.1)-(6.2) can be reformulated as

$$\begin{aligned} V_P \dot{c}_P(t) &= -k_b V_P c_P(t) - k_e V_P c_P(t) + k_a V_T c_T(t) + g(t) \\ V_T \dot{c}_T(t) &= k_b V_P c_P(t) - k_a V_T c_T(t). \end{aligned} \quad (6.66)$$

However, we will take $g(t) = 0$, assuming that after the initial dose, no additional LSD is administered.

In [88] and [113] the authors also estimate that these volumes can be approximated as $V_P = 0.163M$ and $V_T = 0.115M$ liters, where M is the mass of the subject in kilograms. Since plasma and tissue both have a density of about 1 kg per liter, we might also interpret these as equivalent masses. Then the system (6.66) becomes

$$\begin{aligned} 0.163M \dot{c}_P(t) &= -0.163M k_b c_P(t) - 0.163M k_e c_P(t) + 0.115M k_a c_T(t) \\ 0.115M \dot{c}_T(t) &= 0.163M k_b c_P(t) - 0.115M k_a c_T(t). \end{aligned} \quad (6.67)$$

Note that in (6.67) the mass M can be divided out. The initial dose administered to each subject was 2 mg per kg of body mass M for an initial dose of $2M$ mg. If we assume this was quickly distributed uniformly over the plasma volume, we have an initial concentration of $2/0.163 \approx 12.27$ mg per liter of plasma volume.

Modeling Exercise 5.1.1 Show that the system (6.67) can be reformulated as

$$\begin{aligned} \dot{c}_P(t) &= -k_b c_P(t) - k_e c_P(t) + 0.706k_a c_T(t) \\ \dot{c}_T(t) &= 1.407k_b c_P(t) - k_a c_T(t). \end{aligned} \quad (6.68)$$

with initial conditions $c_P(0) = 12.27$ mg per liter and $c_T(0) = 0$ mg per liter (M will drop out). Why is this initial condition for c_T appropriate?

The system (6.68) has a unique solution for any choice of k_a , k_b , and k_e . Our goal in what follows is to estimate k_a , k_b , and k_e by using the data from Table 6.1. We seek the parameter values that give the best fit to the data, in a least-squares sense.

Modeling Exercise 5.1.2 Use a computer algebra system to solve (6.68) symbolically in terms of t , with unspecified parameters k_a , k_b and k_e (it will be quite messy).

Modeling Exercise 5.1.3

- (a) Using whatever computer algebra system you have available, form a least-squares functional

$$S(k_a, k_b, k_e) = \sum_{j=1}^5 \sum_{k=1}^7 (c_P(t_k) - d_{j,k})^2,$$

where t_k denotes the k th time (measured in hours) at which data was taken (so $t_1 = 1/12$, $t_2 = 1/4$, and so on), and $d_{j,k}$ is the LSD plasma concentration of the j th subject at time t_k . Notice that the data point for subject 3 and time t_6 (4 hours) is missing, so you can't include that.

- (b) Minimize this least-squares functional using whatever command is appropriate, such as Maple's `Minimize` command, or Mathematica's `FindMinimum` command. It may be helpful to specify that the parameters k_a , k_b and k_e are all nonnegative.
- (c) Plot the best fit $c_P(t)$ and overlay its graph on a plot of the corresponding data from Table 6.1. Does the model provide a reasonable fit to the data?
- (d) Plot $c_T(t)$ on the time interval $0 \leq t \leq 8$ and overlay it with a plot of the subject's performance scores. It might be helpful to rescale the performance scores so they are on the same general scale as $c_T(t)$, and perhaps use 100 minus the performance score to re-center the data. How does the tissue LSD concentration correlate with the subject's performance on the arithmetic test? What might explain any discrepancies (were they all equally good at arithmetic to begin with)?

6.5.2 Project: Homelessness

In this project, which is based on the SIMIODE project [112], we develop a compartmental model to study eviction trends in a population of non-homeowner households using actual eviction rates. The data was compiled by the Eviction Lab at Princeton University, which has developed a nationwide database of evictions based on 83 million eviction records [85]. The model yields a linear system of ODEs, so we can readily calculate solutions and determine long-term trends. In Section 7.7 we will develop a second, nonlinear model that incorporates a carrying capacity, in this case, the number of rental units, as defined in Section 1.3.

Introduction

According to the National Law Center on Homelessness and Poverty, unaffordable rents and a lack of legal protections for renters have created a national eviction epidemic [94]. Matthew Desmond, author of *Evicted: Poverty and Profit in the American City* and director of the Eviction Lab at Princeton University, estimates that 2.3 million evictions were filed in the U.S. during 2016 (four evictions per minute). Desmond writes, “Eviction is a direct cause of homelessness, but it also is a cause of residential instability, school instability [and] community instability” [41]. In this project you will develop a mathematical model to study eviction trends in a city using an actual eviction rate.

A Linear Model

Suppose a city has 118,000 non-homeowner households and that this number remains constant each year. (For example, if three of these households move to a different city or purchase a home, then

three new non-homeowner households move into this city.) Furthermore, suppose that each of these households is either renting an apartment or house or is not renting due to having been evicted. We can define each of these subpopulations as functions of time t (years) in the following manner:

$R(t)$ is the number of renting households at time t

$E(t)$ is the number of evicted households at time t .

In order to simplify our calculations, we will consider the fraction of households in each category, that is, if N is the total number of non-homeowner households (in our example $N = 118,000$), then

$r(t) = R(t)/N$ is the fraction of renting households at time t

$e(t) = E(t)/N$ is the fraction of evicted households at time t .

In our model we will assume that a fixed proportion $\alpha > 0$ of the renting group become evicted each year. For example, according to data from the Eviction Lab at Princeton University, in 2016 North Charleston, South Carolina, had an eviction rate of 16.5%, so for North Charleston we would set $\alpha = 0.165$. We will also assume that a fixed proportion $\beta > 0$ of the evicted group become renters each year. In this model, the only way a renting household can leave the renting group is by transitioning to the evicted group. Similarly, the only way an evicted household can leave the evicted group is by transitioning to the renting group. It is helpful to represent this scenario with a flow diagram in Figure 6.10.

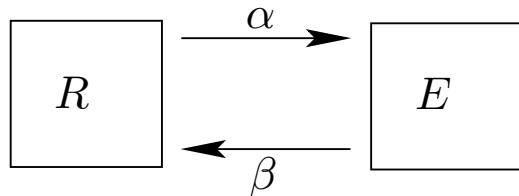


Figure 6.10: Flow diagram for eviction model. Here R is the number of renting households in a city, E is the number of households that have been evicted, α and β represent the rate at which each group moves to the other.

Modeling Exercise 5.2.1 Find equations for $\frac{dr}{dt}$ and $\frac{de}{dt}$ that satisfy these assumptions, with t measured in years. Explain the meanings of $\frac{dr}{dt}$ and $\frac{de}{dt}$ and what each component of your equations represents.

Modeling Exercise 5.2.2 Formulate the system as $\dot{\mathbf{u}} = \mathbf{A}\mathbf{u}$ with $\mathbf{u} = \langle r(t), e(t) \rangle$. Write out the matrix \mathbf{A} explicitly in terms of α and β .

Modeling Exercise 5.2.3 Compute the eigenvalues and eigenvectors for \mathbf{A} , and use them to write out a general solution to $\mathbf{u} = \langle r(t), e(t) \rangle$.

Modeling Exercise 5.2.4 For our model, we are only concerned with initial conditions that satisfy $r(0) + e(0) = 1$. Why? Use your results from Modeling Exercise 5.2.3 to write out the solution to $\dot{\mathbf{u}} = \mathbf{A}\mathbf{u}$ with initial data $\mathbf{u}(0) = \langle r_0, 1 - r_0 \rangle$ (that is, $r(0) = r_0$ and $e(0) = 1 - r_0$) explicitly in terms of α, β , and r_0 .

What does the model predict as t approaches infinity?

Modeling Exercise 5.2.5 Choose a value for α from the 2016 data for city evictions from the Eviction Lab [85], and state which city's information you chose. Set $\beta = \frac{1}{3}\alpha$, as a start. Suppose

that, initially, 95% of non-homeowner households are renters, so 5% are in the evicted group. Solve this initial value problem and plot $r(t)$ and $e(t)$ on one graph. What does this model predict about the percentage of non-homeowner households in each group, renting and evicted, in the long run? What will be the eventual ratio of renting to evicted non-homeowner households?

Modeling Exercise 5.2.6 How might this model be overly simplistic? What are some additional considerations that should be included in an eviction model?

6.5.3 Project: Tuned Mass Dampers

This project is adapted from the SIMIODE projects [77, 78] and the article [29].

Tuned Mass Dampers

The Taipei 101 tower in Taiwan was completed in 2004, and until 2009 it was the tallest skyscraper in the world. The tower is 509 meters tall, roughly 1670 feet. One of the challenges in building such a tall but lightweight structure is that of controlling the structure's tendency to sway. One obvious cause of swaying is an earthquake, but a more likely day-to-day source of swaying is the excitation caused by the wind blowing on the structure. For the comfort of the occupants, this motion must be controlled.

In the 1800s and early 1900s, most large civil engineering infrastructure such as buildings, dams, and bridges, was designed and built using rather conservative design processes that resulted in stiff, rigid, and heavy structures. Vibrations in structural components such as the floor beams caused by dynamic loads were rarely a concern. In the late 1900s, significant improvements in engineering design, engineering science, and construction methods resulted in lighter, more slender structures that were far more susceptible to large deflections resulting from dynamic wind or seismic stimuli. In the worst case, such forcing could result in resonance, as we studied in Chapter 4.

Rather than return to the old methods of overbuilt structures, modern engineers have sought more elegant and insightful ways to control vibration in structures. One solution is to add a tuned mass damper (TMD) to the skyscraper. A tuned mass damper is a small, damped spring-mass system attached to a larger spring-mass system (namely, the skyscraper). The tuned mass damper prevents the larger spring-mass system from vibrating excessively, especially near the resonant frequencies of the larger system. An effective tuned mass damper may have a mass that is a small fraction of the larger system. The tuned mass damper for the Taipei 101 tower has less than one-quarter of one percent of the tower's total mass. This tuned mass damper is shown in Figure 6.11.

The first uses of TMDs in the United States for large structures were in the John Hancock Building in Boston in 1977 [52] and Citigroup Center [91] (formerly known as the Citicorp Center) in New York in 1978. Since that time many different styles, including active TMDs and pendulum TMDs, have been employed, while diverse applications have been found through retrofitting large-span bridges and highways. Indeed, the current TMD exemplar is the 800-ton wind-compensating damper built into the center of the 509 meter tall Taipei 101 in Taiwan. This TMD consists of a huge spherical mass hung as a pendulum, and is visible from the restaurant on the 88th and 89th floors. Another recent use of a TMD is in the construction of the Grand Canyon Skywalk [67]. The use of TMDs is not limited to civil engineering. TMDs are also used in the design of surgery tables to mitigate the vibrations of surroundings during surgery [90]—think eye surgery in a surgical center and the New York City subway rumbling below the building. Many of the advancements in the application of TMDs in structures are found in the field of earthquake engineering. A list of significant structures that utilize TMDs is available from the National Information Service for Earthquake Engineering (NISEE) at UC Berkeley [93]. The Practical Engineering Project offers an excellent narrative, “What is a Tuned Mass Damper?” [47] with both technical, laboratory, and cultural elements at its YouTube site.



Figure 6.11: The TMD atop the Taipei 101 tower (from Wikimedia Commons, [45].)

Modeling a Tuned Mass Damper

A tall structure like a skyscraper may have complex dynamics, but to some approximation we might think of the building as a spring-mass-damper system as described by (4.3) in Chapter 4, just as we did with a single story building in Figure 4.1. Our primary interest is the lateral or back and forth motion of the building in some fixed plane of motion, and we will use $x_1(t)$ to indicate the displacement of the building from equilibrium at some altitude, such as the top of the building. We thus have a model

$$m_1 \ddot{x}_1 + c_1 \dot{x}_1 + k_1 x_1 = f(t), \quad (6.69)$$

where $f(t)$ is a driving force, perhaps the wind or a seismic disturbance, and m_1, c_1 and k_1 are appropriate constants. We will generally assume that $x_1(t_0) = 0$ and $\dot{x}_1(t_0) = 0$ at some initial time $t = t_0$. When stimulated by a driving force $f(t)$, the building will sway. The variables in (6.69) are subscripted with a 1 because we are going to add a second mass.

The role of this second mass m_2 is to help control or dampen any swaying of the building. The mass m_2 is that is attached to the mass m_1 , but with m_2 much less than m_1 . We will model the TMD in a simplified and abstracted version as illustrated in Figure 6.12.

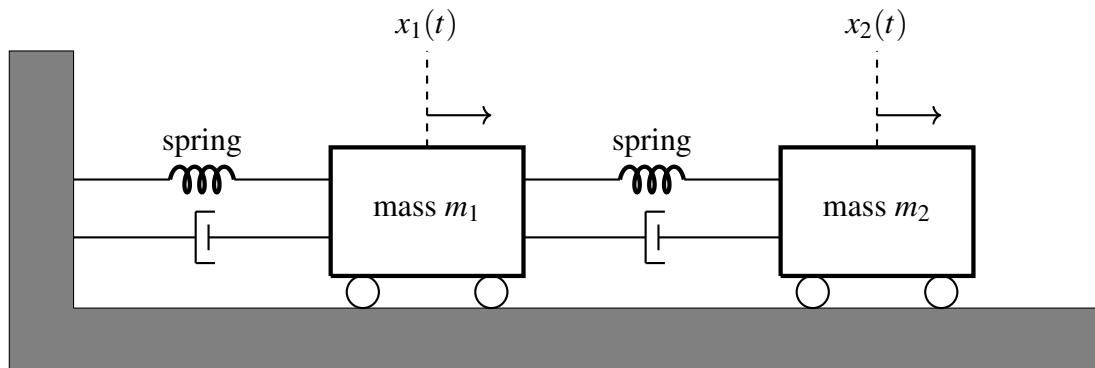


Figure 6.12: Tuned mass damper system abstracted as a smaller mass m_2 attached to a larger mass m_1 .

The situation here is almost exactly that of Example 6.2 in Section 6.1. In particular, as in Example 6.2, let us assume that the masses have negligible widths and that the rest length of spring 1 connecting mass m_1 to the left wall is L_1 . Similarly, the rest length of spring 2 connecting m_1 to m_2 is L_2 . We will use $x_1(t)$ to denote the position of mass m_1 with respect to the wall and $x_2(t)$ for the position of the mass m_2 . We assume that each spring obeys Hooke's law with spring constants k_1 for spring 1 and k_2 for spring 2. Suppose that the frictional force exerted by damper 1 (connecting the wall to mass m_1) on mass m_1 is $-c_1\dot{x}_1$, so this force is proportional and opposed to the rate at which first damper is elongating. In a very slight contrast to Example 6.2, let us suppose the force exerted by the second damper on the mass m_1 is $c_2(\dot{x}_2 - \dot{x}_1)$; this is proportional to the rate at which the second damper is elongating. The force exerted by the second damper on the mass m_2 is precisely the opposite, $-c_2(\dot{x}_2 - \dot{x}_1)$.

Modeling Exercise 5.3.1

- (a) Draw a free body diagram for m_1 to convince yourself these spring and friction forces are reasonable. Then use Newton's second law of motion to show that

$$m_1\ddot{x}_1 = -k_1(x_1 - L_1) + k_2(x_2 - x_1 - L_2) - c_1\dot{x}_1 + c_2(\dot{x}_2 - \dot{x}_1) + f(t), \quad (6.70)$$

where $f(t)$ is the driving force on m_1 . It may be helpful to review Reading Exercise 6.1.4 and the derivation of (6.8)-(6.9).

- (b) Draw a free body diagram for m_2 to convince yourself these spring and friction forces are correct. Then use Newton's second law of motion to show that

$$m_2\ddot{x}_2 = -k_2(x_2 - x_1 - L_2) - c_2(\dot{x}_2 - \dot{x}_1). \quad (6.71)$$

As in Example 6.2, let us make the convenient substitution $u_1(t) = x_1(t) - L_1$ and $u_2(t) = x_2(t) - L_1 - L_2$ (so $x_1(t) = u_1(t) + L_1$, $x_2(t) = u_1(t) + u_2(t) + L_1 + L_2$). Then (6.70)-(6.71) become

$$\begin{aligned} m_1\ddot{u}_1 &= -k_1u_1 + k_2(u_2 - u_1) - c_1\dot{u}_1 + c_2(\dot{u}_2 - \dot{u}_1) + f(t) \\ m_2\ddot{u}_2 &= -k_2(u_2 - u_1) - c_2(\dot{u}_2 - \dot{u}_1). \end{aligned} \quad (6.72)$$

Analysis of the Undamped Case

Modeling Exercise 5.3.2 Consider (6.72) with parameters $m_1 = 10$, $k_1 = 90$, $m_2 = 1$, and damping constants $c_1 = c_2 = 0$. Let us also decouple the second mass from the system by taking $k_2 = 0$, effectively removing the TMD from the system; no forces act on mass 2, so m_2 will remain motionless if it starts at rest. We begin with some analysis of this undamped system.

- (a) Suppose $f(t) = 0$, so the system is unforced. Write out the corresponding pair of second-order equations (6.72) and convert them to a system of four first-order equations in the form $\dot{\mathbf{w}} = \mathbf{Aw}$, with $w_1 = u_1$, $w_2 = \dot{u}_1$, $w_3 = u_2$, and $w_4 = \dot{u}_2$.
- (b) Compute the eigenvalues and eigenvectors for \mathbf{A} in part (a). What can you deduce about the unforced motion of the system—what natural frequencies does it possess?
- (c) Solve the system (6.72) from part (a) (or its first-order equivalent from part (a)) with initial conditions $u_1(0) = 1$, $\dot{u}_1(0) = 0$, $u_2(0) = 0$ and $\dot{u}_2(0) = 0$. Plot $u_1(t)$ (the displacement of mass 1). Reconcile what you see with your answers to part (b).
- (d) Now take $f(t) = \cos(3t)$ as the forcing function in (6.72); perhaps this is excitation from the wind, or seismic in nature. Solve the resulting forced system with zero initial data $u_1(0) = 0$, $\dot{u}_1(0) = 0$ and $u_2(0) = 0$, $\dot{u}_2(0) = 0$. Plot $u_1(t)$ on the range $0 \leq t \leq 20$. In view of what you learned in Section 4.4, what's going on here? Why might it be a problem?

Modeling Exercise 5.3.3 We again consider (6.72) with parameters $m_1 = 10$, $k_1 = 90$, $m_2 = 1$, and damping constants $c_1 = c_2 = 0$, but let us now take $k_2 = 10$, and with driving force $f(t) = \cos(3t)$. Notice that here the TMD mass is 10 percent of the mass of the building mass m_1 , which would be very large in a typical application.

- (a) Solve the system (6.72) with these parameters and with initial conditions $u_1(0) = 0$, $\dot{u}_1(0) = 0$, $u_2(0) = 0$, and $\dot{u}_2(0) = 0$, and plot $u_1(t)$ and $u_3(t)$ (the motion of masses 1 and 2) for $0 \leq t \leq 50$. Compare what you see to the plot obtained in part (d) of Modeling Exercise 5.3.2.
- (b) Formulate the system in part (a) as a system of four first-order ODEs $\dot{\mathbf{u}} = \mathbf{A}\mathbf{u} + \mathbf{f}(t)$, then compute the eigenvalues for \mathbf{A} . Can you explain why the system's response at driving frequency $\omega = 3$ is now less vigorous than it was in part (d) of Modeling Exercise 5.3.2?
- (c) Redo part (a) but vary the stiffness parameter k_2 in the range $k_2 = 1$ to $k_2 = 20$. In each case plot the motion of mass 1. What choice for k_2 gives the best result if the goal is to minimize the motion of mass 1? Defend your choice with plots or analysis.

Modeling Exercise 5.3.4 Again consider (6.72) with parameters $m_1 = 10$, $k_1 = 90$, and damping constants $c_1 = c_2 = 0$, but now with $m_2 = 0.1$, so the TMD mass m_2 is only 1 percent of the mass m_1 , a more realistic scenario. Redo part (c) of Reading Exercise 5.3.3, by experimenting with choices for k_2 in the range 0.5 to 2. What choice for k_2 is best? Defend your conclusion.

Modeling Exercise 5.3.5 Based on your analysis in Reading Exercises 5.3.2 to 5.3.4, offer a description of how to design a TMD to stop resonant phenomena when the system has no damping. A hint: in both Modeling Exercises 5.3.3 and 5.3.4, look at the resonant frequencies $\sqrt{m_1/k_1}$ and $\sqrt{m_2/k_2}$ for the individual spring-mass systems.

Some Analysis of the Damped Case

Modeling Exercise 5.3.6 Let us perform some analysis of the damped case. Take $m_1 = 10$, $k_1 = 90$, $c_1 = 3$, and $m_2 = 0.1$ in each part below.

- (a) With $k_2 = 0$, $c_2 = 0$, and $f(t) = 0$ (so no TMD is present, and no forcing) solve (6.72) with initial data $u_1(0) = 1$ and $u'_1(0) = u_2(0) = u'_2(0) = 0$. How long does it take the building motion to substantially decay?
- (b) With $k_2 = 0$, $c_2 = 0$, and $f(t) = \cos(3t)$ (so no TMD is present, but with forcing) solve (6.72) with zero initial data. What is the amplitude of the motion of mass m_1 ?
- (c) Let $k_2 = 1$, $c_2 = 0$ with $f(t) = \cos(3t)$ and solve (6.72) with zero initial data. Plot the motion of the mass m_1 out to at least time $t = 50$. Experiment with different values for k_2 and c_2 . What are the best choices for these parameters?
- (d) Based on parts (a)-(c), what recommendations can you make for the TMD design parameters?

For a more thorough analysis, especially for the damped case, see [78].

7. Nonlinear Systems of Differential Equations

7.1 Autonomous Nonlinear Systems and Direction Fields

This chapter is devoted to the analysis of autonomous first-order systems of differential equations. These systems are of the form (6.3), but in which the functions f_k have no explicit dependence on t . Let's state this plainly as a definition.

Definition 7.1.1 A system of first-order ODEs of the form

$$\begin{aligned}\dot{x}_1 &= f_1(x_1, \dots, x_n) \\ \dot{x}_2 &= f_2(x_1, \dots, x_n) \\ &\vdots \\ \dot{x}_n &= f_n(x_1, \dots, x_n)\end{aligned}\tag{7.1}$$

is said to be **autonomous**.

When convenient we will write the system (7.1) in the form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ where $\mathbf{x}(t) = \langle x_1(t), \dots, x_n(t) \rangle$ and $\mathbf{f}(\mathbf{x})$ is the vector valued function

$$\mathbf{f}(\mathbf{x}) = \langle f_1(\mathbf{x}), \dots, f_n(\mathbf{x}) \rangle.$$

Autonomous systems of ODEs are important for at least two reasons: they provide good models for many real-world systems, and they are amenable to qualitative analysis using graphical and geometric techniques. We can use these graphical and geometric methods to determine the nature of solutions, for example, their long-term behavior. Do solutions approach an equilibrium, grow without bound, or cycle periodically? The ideas are a natural extension of those developed in Section 2.3 for autonomous scalar ODEs. We'll initially focus on applying these techniques to linear systems and on the relationship of these methods to the eigenvalues for the system. We will then proceed to the analysis of nonlinear systems. In each section we'll concentrate on systems of two ODEs for two unknown functions, but we will indicate how the ideas extend to higher dimensions.

7.1.1 Some Nonlinear ODE Models

We've already encountered models involving linear systems of ODEs, for example, the LSD metabolism compartment model, double spring-mass models, tuned-mass-dampers, multiple-loop RLC circuits, and even models for the homelessness problem. Now let's consider a few physical situations that involve systems of nonlinear ODEs. These will provide motivation and illustration for the techniques we develop in this chapter.

The Struggle for Existence

This material is based on the SIMIODE project [127]. In Section 1.3 we encountered the logistic equation (1.10) for the growth of a species with population $u(t)$ in an environment that can support a maximum population of K individuals. This led to the logistic equation

$$\dot{u}(t) = ru(t)(1 - u(t)/K), \quad (7.2)$$

where r is the intrinsic growth rate parameter and K is the carrying capacity. The solution to this ODE with initial condition $u(0) = u_0$ was given by (1.11). In Exercises 2.2.8 and 3.4.8 you may have examined the fidelity of this model to actual data concerning the growth of a species of yeast.

What would happen if two different yeast species in the same vessel competed for resources, for example, nutrients and space? How would each population grow? Would one dominate, perhaps driving the other to extinction? In the 1930s in the former Soviet Union the scientist G. F. Gause considered this question in the works [50] and [51], in the service of improving vodka production.

Gause adopted a classic model for competing species, sometimes called the **Lotka-Volterra competing species model**. Let $u_1(t)$ denote the population of the first species of yeast and $u_2(t)$ the population of the second species of yeast. In this model we assume that each species would, in the absence of the other competing species, grow according to (7.2). Let's suppose that under these conditions the first species' intrinsic growth rate and carrying capacity are r_1 and K_1 , respectively, while the second species' corresponding parameters are r_2 and K_2 . When present in the same environment, however, the species interact and compete for resources. As a result, the presence of either yeast species should have a negative impact on the population of the other yeast species. One way to model this impact is by altering the logistic ODE (7.2) for $u(t) = u_1(t)$ to be

$$\dot{u}_1(t) = r_1 u_1(t) \left(\frac{K_1 - u_1(t) - au_2(t)}{K_1} \right) \quad (7.3)$$

for some nonnegative constant a ; the constant a quantifies the magnitude of the second species' impact on the population growth of the first. A similar modification is made to the equation governing the growth of the second species, to yield

$$\dot{u}_2(t) = r_2 u_2(t) \left(\frac{K_2 - u_2(t) - bu_1(t)}{K_2} \right) \quad (7.4)$$

for some nonnegative constant b .

Reading Exercise 7.1.1 Justify the modifications of (7.2) that lead to (7.3)-(7.4). In particular (when $a > 0$) how does the right sides of (7.3) behave as u_2 increases, and how would this affect the growth rate of the first yeast species? Similar considerations apply to the ODE (7.4).

Equations (7.3)-(7.4) form an autonomous pair of coupled nonlinear ODEs for the functions $u_1(t)$ and $u_2(t)$. These coupled ODEs typically have no elementary analytical solution. But with the techniques of this chapter we will be able to make certain conclusions concerning the behavior of solutions. We will be able to determine how the behavior of solutions depends on the population parameters r_1 , r_2 , K_1 , K_2 , a , and b . In the project "Parameter Estimation for Competing Species" in Section 7.7.3 we'll consider some of the data collected by Gause and use it to estimate the parameters in (7.3)-(7.4).

Reading Exercise 7.1.2 Show that each pair of constant functions below yields a solution to (7.3)-(7.4); note that in each case $\dot{u}_1 = \dot{u}_2 = 0$.

- $u_1(t) = 0$ and $u_2(t) = 0$ for all t .
- $u_1(t) = K_1$ and $u_2(t) = 0$ for all t .
- $u_1(t) = 0$ and $u_2(t) = K_2$ for all t .

What physical interpretation would you attach to each of these three solutions—what does it mean for the populations involved? Then verify that there is a fourth solution

$$u_1(t) = \frac{K_1 - K_2 a}{1 - ab}, \quad u_2(t) = \frac{K_2 - K_1 b}{1 - ab}, \quad (7.5)$$

in which $u_1(t)$ and $u_2(t)$ are constant, at least if $ab \neq 1$. What condition must be met for this solution to be physically relevant? Hint: u_1 and u_2 are populations.

As in the scalar case, constant solutions $u_1(t) = u_1^*$, $u_2(t) = u_2^*$ are called **equilibrium solutions**, and (u_1^*, u_2^*) is a **fixed point**.

Epidemic Models

This material is drawn from the SIMIODE project [89], which itself draws on material from [92].

A boarding school is a relatively closed community in which all students live on campus, teachers tend to live on or near campus, and students do not regularly interact with people outside the boarding school community. Table 7.1 gives data for an influenza outbreak at a boarding school in England during which there were no fatalities. These data points were compiled from the Communicable Disease Surveillance Center [36, p. 587] and are given as an example by Murray in his book on mathematical biology [92]. The data values were extracted from the graph found in [107].

Time (days)	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Bedridden	1	3	25	72	222	282	256	233	189	123	70	25	11	4

Table 7.1: Total Number of Bedridden Boys on Day t .

There were 763 boys at the English boarding school from which our data were obtained. We can see from Table 7.1 that the initial number of bedridden students is one. The data given is the number of bedridden boys rather than the number of boys with influenza; let us assume that the number of boys who are bedridden on day t consists exactly of all those who are newly infected and all continuing infections (students who are still on bedrest after being infected on a prior day). Let us also assume that no boys are bedridden for any other reason during this period of time and also that all boys with influenza will be not only symptomatic but bedridden. (One may wish to test the effects of this last assumption on the model in the analysis phase. Another, perhaps more realistic, assumption is that the bedridden boys represent a fixed percentage of the infected students.) These assumptions, together with the exclusion of teachers and staff from the population, imply that the number of students infected on day zero is one, or, $I(0) = 1$, where $I(t)$ is the number of students symptomatic on day t (either infectious or infected).

Our goal is to develop and analyze a model for how the influenza epidemic propagates through the boarding school population as a function of time. The progress of the epidemic depends on various parameters that appear in the model. Such a model could be used to examine various strategies for controlling the epidemic, such as isolating sick students for a period of time. One very common approach to modeling epidemics is called the **SIR epidemic model** as illustrated in Figure 7.1. This is a compartmental model with three compartments. Each individual is either susceptible to the disease and hence in the “S” compartment, or the individual is actively infected and so in the

“I” compartment, or the individual has recovered and is in the “R” compartment. The susceptible compartment is for those who have never had the disease and so remain susceptible to infection. In addition to being sick, the individuals in the infected compartment are capable of infecting others. Those in the recovered compartment are considered to have immunity to the disease, which we will assume is permanent. As $I(t)$ indicates the number of infected persons, so $S(t)$ and $R(t)$ will indicate the number of susceptible and recovered students, respectively.

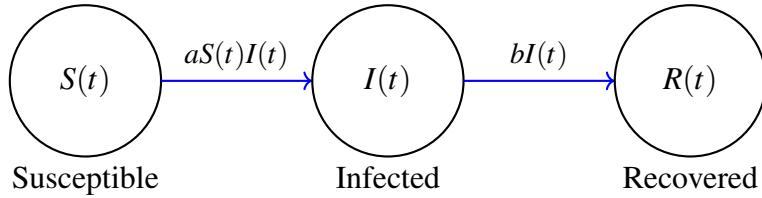


Figure 7.1: Susceptible, recovered, and infected (SIR) compartmental model of an epidemic.

As in any compartment model, we must specify the rate at which the quantities of interest (in this case, people) move between the compartments. In the classic SIR model, susceptible and infected people interact at a rate that is proportional to the product SI . The reasoning is that, for any fixed value of S , if the value of I is doubled then the number of interactions between susceptible and infected people should double, and a similar result should hold if I is fixed but S is doubled; the quantity SI captures this observation. Moreover, we suppose in the model that each such interaction carries a fixed risk of the susceptible person becoming infected and moving from the S to the I compartment. This is captured by the $aS(t)I(t)$ label above the arrow from the S compartment to the I compartment in Figure 7.1: the likelihood of infection is proportional to the number of interactions between the susceptible (S) and the infected (I) populations. The constant of proportionality a depends, for example, on the infectiousness of the disease. Movement from the I compartment to the R compartment is assumed to occur at a rate proportional to the number of infected people: all else being equal, if there are twice as many infected people then the number of people getting better per unit time should double.

With these observations we can posit the model

$$\begin{aligned}\dot{S} &= -aSI \\ \dot{I} &= aSI - bI \\ \dot{R} &= bI.\end{aligned}\tag{7.6}$$

Reading Exercise 7.1.3 In light of the above assumptions, defend the model (7.6). What critiques can you make concerning these assumptions?

Reading Exercise 7.1.4 Compute $\dot{S} + \dot{I} + \dot{R}$ for (7.6), and explain why the result makes sense.

Equations (7.6) constitute a system of three coupled autonomous nonlinear ODEs for the functions $S(t)$, $I(t)$, and $R(t)$. We will analyze and improve on this model in the sections to come.

The Nonlinear Pendulum

The equation of a damped pendulum of length L swinging under the influence of gravity was derived in the Modeling Project “The Pendulum 2” in Section 4.6. The angle $\theta(t)$ that the pendulum makes with the vertical as it swings obeys the nonlinear second-order ODE

$$\ddot{\theta}(t) + c\dot{\theta}(t) + \frac{g}{L} \sin(\theta(t)) = 0,\tag{7.7}$$

where c is a nonnegative frictional or damping constant. Equation (7.7) has no analytical solution. However, the behavior of solutions can be deduced using the techniques of this chapter. To proceed,

we convert (7.7) into an autonomous pair of coupled first-order ODEs using the procedure of Section 6.1, by letting $x_1(t) = \theta(t)$ and $x_2(t) = \dot{\theta}(t)$. We obtain

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{g}{L} \sin(x_1) - cx_2.\end{aligned}\tag{7.8}$$

We will examine the behavior of this system, for both the damped ($c > 0$) and undamped ($c = 0$) cases later in this chapter.

Reading Exercise 7.1.5 If the angle $\theta(t)$ that the pendulum makes with respect to the vertical remains close to zero then $x_1 \approx 0$ and the approximation $\sin(x_1) \approx x_1$ is reasonable. Make this substitution in (7.8) and show that the resulting linear system can be formulated as $\dot{\mathbf{x}} = \mathbf{Ax}$ with

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -g/L & -c \end{bmatrix}$$

and that \mathbf{A} has eigenvalues

$$-\frac{c}{2} \pm \frac{\sqrt{c^2 - 4g/L}}{2}.$$

Show that if $c > 0$ then the eigenvalues are always either both real and negative or complex conjugate with negative real part. What significance does this have for the motion of the pendulum?

7.1.2 Direction Fields

Recall the technique of phase portraits (or *phase line portraits*) for an autonomous scalar ODE, $u' = f(u)$ from Section 2.3. This graphical technique provides a simple method for determining the behavior of solutions to the ODE, without actually solving the equation. We can do something similar for an autonomous system of ODEs. We'll first illustrate the technique with systems of two ODEs in two unknown functions $x_1(t)$ and $x_2(t)$.

A Spring-Mass Example

Let's begin with a concrete example involving a spring-mass-damper system, since we have accumulated some intuition about the behavior of such a system. Consider the unforced, underdamped spring-mass system with mass $m = 1$, damping constant $c = 1$, and spring constant $k = 1$ governed by $\ddot{u}(t) + \dot{u}(t) + u(t) = 0$, where $u(t)$ denotes the displacement of mass from equilibrium. If we let $x_1 = u$ and $x_2 = \dot{u}$, this ODE is equivalent to the linear ODE system

$$\begin{aligned}\dot{x}_1 &= \underbrace{x_2}_{f_1(x_1, x_2)} \\ \dot{x}_2 &= \underbrace{-x_1 - x_2}_{f_2(x_1, x_2)},\end{aligned}\tag{7.9}$$

where $f_1(x_1, x_2) = x_2$ and $f_2(x_1, x_2) = -x_1 - x_2$. The function $x_1(t)$ is the position of the mass as a function of time and $x_2(t)$ is its velocity. The set of all points (x_1, x_2) in \mathbb{R}^2 is called the **phase space** for this system (or **phase plane**, since we're in two dimensions). If we know that a solution $\langle x_1(t), x_2(t) \rangle$ to (7.9) passes through a point $(x_1, x_2) = (a_1, a_2)$ in the phase space at a time $t = t^*$, we can determine the past and future trajectory of the solution by finding the solution to (7.9) that satisfies the conditions $x_1(t^*) = a_1$ and $x_2(t^*) = a_2$.

But even without solving, we can determine something about the solution to (7.9) that passes through the point $(x_1, x_2) = (a_1, a_2)$ at a time $t = t^*$. Let us write $\mathbf{x}(t) = \langle x_1(t), x_2(t) \rangle$ for a vector-valued description of a solution pair $x_1(t)$ and $x_2(t)$ to (7.9). Recall from multivariable calculus

that the vector

$$\dot{\mathbf{x}}(t^*) = \langle \dot{x}_1(t^*), \dot{x}_2(t^*) \rangle$$

is tangent to the curve parameterized by $\mathbf{x}(t)$ at the point $\mathbf{x}(t^*)$. Thus the vector $\dot{\mathbf{x}}(t^*)$ tells us in what direction the solution is moving as it passes through the point $\mathbf{x}(t^*) = \langle a_1, a_2 \rangle$. We can compute this vector by using (7.9), to find

$$\begin{aligned} \dot{\mathbf{x}}(t^*) &= \langle \dot{x}_1(t^*), \dot{x}_2(t^*) \rangle \\ &= \langle f_1(x_1(t^*), x_2(t^*)), f_2(x_1(t^*), x_2(t^*)) \rangle \\ &= \langle f_1(a_1, a_2), f_2(a_1, a_2) \rangle. \end{aligned} \quad (7.10)$$

Because the system is autonomous, the functions f_1 and f_2 do not depend on t . The vector $\dot{\mathbf{x}}(t^*)$ defined by (7.10) therefore does not depend on t^* , but only on the point (a_1, a_2) itself.

■ **Example 7.1** Suppose that a solution for the system (7.9) passes through the point $(1, 2)$ in the phase plane; this corresponds physically to a mass with position $u = 1$ and velocity $\dot{u} = 2$. From (7.10) the solution curve at this point is tangent to the vector

$$\dot{\mathbf{x}} = \langle f_1(1, 2), f_2(1, 2) \rangle = \langle 2, -3 \rangle.$$

In physical terms, $\dot{\mathbf{x}} = \langle 2, -3 \rangle$ tells us that $\dot{x}_1 = 2$ (the mass is moving with velocity 2, which we already knew) and $\dot{x}_2 = -3$; since $\dot{x}_2 = \ddot{u}$, the mass has acceleration -3 .

■

Reading Exercise 7.1.6 Use (7.10) to compute the vector $\dot{\mathbf{x}}$ for a solution that passes through the point $(-1, 1)$ in the phase plane (recall f_1 and f_2 are defined in (7.9)). Use this vector to express the direction in which the solution is moving, as a unit vector. Then interpret the situation physically; if $(x_1, x_2) = (-1, 1)$, what are the position and velocity of the mass? What does the value of $\dot{\mathbf{x}}$ tell you about the velocity and acceleration of the mass at this instant?

Plotting The Direction Field and Solution Curves

We can plot the vector $\dot{\mathbf{x}}$ in Example 7.1 with its tail at the point $(1, 2)$ to indicate the direction in which the solution to (7.9) is moving as it passes through $(1, 2)$. We can perform the same computation for the vector with its tail at $(-1, 1)$ found in Reading Exercise 7.1.6, and so indicate the direction that a solution to (7.9) moves as it passes through the point $(-1, 1)$. Performing this same computation for many more points in the phase plane and plotting them all at once gives a more complete picture of how solutions move. In Figure 7.2 we show two graphical variations obtained by choosing many points (a_1, a_2) in the range $-2 \leq a_1, a_2 \leq 2$, computing the vector $\dot{\mathbf{x}}(a_1, a_2) = \langle f_1(a_1, a_2), f_2(a_1, a_2) \rangle$ and then plotting each such vector with its tail at the point (a_1, a_2) . The left panel of Figure 7.2 shows these vectors scaled to $1/5$ of their actual length; this has the benefit of making the vectors fit into the picture but without changing their direction. This figure is an example of a **direction field** for an autonomous system. The vectors vary widely in length, because $\dot{\mathbf{x}}(a_1, a_2)$ varies widely in magnitude. Since our primary interest is the direction in which solutions move, it can be more insightful (and more aesthetically pleasing) to scale all vectors to the same length. This is shown in the right panel of Figure 7.2, in which all vectors are scaled to a fixed length of 0.2. We will generally adopt the convention of scaling all vectors to the same length.

Recall Remark 6.1.1 in Section 6.1; a solution to the ODE system (7.9) that passes through a point may be interpreted as a parameterized curve $x_1 = x_1(t), x_2 = x_2(t)$ in the phase plane. Based on either panel in Figure 7.2, we can sketch the solution curve that passes through any point (a_1, a_2) by starting at this point and following the arrows of the direction field. Figure 7.3 provides an illustration in which the direction field (as presented in the right panel of Figure 7.2) is shown but now overlayed with a solution curve to (7.9) that passes through the point $(x_1, x_2) = (2, 0)$.

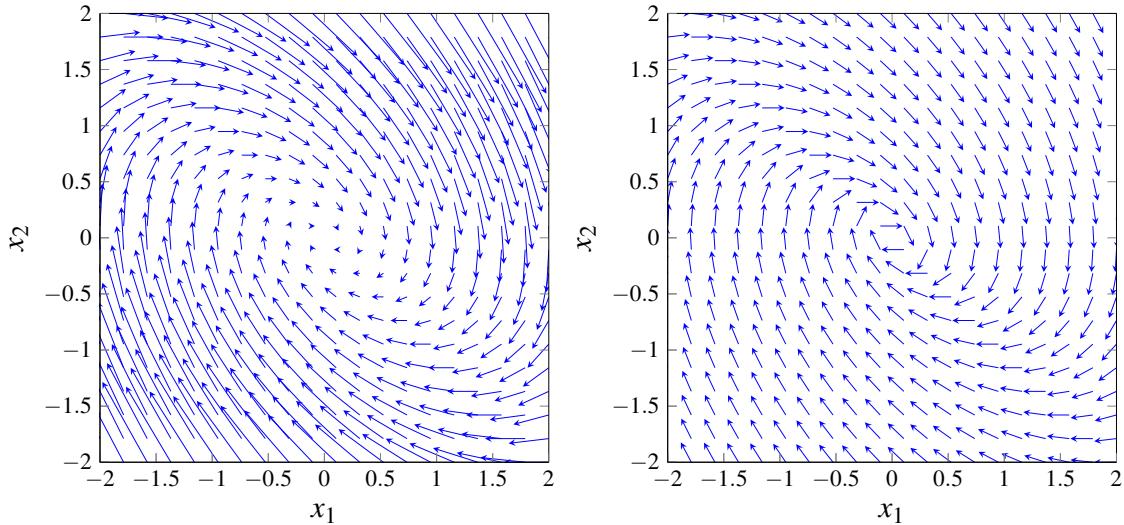


Figure 7.2: Left panel: direction field for (7.9) with vectors (7.10), scaled to $1/5$ actual length. Right panel: direction field for (7.9) with vectors (7.10), but with all vectors scaled to fixed length 0.2.

Interpretation of the Direction Field and Solution Curves

It is important to note that although this curve doesn't yield the precise value of the solution at any time, the curve does indicate the long-term behavior of the solution. Specifically, it is clear that the solution curve parameterized by $\mathbf{x}(t)$ that starts at $(2, 0)$ in Figure 7.3 spirals into the origin. This observation gives us essential information about the physical behavior of the damped spring-mass system $\ddot{u}(t) + \dot{u}(t) + u(t) = 0$ on which Figure 7.3 is based. The position of the mass is $x_1(t) = u(t)$ and its velocity is $x_2(t) = \dot{u}(t)$. The graph of the solution curve indicates how the mass moves. The mass starts at $(x_1, x_2) = (u, \dot{u}) = (2, 0)$ (the time is irrelevant). That is, the spring was stretched 2 units to the right (elongated) and then released with no initial velocity. As t increases and we move along the curve, $x_1 = u$ decreases toward 0 and $x_2 = \dot{u}$ becomes negative; in the actual spring-mass system, the spring is contracting and the mass is moving in the direction of decreasing u . Eventually we pass through a point where $x_1 = u \approx -0.3 < 0$ and $x_2 = \dot{u} = 0$; the spring is in compression and the mass is momentarily stopped. Subsequently $x_1 = u$ begins to increase and $x_2 = \dot{u} > 0$, as the mass moves in the direction of increasing u , until it momentarily comes to rest, and the pattern repeats. The mass oscillates back and forth. In the long run the solution curve approaches the point $(x_1, x_2) = (0, 0)$, or $(u, \dot{u}) = (0, 0)$, which indicates that the mass asymptotically approaches a resting position at the equilibrium length of the spring. It should be clear from Figure 7.3 that this is the fate of any solution to $\ddot{u} + \dot{u} + u = 0$.

Reading Exercise 7.1.7 Solve the system $\dot{x}_1 = x_2$, $\dot{x}_2 = -x_1 - x_2$ with $x_1(0) = 2$ and $x_2(0) = 0$ (or equivalently, $\ddot{u} + \dot{u} + u = 0$ with $u(0) = 2$ and $\dot{u}(0) = 0$). Plot $x_1(t)$ and $x_2(t)$ or $u(t)$ and $\dot{u}(t)$ for $0 \leq t \leq 10$ and reconcile these graphs with the solution trajectory shown in Figure 7.3.

7.1.3 A Nonlinear Direction Field Example

The procedure for constructing a direction field can be carried out for any autonomous system of two coupled ODEs in two unknowns. To illustrate, consider the competing species equations (7.3)-(7.4) in the specific case $r_1 = 1$, $K_1 = 2$, $r_2 = 2$, $K_2 = 3$, $a = 0.2$, and $b = 0.45$. With these

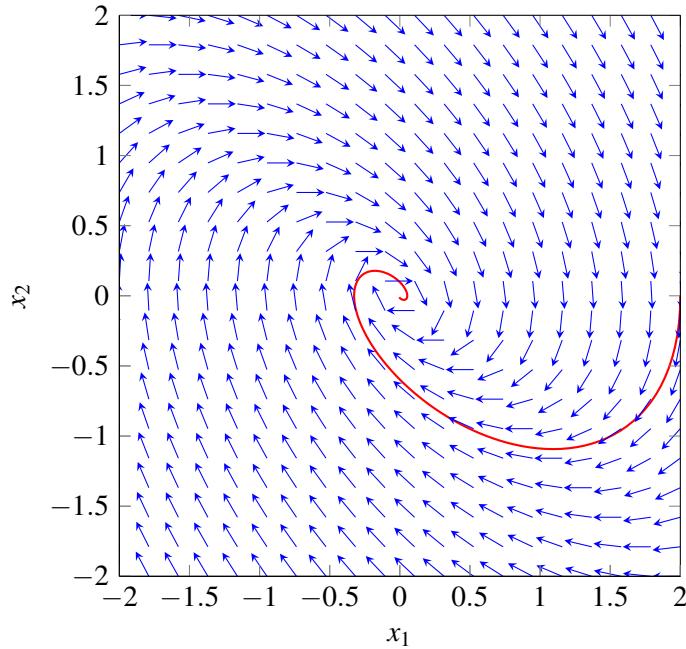


Figure 7.3: Direction field and a solution curve to (7.9); the solution passes through $(x_1, x_2) = (2, 0)$.

parameters the equations for the species populations $u_1(t), u_2(t)$ are

$$\begin{aligned}\dot{u}_1 &= \underbrace{u_1(1 - u_1/2) - 0.1u_1u_2}_{f_1(u_1, u_2)} \\ \dot{u}_2 &= \underbrace{2u_2(1 - u_2/3) - 0.3u_1u_2}_{f_2(u_1, u_2)},\end{aligned}\tag{7.11}$$

with f_1 and f_2 as indicated. Figure 7.4 shows the direction field for the system (7.11) in the region $0 \leq u_1, u_2 \leq 5$; negative values for u_1 or u_2 are not physically relevant here. As in the previous example, this direction field is obtained by choosing a large number of points (u_1, u_2) in this region, computing the vector $\langle \dot{u}_1, \dot{u}_2 \rangle = \langle f_1(u_1, u_2), f_2(u_1, u_2) \rangle$ at each point, and then plotting each such vector with its tail at (u_1, u_2) . This computation does not require us to solve the ODEs (7.11), which we can't do anyway; it merely requires arithmetic.

By using the direction field we can sketch a solution with any initial condition by following the arrows. Several solution curves are shown in Figure 7.4, starting at the (u_1, u_2) points $(3, 3), (1, 5)$, and $(4, 0.2)$, respectively. It appears that in each case the solution approaches a fixed point somewhere near $(u_1, u_2) \approx (1.6, 2.3)$. Furthermore, based on the appearance of the direction field it seems that for any initial populations the corresponding solution will approach this fixed point. The u_1 species population stabilizes at $u_1 \approx 1.6$ and the u_2 species population stabilizes at $u_2 \approx 2.3$, and the species coexist.

Reading Exercise 7.1.8 Find the precise point (p_1, p_2) to which the solution trajectories in Figure 7.4 converge. Hint: It should be the case that if (u_1, u_2) converges to (p_1, p_2) then \dot{u}_1 and \dot{u}_2 both converge to 0, so that from (7.11) we have $p_1(1 - p_1/2) - 0.1p_1p_2 = 0$ and $2p_2(1 - p_2/3) - 0.3p_1p_2 = 0$. Find all points (p_1, p_2) that satisfy this pair of algebraic equations; there should be four points, one of which is the point to which solutions in Figure 7.4 converge. What are the other three points, and what physical significance do they have? Compare to Reading Exercise 7.1.2.

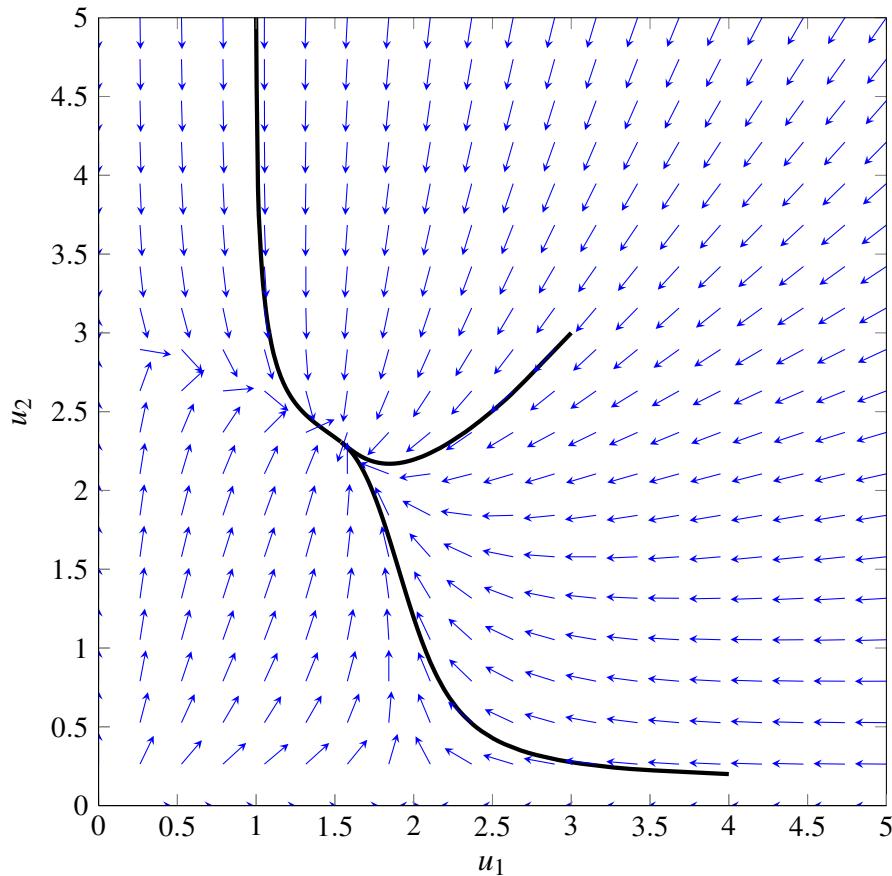


Figure 7.4: Direction field for the competing species system (7.11) with $r_1 = 1$, $K_1 = 2$, $r_2 = 2$, $K_2 = 3$, $a = 0.2$, and $b = 0.45$, with several solution trajectories.

If we increase the competition parameters a and b from $a = 0.2$ and $b = 0.45$ to $a = 2$ and $b = 1.5$ and recompute the direction field with the same initial conditions for the solutions, we obtain the result in Figure 7.5. It now appears that all solutions approach the equilibrium point $(u_1, u_2) = (0, 3)$ and coexistence does not occur—the first species is driven to extinction and the second species' population stabilizes at $u_2 = 3$.

The direction field allows us to make these conclusions without solving the ODEs, and even without having specific choices for system parameters like r_1 , K_1 , r_2 , K_2 , a , and b . The techniques we develop will allow us to determine not only how solutions behave, but how this behavior is influenced by the parameters in the ODE. In the competing species model we will be able to determine what choices for the parameters allow for coexistence of the species and what values lead to the certain extinction of one species or the other. These ideas can be used to analyze many autonomous systems.

Reading Exercise 7.1.9 Repeat Reading Exercise 7.1.8 with the parameters $a = 2$ and $b = 1.5$ in (7.11).

7.1.4 Direction Fields in Higher Dimensions

Consider an autonomous system of the form (7.1) with n ODEs and n functions $x_1(t), \dots, x_n(t)$. The vector-valued function

$$\mathbf{f}(\mathbf{x}) = \langle f_1(\mathbf{x}), \dots, f_n(\mathbf{x}) \rangle$$

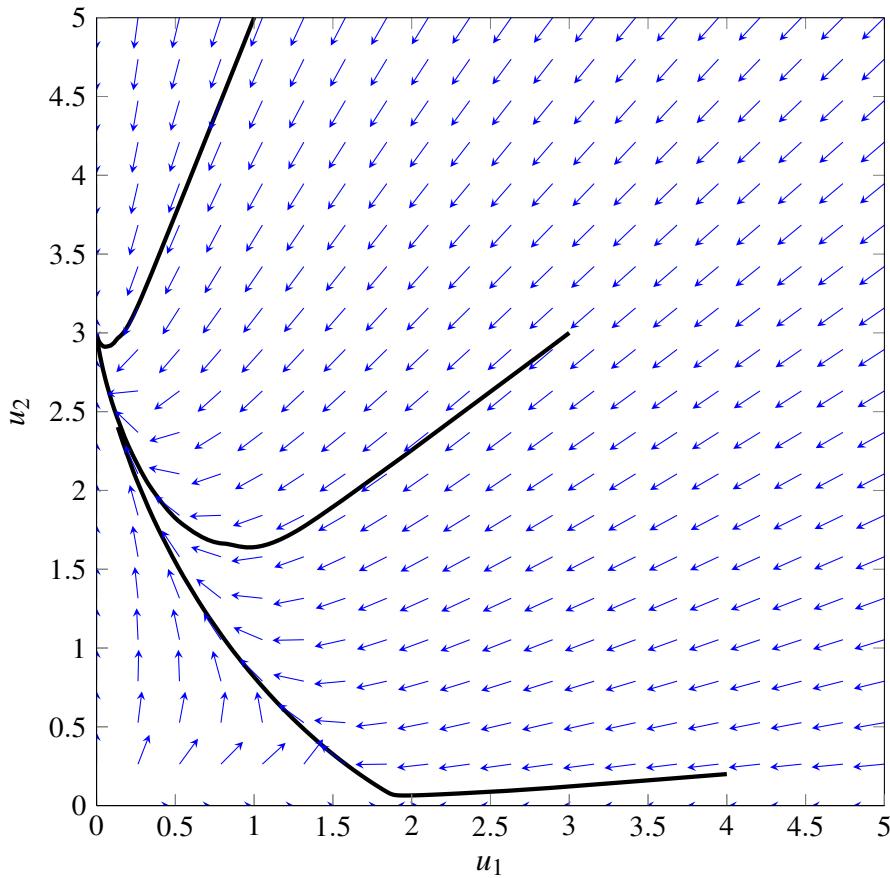


Figure 7.5: Direction field for the competing species system (7.11) but with parameters $r_1 = 1$, $K_1 = 2$, $r_2 = 2$, $K_2 = 3$, $a = 2$ and $b = 1.5$, embodying more intense competition, with several solution trajectories.

with $\mathbf{x} = \langle x_1, \dots, x_n \rangle$ dictates the direction that a solution moves when passing through the point $x_1 = a_1, \dots, x_n = a_n$ in \mathbb{R}^n , in the same manner that (7.10) does in \mathbb{R}^2 . The difficulty, of course, is that we can't sketch or easily visualize this direction field in more than three dimensions. Indeed, even though many software packages can sketch direction fields in three dimensions, the result is usually a confusing jumble of arrows that yield little insight. Nonetheless, the main idea behind the direction field, that the functions f_k on the right side (7.1) specify the direction in which solutions move, is an important intuitive and geometric insight to keep in mind.

7.1.5 Exercises

Exercise 7.1.1 For each system of ODEs $\dot{x}_1 = f_1(x_1, x_2)$, $\dot{x}_2 = f_2(x_1, x_2)$ and each of the points $(a_1, a_2) = (1, 1), (1, 2), (2, 1), (2, 2)$:

- Compute the vector $\langle f_1(a_1, a_2), f_2(a_1, a_2) \rangle$ (recall (7.10)).
 - Sketch the vector you obtain, with tail at (a_1, a_2) , on axes that span the range $0 \leq x_1, x_2 \leq 4$, to form a (crude) direction field. Scale the vectors to have length 1.
- $f_1(x_1, x_2) = (x_1 + x_2)/2$, $f_2(x_1, x_2) = x_2/2$.
 - $f_1(x_1, x_2) = x_1^2/2 - x_2$, $f_2(x_1, x_2) = x_1 + 1$.
 - $f_1(x_1, x_2) = x_2 - 3/2$, $f_2(x_1, x_2) = -x_1 + 3/2$.

(d) $f_1(x_1, x_2) = x_1 - 3/2, f_2(x_1, x_2) = x_2 - 3/2.$

Exercise 7.1.2 Use whatever technology you have to sketch a direction field for equations (7.3)-(7.4) with parameter values $r_1 = 1, r_2 = 1, K_1 = 3, K_2 = 3, a = 2$, and $b = 2$, on the region $0 \leq u_1, u_2 \leq 5$. Sketch a few representative solutions. To what point(s) do solutions converge? Does the long-term solution behavior depend on the initial condition? What does this say about the populations?

Exercise 7.1.3 Use whatever technology you have to sketch a direction field for (7.3)-(7.4) with parameter values $r_1 = 1, r_2 = 1, K_1 = 3, K_2 = 3, a = 0$, and $b = 0$, on the range $0 \leq u_1, u_2 \leq 5$. This is the situation in which there is no competition at all. Sketch a few representative solutions. How do solutions behave? Why does this make sense?

Exercise 7.1.4 Use whatever technology you have to sketch a direction field for the damped nonlinear pendulum system (7.8) with $L = 1, c = 1$, and $g = 9.8$ on the range $-2 \leq x_1, x_2 \leq 2$. Sketch a few solution trajectories. What do solutions do as t approaches infinity? What does this say about the motion of the pendulum, and why does this make sense?

Exercise 7.1.5 Use whatever technology you have to sketch a direction field for the undamped nonlinear pendulum system (7.8) with $L = 1, c = 0$, and $g = 9.8$ on the range $-2 \leq x_1, x_2 \leq 2$. Sketch a few solution trajectories. What do solutions do as t approaches infinity? What does this say about the motion of the pendulum, and why does this make sense?

Exercise 7.1.6 The SIR epidemic model (7.6) might be considered a system of two ODEs $\dot{S} = -aSI$ and $\dot{I} = aSI - bI$ in unknowns S and I (we simply ignore the third equation for R , since R doesn't appear in either of the first two equations.) Use whatever technology you have to sketch a direction field for $\dot{S} = -aSI$ and $\dot{I} = aSI - bI$ with $a = b = 1$ on the range $0 \leq x_1, x_2 \leq 2$. Sketch a few solution trajectories. What do solutions do as t approaches infinity? What does this say about the epidemic, in particular, about the number of susceptible, infected, and recovered students over time?

7.2 Direction Fields and Phase Portraits for Linear Systems

In this section we'll take a closer look at direction fields for autonomous linear systems of the form $\dot{\mathbf{x}} = \mathbf{Ax}$ and also $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{b}$. If the system is autonomous then the vector \mathbf{b} is necessarily constant. We will focus on how the eigenvalues of \mathbf{A} affect the direction field and the behavior of solutions. Our primary interest is the case in which the system is two-dimensional and \mathbf{A} is invertible, but we will indicate how these conclusions can be extended to higher dimensions and summarize results for the case in which \mathbf{A} is singular. Finally, we'll begin to consider how one can glean information about direction fields and the behavior of solutions even when the ODEs contain unspecified parameters. These powerful techniques will be invaluable later for analyzing nonlinear systems of ODEs.

7.2.1 Direction Fields for Homogeneous Linear Systems

Consider a linear constant-coefficient autonomous system of n ODEs of the form

$$\dot{\mathbf{x}} = \mathbf{Ax}.$$

As already noted we will initially consider the two-dimensional case, so \mathbf{A} is 2×2 , and for now we assume that \mathbf{A} is invertible. As discussed in Appendix B, a matrix is invertible if and only if all of its eigenvalues are nonzero. We will first focus on the case in which $\mathbf{b} = \mathbf{0}$, the zero vector.

The Case in Which \mathbf{A} Is Invertible With Real Eigenvalues

Suppose \mathbf{A} is invertible. In this case the only equilibrium solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ is the constant vector \mathbf{x} that satisfies $\dot{\mathbf{x}} = \mathbf{Ax} = \mathbf{0}$. This means that $\mathbf{x} = \mathbf{A}^{-1}\mathbf{0} = \mathbf{0}$.

Let's start with the case in which the eigenvalues λ_1 and λ_2 for \mathbf{A} are real, with linearly independent eigenvectors \mathbf{v}_1 and \mathbf{v}_2 ; equivalently, we assume that neither \mathbf{v}_1 nor \mathbf{v}_2 is a scalar multiple of the other. Since \mathbf{A} is invertible, both λ_1 and λ_2 are nonzero. From Section 6.2 a general solution $\mathbf{x}_h(t)$ to the homogeneous system $\dot{\mathbf{x}} = \mathbf{Ax}$ is

$$\mathbf{x}_h(t) = c_1 e^{\lambda_1 t} \mathbf{v}_1 + c_2 e^{\lambda_2 t} \mathbf{v}_2. \quad (7.12)$$

Our primary interest is how these eigenvalues shape the direction field for the system.

There are three main cases to consider:

- If λ_1 and λ_2 are both negative then from (7.12) we see that all nontrivial solutions $\mathbf{x}_h(t)$ approach the origin as t approaches infinity. Also, $|\mathbf{x}_h(t)|$ grows without bound as t decreases to minus infinity. This case is illustrated by the direction field in the top left panel of Figure 7.6, in which the direction field is shown with a few solution trajectories; all solutions approach the origin as t increases. The origin here is called an **asymptotically stable node** or **sink**. In the special case that $\lambda_1 = \lambda_2$ the origin is called a **stable star point**.
- If λ_1 and λ_2 are both positive then from (7.12) we see that nonzero solutions grow without limit as t increases; that is, the magnitude $|\mathbf{x}_h(t)|$ approaches infinity as t approaches infinity. Solutions approach the origin as t decreases to minus infinity. This case is illustrated by the direction field in the top right panel of Figure 7.6, in which the direction field is shown with a few solution trajectories; all solutions here radiate away from the equilibrium point $\langle 0, 0 \rangle$. In this case the origin $\langle 0, 0 \rangle$ is called an **unstable node** or **source**. In the special case that $\lambda_1 = \lambda_2$ the origin is called an **unstable star point**.
- If $\lambda_1 < 0$ and $\lambda_2 > 0$ then $\mathbf{x}_h(t)$ approaches a larger and larger multiple of the eigenvector \mathbf{v}_2 as t increases. That is, $\mathbf{x}_h(t) \rightarrow c_2 e^{\lambda_2 t} \mathbf{v}_2$ as $t \rightarrow \infty$, as long as $c_2 \neq 0$ (generically the case). But as $t \rightarrow -\infty$ then we find that $\mathbf{x}_h(t)$ approaches $c_1 e^{\lambda_1 t} \mathbf{v}_1$, a multiple of \mathbf{v}_1 . If $\lambda_1 > 0$ and $\lambda_2 < 0$ the roles of \mathbf{v}_1 and \mathbf{v}_2 are reversed. The equilibrium solution $\langle 0, 0 \rangle$ in this case is a **saddle point**. This situation is illustrated by the direction field in the bottom panel of Figure 7.6. For this particular direction field the corresponding matrix has eigenvectors $\mathbf{v}_1 = \langle 1, 1 \rangle$ (eigenvalue -1) and $\mathbf{v}_2 = \langle 1, -2 \rangle$ (eigenvalue 1); you can see these eigenvectors in the behavior in the direction field and solutions.

■ **Example 7.2** Consider an overdamped spring-mass system with mass position $u(t)$ governed by $m\ddot{u} + c\dot{u} + ku = 0$, where m, c , and k are all positive. Since the system is overdamped we have $c^2 - 4mk > 0$. We can formulate this as a system by taking $x_1 = u$ and $x_2 = \dot{u}$ to obtain $\dot{\mathbf{x}} = \mathbf{Ax}$ in matrix form, where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -k/m & -c/m \end{bmatrix}.$$

The eigenvalues for \mathbf{A} are

$$\lambda_1 = \frac{-c + \sqrt{c^2 - 4mk}}{2m} \quad \text{and} \quad \lambda_2 = \frac{-c - \sqrt{c^2 - 4mk}}{2m}. \quad (7.13)$$

It's not hard to see that λ_2 is always negative since $\sqrt{c^2 - 4mk}$ is positive. The eigenvalue λ_1 is also negative, because $c^2 - 4mk < c^2$ implies that $\sqrt{c^2 - 4mk} < c$ and then $-c + \sqrt{c^2 - 4mk} < 0$. The numerator of λ_1 is thus negative and so is λ_1 , since $2m > 0$.

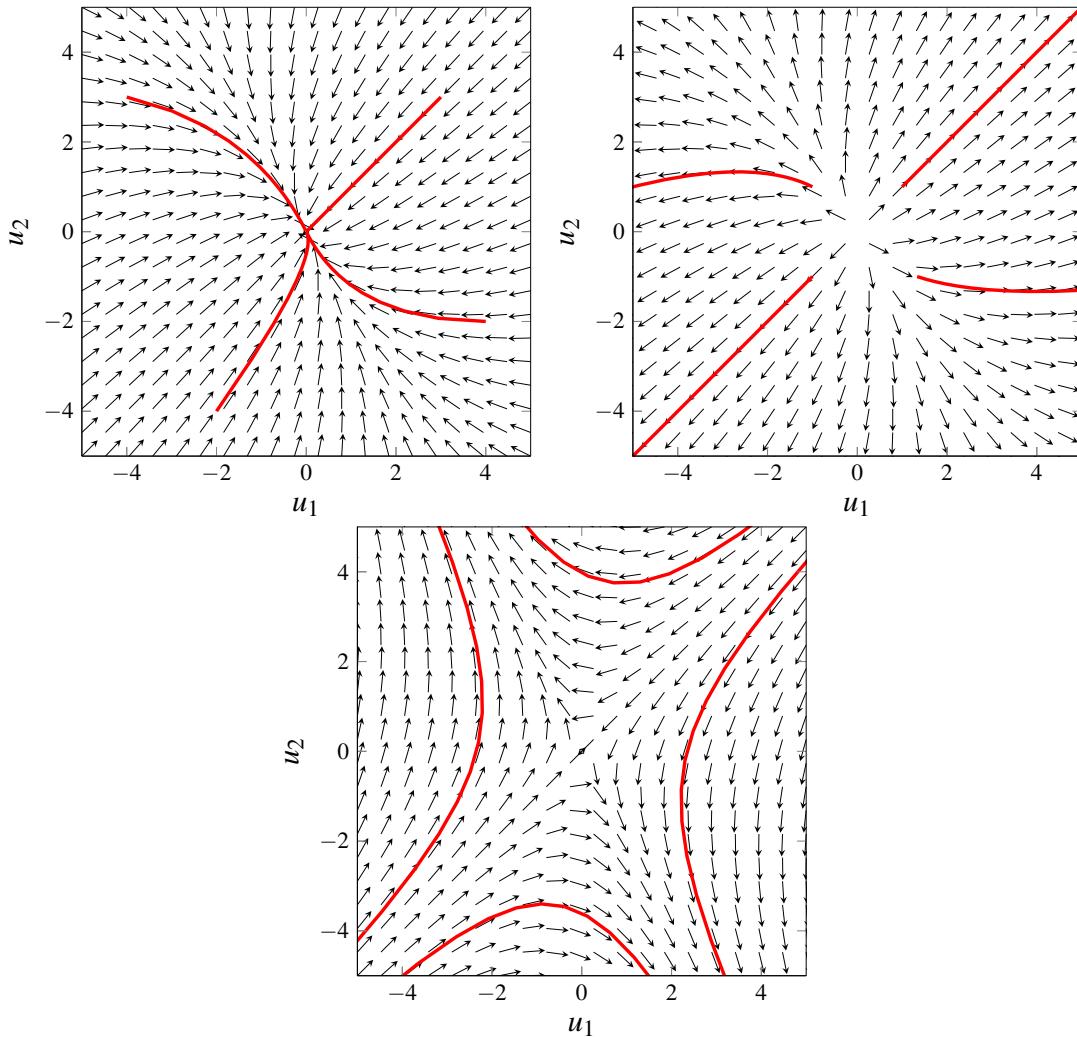


Figure 7.6: Typical direction fields and solution curves for a linear system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ in which \mathbf{A} has real nonzero eigenvalues. Top left panel: both eigenvalues negative (asymptotically stable node). Top right panel: both eigenvalues positive (unstable node). Bottom panel: One positive eigenvalue, one negative eigenvalue (saddle point).

We conclude that the equilibrium solution $x_1 = x_2 = 0$ is always an asymptotically stable node or sink, to which all solutions decay. Of course this means that for any initial data both $u(t)$ and $\dot{u}(t)$ both decay to zero as t increases. The mass approaches a rest state at equilibrium, as expected. Moreover, since $u(t)$ is a superposition of decaying real exponential functions, the motion is not oscillatory. ■

Reading Exercise 7.2.1 Suppose the matrix \mathbf{A} that governs $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ has two equal eigenvalues $\lambda_1 = \lambda_2 = \lambda$ and only one eigenvector \mathbf{v} , so the matrix \mathbf{A} is defective.

- Use (6.31) to argue that if $\lambda < 0$ then all nontrivial solutions decay to the origin as $t \rightarrow \infty$. In this case the origin is called an **asymptotically stable improper node**.
- Use (6.31) to argue that if $\lambda > 0$ then all nontrivial solutions are unbounded as $t \rightarrow \infty$. In this case the origin is called an **unstable improper node**.

The Case in Which A Is Invertible With Complex Eigenvalues

Suppose \mathbf{A} is invertible and has complex eigenvalues. The eigenvalues are necessarily conjugate and distinct, and hence \mathbf{A} has two linearly independent eigenvectors. A general solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ is given by (7.12).

There are three cases to consider, depending on whether the real parts of the eigenvalues for \mathbf{A} are negative, positive, or zero.

- If the eigenvalues are of the form $\lambda = -\alpha \pm \omega i$ with $\alpha > 0$ then the general solution (7.12) can be written as

$$x_h(t) = c_1 e^{-\alpha t} \cos(\omega t) + c_2 e^{-\alpha t} \sin(\omega t).$$

In this case all solutions (7.12) decay to $\langle 0, 0 \rangle$, but spiral infinitely many times around the origin as they do so. A typical direction field in this case is shown in the top left panel of Figure 7.7, with a few solution trajectories. The equilibrium point $\langle 0, 0 \rangle$ here is called an **asymptotically stable spiral point** or **spiral sink**.

- If the eigenvalues are of the form $\lambda = \alpha \pm \omega i$ with $\alpha > 0$ then the general solution (7.12) can be written as

$$x_h(t) = c_1 e^{\alpha t} \cos(\omega t) + c_2 e^{\alpha t} \sin(\omega t).$$

In this case all solutions (7.12) spiral away from $\langle 0, 0 \rangle$ as $t \rightarrow \infty$. A typical direction field in this case is shown in the top right panel of Figure 7.7, with a few solution trajectories. The equilibrium point $\langle 0, 0 \rangle$ here is called an **unstable spiral point** or **spiral source**.

- If the eigenvalues are purely imaginary, of the form $\lambda = \pm \omega i$, then the general solution (7.12) can be written as

$$x_h(t) = c_1 \cos(\omega t) + c_2 \sin(\omega t).$$

In this case all solutions (7.12) form closed elliptical trajectories around $\langle 0, 0 \rangle$. A typical direction field in this case is shown in the bottom panel of Figure 7.7, with a few solution trajectories. The equilibrium point $\langle 0, 0 \rangle$ here is called a **center**.

■ **Example 7.3** Let's revisit Example 7.2, but in the case that the spring-mass ODE $m\ddot{u} + c\dot{u} + ku = 0$ is underdamped, or even undamped, so $c^2 - 4mk < 0$. As in Example 7.2 we formulate this as a system $\dot{\mathbf{x}} = \mathbf{Ax}$; the matrix \mathbf{A} and eigenvalues are unchanged, but now the eigenvalues in (7.13) are complex and can be written as

$$\lambda_1 = \frac{-c}{2m} + i \frac{d}{2m} \quad \text{and} \quad \lambda_2 = \frac{-c}{2m} - i \frac{d}{2m}$$

where $d = \sqrt{4mk - c^2}$ is real, since $4mk - c^2 > 0$. If c is positive then these eigenvalues have negative real part; the origin $x_1 = x_2 = 0$ is a spiral sink, and as $t \rightarrow \infty$ we see that $u(t)$ and $\dot{u}(t)$ both approach zero in an oscillatory manner. If $c = 0$ the origin is a center; $u(t)$ and $\dot{u}(t)$ are periodic and do not decay in amplitude. ■

The Case in Which A Is Singular

Although our primary interest is the case in which \mathbf{A} is invertible, let's briefly look at the possibilities when \mathbf{A} is singular. In this case at least one eigenvalue of \mathbf{A} equals zero. Suppose \mathbf{A} has exactly one zero eigenvalue, say $\lambda_1 = 0$ and $\lambda_2 = \lambda \neq 0$, with eigenvectors \mathbf{v}_1 and \mathbf{v}_2 (note that λ is necessarily real). A general solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ is still given by (7.12) (just with $\lambda_1 = 0$, so $e^{\lambda_1 t} = 1$) and is

$$x_h(t) = c_1 \mathbf{v}_1 + c_2 e^{\lambda t} \mathbf{v}_2.$$

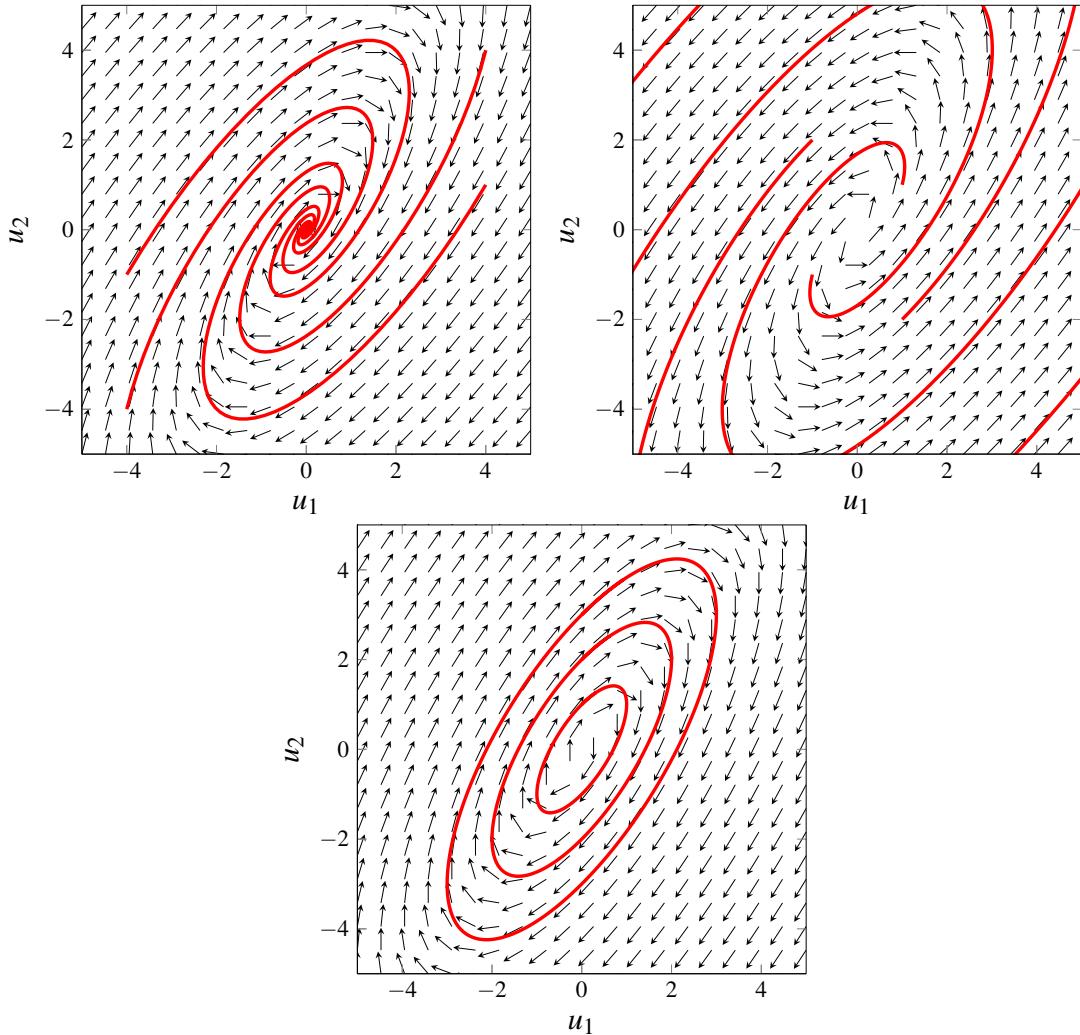


Figure 7.7: Typical direction fields and solution curves for a linear system $\dot{\mathbf{x}} = \mathbf{Ax}$ in which \mathbf{A} has complex conjugate eigenvalues. Top left panel: both eigenvalues have negative real part (asymptotically stable spiral point). Top right panel: both eigenvalues have positive real part (unstable spiral point). Bottom panel: Both eigenvalues are purely imaginary (zero real part), a center.

Every point on the line $\mathbf{x} = c_1\mathbf{v}_1$ with $-\infty < c_1 < \infty$ is a fixed point. If λ is negative then all solutions decay to this line, while if λ is positive all solutions radiate away from this line. The case in which $\lambda < 0$ is illustrated in Figure 7.8, and the line consisting of multiples $c_1\mathbf{v}_1$ is shown as a dashed blue line. The case in which $\lambda > 0$ is similar, just reverse the direction arrows.

If \mathbf{A} has only 0 as a double eigenvalue, there are two possibilities. If there are two linearly independent eigenvectors then \mathbf{A} is the zero matrix, and every point in the phase plane is a fixed point. The direction field would appear as a plane of dots. Finally, if \mathbf{A} has only zero as an eigenvalue and is defective (that is, there is only one eigenvector for $\lambda = 0$), then the analysis of Section 6.2 yields a general solution $\mathbf{x}_h(t)$ given by (6.31), which in this case becomes

$$\mathbf{x}_h(t) = (c_1 + c_2 t)\mathbf{v}_2 + c_2\mathbf{v}_1.$$

The solution curves here are straight lines parallel to the vector \mathbf{v}_2 . A typical direction field is shown in the right panel of Figure 7.8, along with the line spanned by $\mathbf{v}_2 = \langle 1, 2 \rangle$ here.

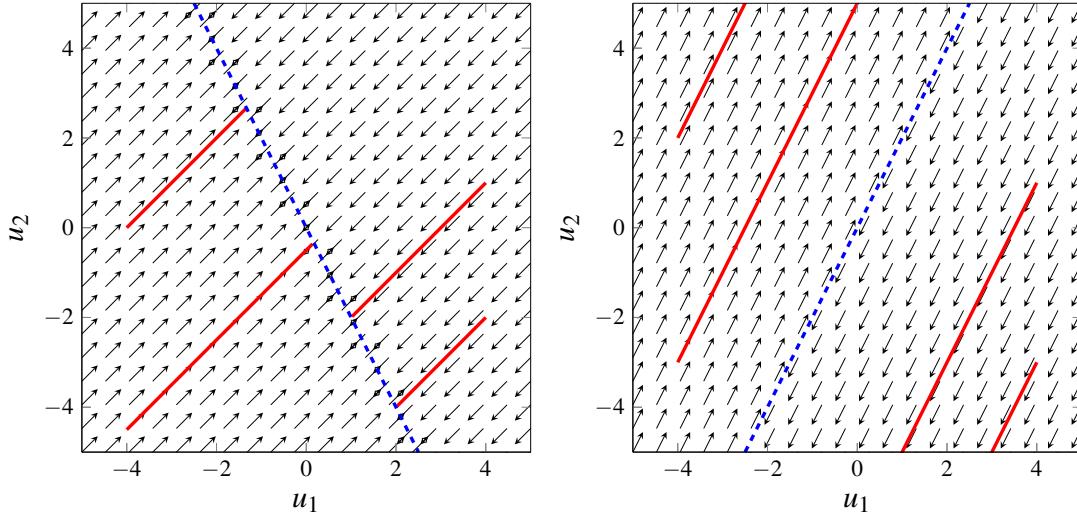


Figure 7.8: Left panel: Direction field for $\dot{\mathbf{x}} = \mathbf{Ax}$ where \mathbf{A} has eigenvalues 0 and λ with $\lambda < 0$. Right panel: Direction field for $\dot{\mathbf{x}} = \mathbf{Ax}$ where \mathbf{A} double defective eigenvalue 0.

7.2.2 Application to the LSD Model

In Examples 7.2 and 7.3 we were able to make conclusions about the behavior of a spring-mass system even in the absence of specific values for the parameters m, c , and k , aside from the physically relevant assumptions that $m, k > 0$ and $c \geq 0$. In fact we can often make conclusions about the qualitative and long-term behavior of solutions to ODEs without choosing specific values for the parameters that appear in the ODEs. In this section we will illustrate this by applying this eigenvalue analysis to the LSD metabolism model from Section 6.1, the pair of linear ODEs (6.1)-(6.2). This analysis can be carried out even if the physical constants k_a, k_b , and k_e are not specified. We'll also show how one can sketch a direction field for that system without specifying these parameters. These techniques will be extended to nonlinear systems in Sections 7.3 and 7.4.

For convenience, we reproduce here the LSD metabolism model from Section 6.1, the linear ODEs (6.1)-(6.2), but with a minor change of notation: we use $x_1 = u_P$ and $x_2 = u_T$. The model is then

$$\dot{x}_1 = -(k_b + k_e)x_1 + k_a x_2 \quad (7.14)$$

$$\dot{x}_2 = k_b x_1 - k_a x_2, \quad (7.15)$$

where we have taken $g(t) = 0$ in (6.1)-(6.2); recall $g(t)$ is the rate at which LSD is administered after the initial dose. Here k_a, k_b , and k_e are all positive constants. In (7.14) and (7.15), $x_1(t)$ is the amount of LSD in the subject's plasma at time t and $x_2(t)$ is the amount of LSD in the subject's tissue at time t . Given the physical nature of the problem we can confine our attention to the first quadrant in the phase plane in which x_1 and x_2 are nonnegative.

Matrix Formulation and Eigenvalue Analysis

The system (6.1)-(6.2) can be formulated as $\dot{\mathbf{x}} = \mathbf{Ax}$, where $\mathbf{x} = \langle x_1, x_2 \rangle$ and

$$\mathbf{A} = \begin{bmatrix} -(k_b + k_e) & k_a \\ k_b & -k_a \end{bmatrix}.$$

It's easy to check that the matrix \mathbf{A} is invertible (for example, it has determinant $k_a k_b > 0$), so the origin $\mathbf{x} = \mathbf{0}$ is the only equilibrium solution for this system.

The stability of the origin is dictated by the eigenvalues of \mathbf{A} . Despite the simplicity of the matrix \mathbf{A} , these eigenvalues are rather complicated. Determining the stability of the origin by direct

examination of these eigenvalues is difficult. However, the trace/determinant analysis of Theorem B.4.1 in Appendix B makes this task straightforward. The determinant and trace of \mathbf{A} are given by

$$\det(\mathbf{A}) = k_a k_b \quad \text{and} \quad \text{tr}(\mathbf{A}) = -(k_a + k_b + k_e).$$

Since each of k_a , k_b , and k_e are positive, we have $\det(\mathbf{A}) > 0$ (noted above) and $\text{tr}(\mathbf{A}) < 0$. Based on the analysis in Section B.4 we can immediately conclude that either both eigenvalues are real and negative, or the eigenvalues are complex-conjugates with negative real part. Either way, the origin is stable. In Exercise 7.2.5 you are asked to show that the eigenvalues are in fact both real, negative, and distinct.

Based on these facts about the eigenvalues of \mathbf{A} , we can conclude that all solutions decay to the origin. Physically, the amount of LSD in both the plasma and tissue decreases to zero for any positive choices of the rate constants k_a , k_b , and k_e , which is quite expected. This eigenvalue analysis lets us make a strong conclusion concerning the long-term behavior of the system, but we can say a bit more and reaffirm this conclusion in another way: graphically. These techniques are simple but powerful, especially when applied to nonlinear systems. But let us begin by illustrating these graphical methods on (7.14)-(7.15), a linear system.

The Nullclines

First, consider (7.14) and suppose we are at a point in the x_1x_2 phase plane where $\dot{x}_1 = 0$. Geometrically, this means that a direction field vector with tail at such a point must have a zero horizontal component and so is vertical. A solution curve passing through such a point has no component of motion in the x_1 -direction, since $\dot{x}_1 = 0$, and so is moving vertically (unless $\dot{x}_2 = 0$ too, in which case we are at an equilibrium point). Physically, this corresponds to a point in the x_1x_2 phase plane where the amount of LSD in the subject's plasma is (momentarily) not changing. From (7.14) the set of (x_1, x_2) points in the phase plane where $\dot{x}_1 = 0$ is described by the equation

$$-(k_b + k_e)x_1 + k_a x_2 = 0, \tag{7.16}$$

or equivalently, by $x_2 = \frac{k_b + k_e}{k_a}x_1$. This curve is called the x_1 **nullcline** for this ODE system.

In this example the x_1 nullcline is a line through $(0,0)$ with positive slope $(k_b + k_e)/k_a$, as illustrated by the curve labeled $\dot{x}_1 = 0$ in the left panel of Figure 7.9. The short vertical tick marks on the nullcline are there to indicate that any direction field vector with its tail on this line must be oriented vertically, and so solutions move vertically as they cross this nullcline.

Reading Exercise 7.2.2 Show that if $x_2 > \frac{k_b + k_e}{k_a}x_1$ (so we're above the x_1 nullcline in the left panel of Figure 7.9) then $-(k_b + k_e)x_1 + k_a x_2 > 0$. Show that if $x_2 < \frac{k_b + k_e}{k_a}x_1$ (below the \dot{x}_1 nullcline) then $-(k_b + k_e)x_1 + k_a x_2 > 0$. Then use (7.14) to conclude that $\dot{x}_1 > 0$ (solutions move generally to the right) above the \dot{x}_1 nullcline and $\dot{x}_1 < 0$ (solutions move generally to the left) below the x_1 nullcline. Physically this means that if $x_2 > \frac{k_b + k_e}{k_a}x_1$ at some instant then $\dot{x}_1 > 0$ and the amount of LSD in the plasma is increasing, while if $x_2 < \frac{k_b + k_e}{k_a}x_1$ then $\dot{x}_1 < 0$ and the amount of LSD in the plasma is decreasing.

We can determine the x_1 nullcline and how solutions behave on either side of this nullcline without specific values for the constants k_a , k_b , and k_e ; all we need to know is that these constants are positive.

We can perform the same analysis for (7.15). Specifically, $\dot{x}_2 = 0$ at those points in the phase plane where

$$k_b x_1 - k_a x_2 = 0 \tag{7.17}$$

or equivalently $x_2 = \frac{k_b}{k_a}x_1$. Equation (7.17) describes the x_2 **nullcline** for this system, and this nullcline is a line through the origin with slope k_b/k_a , illustrated as the line labeled $\dot{x}_2 = 0$ in

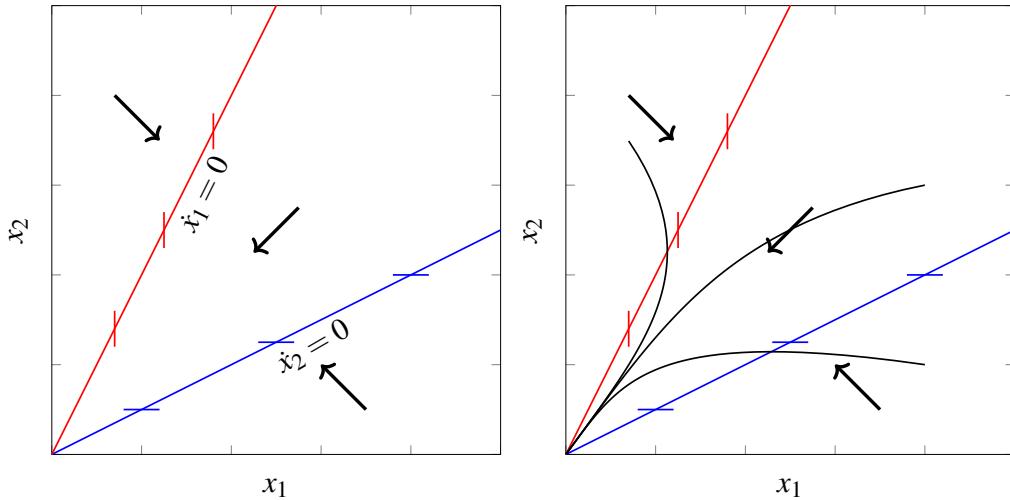


Figure 7.9: Left panel: The x_1 nullcline (labeled $\dot{x}_1 = 0$, red) and x_2 nullcline (labeled $\dot{x}_2 = 0$, blue) for the LSD metabolism system (7.16)-(7.17), with direction arrows in each region. Right panel: Some representative solution curves that follows the direction arrows and nullclines. All solutions converge to the origin.

the left panel of Figure 7.9. The short horizontal tick marks on the nullcline indicate that the direction field vectors with tail on this nullcline are horizontal, and solution curves crossing the x_2 nullcline must move horizontally. Physically, this indicates those points in the phase plane where the amount of LSD in the subject's tissue is not changing, at least momentarily. If $x_2 < k_b x_1 / k_a$ then $\dot{x}_2 = k_b x_1 - k_a x_2 > 0$ (solutions below the x_2 nullcline move generally upward, corresponding to increasing LSD concentration in the tissue) while if $x_2 > k_b x_1 / k_a$ then $\dot{x}_2 < 0$; here solutions move downward, and tissue LSD concentration is decreasing. We can also say definitively that the x_2 nullcline has a slope less than that of the x_1 nullcline, since $k_b/k_a < (k_b + k_e)/k_a$ (because $k_e > 0$) and this is reflected in the qualitatively correct graphs in the left panel of Figure 7.9. No assumptions about the scale on the x_1 or x_2 axes is needed to sketch these nullclines.

Sketching the Phase Portrait

Based on the above analysis we can make a few conclusions about the direction field for (7.14)-(7.15) that allows us to sketch a qualitatively correct direction field and deduce how solutions behave without solving the ODEs and without assuming anything about the constants k_a, k_b , and k_e . We have deduced

1. The direction field arrows with tail on the x_1 nullcline (given by (7.16)) satisfy $\dot{x}_1 = 0$ at that point, and so are vertically oriented. Solution curves that cross the x_1 nullcline do so vertically.
2. Direction field arrows based at points not on the x_1 nullcline satisfy $\dot{x}_1 > 0$ or $\dot{x}_1 < 0$. As a result solutions passing through such points are moving generally to the right ($\dot{x}_1 > 0$) or to the left ($\dot{x}_1 < 0$). This says nothing about their motion in the x_2 direction, though.
3. The direction field arrows with tail on the x_2 nullcline (given by (7.17)) satisfy $\dot{x}_2 = 0$ at that point, and so are horizontally oriented. Solution curves that cross the x_2 nullcline do so horizontally.
4. Direction field arrows based at points not on the x_2 nullcline satisfy $\dot{x}_2 > 0$ or $\dot{x}_2 < 0$. As a result solutions passing through such points are moving generally up ($\dot{x}_2 > 0$) or down ($\dot{x}_2 < 0$). This says nothing about their motion in the x_1 direction, though.

Each of these facts is represented in the left panel of Figure 7.9. In particular, both the x_1

nullcline and the x_2 nullcline are shown, with vertical or horizontal tick marks, respectively, to indicate the nature of the corresponding nullcline. Between the nullclines solutions move with either $\dot{x}_1 > 0$ (right) or $\dot{x}_1 < 0$ (left) and either $\dot{x}_2 > 0$ (up) or $\dot{x}_2 < 0$ (down). This is indicated by the arrows in each region, to provide a rough indication of whether solutions are moving generally up and right, up and left, down and right, or down and left. The left panel of Figure 7.9 is a supremely economical direction field that shows in which direction the solution curves move in the phase plane, in this case using only three arrows.

Sketching Solution Curves

By using the left panel of Figure 7.9, we can sketch plausible solution trajectories, with the nullclines and the few arrows as a guide. See the right panel in Figure 7.9. Such solution curves can easily be sketched by hand. You should visually check that these curves obey the direction arrows and cross the nullclines with the proper behavior, and perhaps sketch a few such curves yourself. Based on this graphical analysis, the inescapable conclusion is that all solutions converge to $(x_1, x_2) = (0, 0)$ or, in the original dependent variables, the point $(u_P, u_T) = (0, 0)$. In plain English, the amount of LSD in the plasma and tissue decays to zero over time, which seems entirely reasonable. This is in accord with the eigenvalue analysis.

The right panel in Figure 7.9 is called a **phase portrait** for the ODE system. In such a phase portrait we indicate the behavior of solutions to a system of ODEs by drawing direction arrows (whether manually or with a computer) and then show typical solution trajectories. We may also indicate fixed points and their stability. Compare to the one-dimensional analog in Figure 2.6.

7.2.3 The Equation $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}$

For the analysis of $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}$ we assume that the matrix \mathbf{A} is invertible, or equivalently that all eigenvalues for \mathbf{A} are nonzero. This will be the primary interest in the next section. When \mathbf{A} is invertible the unique equilibrium solution \mathbf{x}_p for the autonomous nonhomogeneous equation is obtained by solving $\mathbf{A}\mathbf{x}_p + \mathbf{b} = \mathbf{0}$ for \mathbf{x}_p and is

$$\mathbf{x}_p = -\mathbf{A}^{-1}\mathbf{b}. \quad (7.18)$$

A general solution to $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}$ is then as in (6.39) and is $\mathbf{x}(t) = \mathbf{x}_p + \mathbf{x}_h(t)$ or

$$\mathbf{x}(t) = -\mathbf{A}^{-1}\mathbf{b} + \mathbf{x}_h(t), \quad (7.19)$$

where $\mathbf{x}_h(t)$ is a general solution to the homogeneous system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$. The analysis for the homogeneous case based on (7.12), in conjunction with (7.19), shows that the direction field for the nonhomogeneous system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}$ depends on the eigenvalues for \mathbf{A} in a manner that parallels the homogeneous case. The only difference is that the unique equilibrium solution $\mathbf{x}_p = \mathbf{0}$ for the homogeneous system is replaced by $\mathbf{x}_p = -\mathbf{A}^{-1}\mathbf{b}$ in (7.18). Thus, for example, if the eigenvalues for \mathbf{A} are both negative (or have negative real part), solutions decay to \mathbf{x}_p .

7.2.4 Direction Fields for Larger Systems of ODEs

The fundamental idea behind direction fields for autonomous systems of ODEs works, in principle, for systems of any size. Specifically, because $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, the vector

$$\mathbf{f}(x_1, \dots, x_n) = \begin{bmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_n(x_1, \dots, x_n) \end{bmatrix}$$

indicates the direction that a solution to (7.1) moves at any point (x_1, \dots, x_n) in n -dimensional phase space. In the linear homogeneous case the function $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, and so the direction depends on the

matrix \mathbf{A} . The difficulty, of course, is visualizing the situation. Even in three dimensions when we use technology to draw a direction field, the result is usually an unintelligible cloud of arrows.

Still, some conclusions can be drawn. In the linear case, if the eigenvectors of \mathbf{A} span \mathbb{R}^n (that is, if \mathbf{A} is diagonalizable as discussed in Appendix B), then a general solution $\mathbf{x}_h(t)$ to the homogeneous system $\dot{\mathbf{x}} = \mathbf{Ax}$ was derived in (6.17) and is

$$\mathbf{x}_h(t) = \sum_{k=1}^n c_k e^{\lambda_k t} \mathbf{v}_k$$

where the \mathbf{v}_k are the eigenvectors for \mathbf{A} and the λ_k are the corresponding eigenvalues. We can make a couple of easy conclusions:

- If all λ_k have negative real part then all solutions $\mathbf{x}_h(t)$ decay to the origin as t increases.
- If any λ_k has positive real part then typical solutions grow without bound as t increases, unless the corresponding c_k just happens to be zero (this would depend on the initial condition).

However, the case in which all $\lambda_k \leq 0$ with some equal to zero (and possibly with some defective eigenvalues, so \mathbf{A} is not diagonalizable) is more subtle and we won't pursue this analysis here. See [56] for more information.

7.2.5 Exercises

Exercise 7.2.1 Each pair of ODEs below is of the form

$$\begin{aligned}\dot{x}_1 &= A_{1,1}x_1 + A_{1,2}x_2 \\ \dot{x}_2 &= A_{2,1}x_2 + A_{2,2}x_2\end{aligned}$$

or $\dot{\mathbf{x}} = \mathbf{Ax}$ for the appropriate matrix \mathbf{A} . Use technology to sketch a direction field for the system on the range $-3 \leq x_1, x_2 \leq 3$. Based on what you see, deduce what you can about the eigenvalues of \mathbf{A} (refer to Figures 7.6, 7.7, and 7.8.) Then confirm your conclusions by computing the eigenvalues of \mathbf{A} .

- (a) $\mathbf{A} = \begin{bmatrix} -3 & 1 \\ 1 & -3 \end{bmatrix}$
- (b) $\mathbf{A} = \begin{bmatrix} -1 & -2 \\ -4 & 2 \end{bmatrix}$
- (c) $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$
- (d) $\mathbf{A} = \begin{bmatrix} -1 & 1 \\ -2 & 1 \end{bmatrix}$
- (e) $\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 4 & -3 \end{bmatrix}$
- (f) $\mathbf{A} = \begin{bmatrix} 3 & -2 \\ 4 & -1 \end{bmatrix}$

Exercise 7.2.2 Apply the technique used on (7.14)-(7.15) in Section 7.2.2 to sketch phase portraits and solution trajectories for the systems $\dot{\mathbf{x}} = \mathbf{Ax}$ below with the given initial conditions. Each system is linear. Limit your sketch to the range $-5 \leq x, y \leq 5$. In particular, for each system do the following:

- Find the nullclines $\dot{x} = 0$ and $\dot{y} = 0$; they should be straight lines through the origin.
- The nullclines here should divide the plane into four regions. In each region decide

whether solutions move generally up and left, up and right, down and left, or down and right, and then sketch an appropriate arrow.

- Based on your picture, sketch solution trajectories with the given initial conditions; make sure to follow the arrows, and cross the x nullcline vertically, the y nullcline horizontally. What is the long-term fate of each solution?
- Confirm your work by analytically solving the system. Pay attention to the correspondence between solution behavior and the eigenvalues of the matrix \mathbf{A} .

- $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 12 & -1 \end{bmatrix}$, initial conditions $x(0) = 2, y(0) = -2$ and $x(0) = 1, y(0) = -4$.
- $\mathbf{A} = \begin{bmatrix} 1 & -1 \\ 6 & -4 \end{bmatrix}$, initial conditions $x(0) = 2, y(0) = -2$ and $x(0) = -2, y(0) = 3$.
- $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$, initial conditions $x(0) = 1, y(0) = -1$ and $x(0) = 0, y(0) = 3$.
- $\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 4 & -3 \end{bmatrix}$, initial conditions $x(0) = 1, y(0) = -1$ and $x(0) = 1, y(0) = -2$.
- $\mathbf{A} = \begin{bmatrix} 3 & -2 \\ 4 & -1 \end{bmatrix}$, initial conditions $x(0) = 1, y(0) = -1$ and $x(0) = 1, y(0) = -2$.

Exercise 7.2.3 Consider the spring-mass-damper system $m\ddot{x} + c\dot{x} + kx = 0$.

- Show this system can be converted to the equations $\dot{x} = y$ and $\dot{y} = -kx/m - cy/m$ by letting $y = \dot{x}$.
- Show that the x nullcline is the x axis in the phase plane, and the y nullcline is the line $y = -kx/c$.
- Sketch both nullclines in the plane (note $k/c > 0$), in the manner of Figure 7.9 (left panel), in some region around the origin. Assume that $k/c = 1$; you don't need to assume any specific value for m , merely that m is positive. The nullclines should divide the plane into 4 regions; sketch direction arrows in each (up and left, up and right, down and left, or down and right).
- Use part (c) to sketch some solution trajectories. Does your sketch conform to how a damped spring-mass system should behave?
- Repeat parts (c)-(d) assuming c is close to zero, so k/c is large. Does your sketch make sense?
- Repeat parts (c)-(d) assuming c is very large, so k/c is close to zero. Does your sketch make sense?
- Repeat parts (c)-(d) assuming $c = 0$ (so the ODE is $m\ddot{x} + kx = 0$). What does the y nullcline become? Does your sketch and solution curves reflect how an undamped system should behave?

Exercise 7.2.4 In Exercise 6.1.6 a double salt tank was presented, and we derived the equations

$$\begin{aligned}\dot{x}_1 &= \frac{1}{2} - \frac{x_1}{120} + \frac{x_2}{100} \\ \dot{x}_2 &= \frac{x_1}{120} - \frac{x_2}{50}\end{aligned}$$

for the amounts $x_1(t)$ and $x_2(t)$ of salt in each tank.

- Show that the x_1 nullcline here is the line $x_2 = -50 + 5x_1/6$ in the x_1x_2 phase plane. Also

show that the x_2 nullcline here is the line $x_2 = 5x_1/12$ in the x_1x_2 phase plane. Sketch both of these lines on a pair of x_1x_2 axes, on the range $0 \leq x_1, x_2 \leq 200$. (Only the range $x_1, x_2 \geq 0$ is physically relevant here.)

- Find the point where the nullclines intersect, and show that this is an equilibrium solution for the system.
- The nullclines should divide the first quadrant into four regions. Sketch direction arrows in each (up/left, up/right, down/left, down/right), and then sketch some plausible solution trajectories. What do solutions seem to do as t increases to infinity?
- Find the general solution for the system analytically and reconcile this solution with your sketch.

Exercise 7.2.5 Show that the matrix

$$\mathbf{A} = \begin{bmatrix} -(k_b + k_e) & k_a \\ k_b & -k_a \end{bmatrix}$$

that governs the system (6.1)-(6.2) with $g = 0$ (or (7.14)-(7.15)) has negative real eigenvalues for any choice of $k_a, k_b, k_e > 0$, as strongly suggested by Figure 7.9, by following these steps and using the analysis of Section B.4:

- Let $D = \det(\mathbf{A})$ and $T = \text{tr}(\mathbf{A})$. Show that we can write

$$T^2/4 - D = \frac{(k_a - k_e)^2}{4} + \frac{k_b(2k_a + k_b + 2k_e)}{4}.$$

Why does this imply that $T^2/4 - D > 0$?

- Use part (a) and the results in Section B.4 to conclude that the eigenvalues of \mathbf{A} are negative and real.

7.3 Autonomous Nonlinear Systems and Phase Portraits

The approach of graphing nullclines, determining solution directions, then sketching solutions that we used in the last section is applicable to nonlinear systems of ODEs as well. In this section we will refine and extend those techniques and use them to determine the behavior of solutions to the competing species model (7.3)-(7.4). This will illustrate the power of this approach for analyzing systems of ODEs. The methods are of general applicability for two-dimensional systems, and some of these techniques extend to higher-dimensional systems.

7.3.1 Sketching Phase Portraits for Nonlinear Systems

The Struggle for Existence Continues

In Section 7.1 we encountered a population model for two competing species of yeast growing in a single vessel, a coupled system (7.3)-(7.4) of ODEs. Let us revisit the specific case in which the parameters are $r_1 = 1$, $K_1 = 2$, $r_2 = 2$, $K_2 = 3$, $a = 0.2$, and $b = 0.45$. In this case the system becomes $\dot{u}_1 = f_1(u_1, u_2)$, $\dot{u}_2 = f_2(u_1, u_2)$ with

$$f_1(u_1, u_2) = u_1(1 - u_1/2) - 0.1u_1u_2 \tag{7.20}$$

$$f_2(u_1, u_2) = 2u_2(1 - u_2/3) - 0.3u_1u_2. \tag{7.21}$$

(These were given in (7.11).) A direction field for this system as drawn by a computer was shown in Figure 7.4. Let's now reproduce something like Figure 7.4 that will allow us to determine how

solutions behave and how both populations change over time, but without the aid of a computer. This will also be a steppingstone to determining how solutions to (7.3)-(7.4) behave for any choice of the parameters r_1, K_1, r_2, K_2, a , and b , a task for which a computer isn't terribly helpful.

Our ultimate goal is a sketch that illustrates how solutions behave in the $u_1 u_2$ phase plane, which can be used to make conclusions about how the two populations are changing. Some parameter values allow the populations to coexist, others doom one or the other species to extinction. We begin by sketching the nullclines for the system.

The u_1 Nullcline

The first step is to determine and sketch the u_1 nullcline, the set of points in the phase plane at which $\dot{u}_1 = 0$. Since $\dot{u}_1 = f_1(u_1, u_2)$ with $f_1(u_1, u_2)$ as in (7.20) this leads to the equation

$$u_1(1 - u_1/2) - 0.1u_1u_2 = 0.$$

This equation is equivalent to $u_1(1 - u_1/2 - 0.1u_2) = 0$, which means that either $u_1 = 0$ or $1 - u_1/2 - 0.1u_2 = 0$. The u_1 nullcline thus consists of two pieces, the vertical line $u_1 = 0$ (the u_2 axis) and the line $1 - u_1/2 - 0.1u_2 = 0$, or equivalently $u_2 = 10 - 5u_1$. This nullcline is shown in the left panel of Figure 7.10, plotted on the range $0 \leq u_1, u_2 \leq 10$ to show the entire relevant portion of the nullcline. Any solution to (7.20)-(7.21) that crosses or touches this nullcline must move with $\dot{u}_1 = 0$, that is, vertically in the $u_1 u_2$ phase plane. This is indicated with a few vertical tick marks on the u_1 nullcline, although there's little point in drawing these tick marks on that portion of the nullcline that coincides with the u_2 axis. Physically, if the populations (u_1, u_2) are on the u_1 nullcline, the u_1 population is not changing, at least momentarily.

Reading Exercise 7.3.1 In the left panel of Figure 7.10 it appears that if $u_1(t_0) = 0$ and $u_2(t_0) > 0$ (the solution is on the positive u_2 axis at time $t = t_0$) then the solution to (7.20)-(7.21) with this data moves vertically, and so remains on the u_2 axis, and hence $u_1(t) = 0$ for all t . Show this is the case, by showing that if $u_1(t) = 0$ for all t and $u_2(t)$ obeys the logistic equation $\dot{u}_2 = 2u_2(1 - u_2/3)$ with $u_2(t_0) = u_0 > 0$ then these functions provide a solution to (7.20)-(7.21) with data $u_1(t_0) = 0$, $u_2(t_0) = u_0$. Provide a physical interpretation of this situation.

The u_1 nullcline in the left panel of Figure 7.10 divides the first quadrant into two distinct pieces, one of which lies below the line $1 - u_1/2 - 0.1u_2 = 0$, and one above. If a solution is not on the nullcline $\dot{u}_1 = 0$ then either $\dot{u}_1 < 0$ or $\dot{u}_1 > 0$. This means the solution is moving generally to the left or to the right, or in terms of the physical system itself, the u_1 species population is decreasing or increasing, respectively. This is indicated in each region by drawing an arrow that points left or right, as in the left panel of Figure 7.10. Whether $\dot{u}_1 < 0$ or $\dot{u}_1 > 0$ in each region can be determined by choosing a test point in each region. For example, the point $(u_1, u_2) = (1, 2)$ lies below $1 - u_1/2 - 0.1u_2 = 0$, and here $f_1(1, 2) = 0.3 > 0$, so solutions in this region move generally to the right. Similarly $(u_1, u_2) = (3, 3)$ lies above $1 - u_1/2 - 0.1u_2 = 0$, and here $f_1(3, 3) = -2.4 < 0$, so solutions in this region move generally to the left. This says nothing about the motion of the solution in the u_2 direction. This will be determined using the $\dot{u}_2 = 0$ nullcline.

Reading Exercise 7.3.2 Consider the left panel in Figure 7.10. If a solution $\langle u_1(t), u_2(t) \rangle$ passes through the point $u_1 = 3, u_2 = 2$, what does \dot{u}_1 equal? Is the u_1 population increasing or decreasing?

The u_2 Nullcline

A similar analysis is used to sketch the u_2 nullcline. Setting $f_2(u_1, u_2) = 0$ yields $2u_2(1 - u_2/3) - 0.3u_1u_2 = 0$.

$$u_2(2 - 2u_2/3 - 0.3u_1) = 0.$$

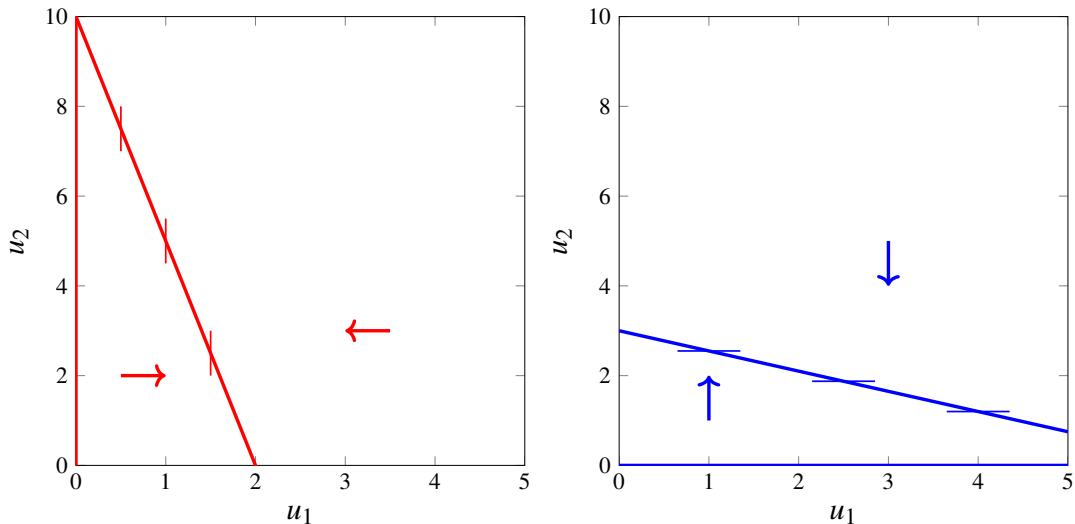


Figure 7.10: Left panel: sketch of u_1 nullcline for the competing species system (7.20)-(7.21). Right panel: sketch of u_2 nullcline for the competing species system (7.20)-(7.21).

The points in the phase plane that satisfy this equation consist of the horizontal line $u_2 = 0$ and the line $2 - 2u_2/3 - 0.3u_1 = 0$. These two pieces of the nullcline are shown in the right panel of Figure 7.10. Any solution curve that touches this nullcline does so with $\dot{u}_2 = 0$, and so is moving horizontally. In terms of the populations, the second species population is momentarily not changing at such a point. The nullcline divides the area of interest in the phase plane into two regions, and in each region either $\dot{u}_2 = f_2(u_1, u_2) < 0$ or $\dot{u}_2 = f_2(u_1, u_2) > 0$, so the u_2 population is either decreasing or increasing. To figure out which, substitute a test point into $f_2(u_1, u_2)$. For example, $f_2(1, 1) = 1.03 > 0$, and so we draw an arrow with tail at $(1, 1)$ pointing upward (the length is irrelevant). Also, $f_2(3, 5) = -11.17 < 0$, so draw an arrow with tail at $(3, 5)$ pointing downward.

Reading Exercise 7.3.3 Suppose that at some instant in time $u_1 = 2$ and $u_2 = 4$. Use the left and right panels in Figure 7.10 to determine if \dot{u}_1 positive, negative, or zero. Is \dot{u}_2 positive, negative, or zero? What is each population doing at this point—increasing, decreasing, or remaining constant?

Reading Exercise 7.3.4 In the right panel of Figure 7.10 it appears that if $u_2(t_0) = 0$ and $u_1(t_0) > 0$ (the solution is on the positive u_1 axis at time $t = t_0$) then the solution to (7.20)-(7.21) with this data moves horizontally, and so remains on the u_1 axis, and hence $u_2(t) = 0$ for all t . Show this is the case, by showing that if $u_2(t) = 0$ for all t and $u_1(t)$ obeys the logistic equation $\dot{u}_1 = u_1(1 - u_1/2)$ with $u_1(t_0) = u_0$ then these functions provide a solution to (7.20)-(7.21) with data $u_1(t_0) = u_0, u_2(t_0) = 0$. What is the physical interpretation of this situation?

Reading Exercise 7.3.5 Use Reading Exercises 7.3.1 and 7.3.4 in conjunction with the existence-uniqueness theorem (Theorem 6.1.1) to argue that solutions to (7.20)-(7.21) that start with initial data $u_1(t_0) \geq 0, u_2(t_0) \geq 0$ can never leave the first quadrant of the phase plane. Thus in this model if the populations start with nonnegative initial data then the populations never assume negative values, a reassuring feature for a model.

Putting It All Together

The same procedure that led to Figure 7.9 can be used here. In the left panel of Figure 7.11 we show the nullclines superimposed, with short vertical or horizontal tick marks to indicate the direction solutions travel as they cross the nullclines. The nullclines divide the first quadrant into a number

of distinct regions and in each region we use the nullcline figures to sketch arrows that point up and left, up and right, down and left, or down and right, to indicate the general motion of solutions.

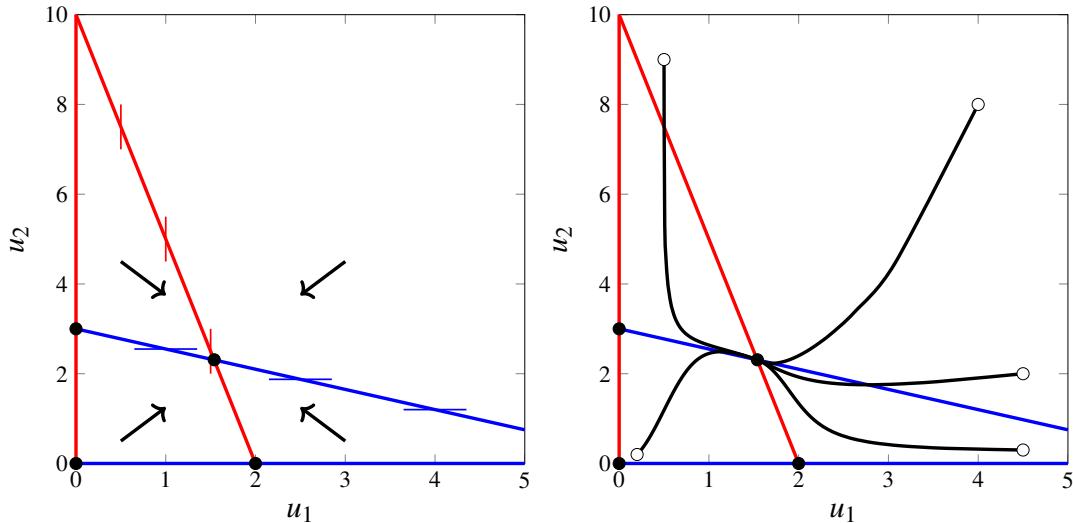


Figure 7.11: Left panel: nullclines and direction arrows for the competing species system (7.20)-(7.21), with equilibrium solutions shown as black dots. Right panel: typical solution trajectories, with equilibrium solutions shown as black dots, initial conditions as circles.

Observe that in the left panel of Figure 7.11 that there are four fixed points or equilibrium solutions where the \dot{u}_1 and \dot{u}_2 nullclines intersect (points where both $\dot{u}_1 = 0$ and $\dot{u}_2 = 0$): $(u_1, u_2) = (0, 0), (2, 0), (0, 3)$, and approximately $(1.54, 2.31)$. These points are highlighted with black dots and can be found by solving the equations $f_1(u_1, u_2) = 0$ and $f_2(u_1, u_2) = 0$ simultaneously, where f_1 and f_2 are given by (7.20) and (7.21).

Reading Exercise 7.3.6 What is the physical interpretation of each fixed point?

The Behavior of Solutions and Physical Interpretation

Using the arrows in the left panel of Figure 7.11 and the nullclines as a guide, we can sketch solution trajectories for (7.20)-(7.21) from any initial point. Although the curves in the text were drawn using a computer, we can easily do this by hand to produce qualitatively correct solutions. A few typical solution curves one might draw are shown in the right panel of Figure 7.11; the start of each trajectory is marked with a circle. In each case the trajectory follows the general up and left, up and right, down and left, or down and right directions, as indicated in the left panel. Moreover, when a solution crosses a nullcline it does so with the proper vertical or horizontal motion. For example, the solution that starts near $(u_1, u_2) = (0.5, 9)$ crosses the u_1 nullcline vertically. The sketch in the right panel of Figure 7.11 is a **phase portrait** for the system (7.20)-(7.21). The nullclines, direction arrows, and representative solution trajectories illustrate the behavior of solutions in a manner similar to that of a direction field, but without the brute force approach of drawing a lot of arrows.

In Figure 7.11 it is clear that all solutions approach the fixed point at $u_1 \approx 1.54, u_2 \approx 2.31$ as t increases. This fixed point appears to be asymptotically stable, a term we will define more precisely shortly. In a nutshell, any solution that starts close to this fixed point approaches it. For this ODE system it appears that all solutions approach this fixed point, whether they start nearby or not. The fixed point at $(0, 0)$ appears to be unstable, for solutions that start near (but not at) $(0, 0)$ move away from $(0, 0)$. The fixed points at $(2, 0)$ and $(0, 3)$ are slightly more subtle, but a little experimentation with drawing trajectories should convince you that although solutions may initially approach these fixed points, solutions typically veer away toward the stable fixed point at $(1.54, 2.31)$.

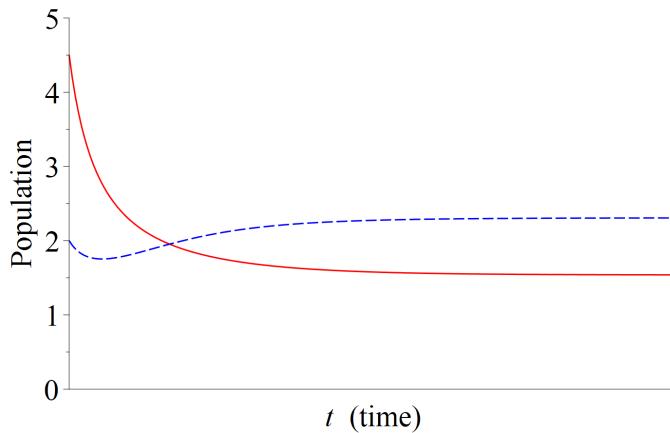


Figure 7.12: Species 1 population $u_1(t)$ (red, solid) and species 2 population $u_2(t)$ (blue, dashed) over time for competing species system (7.20)-(7.21) with initial populations $u_1(0) = 4.5, u_2(0) = 2$ and parameters $r_1 = 1, K_1 = 2, r_2 = 2, K_2 = 3, a = 0.2$, and $b = 0.45$.

This analysis supports an important conclusion regarding the physical behavior of the system (7.3)-(7.4) with the parameters $r_1 = 1, K_1 = 2, r_2 = 2, K_2 = 3, a = 0.2$, and $b = 0.45$: For any nonzero starting populations the species will approach stable coexistence.

The phase portrait can also be used to sketch qualitatively correct graphs of the solution components $u_1(t)$ and $u_2(t)$ as functions of time t . Consider, for example, the solution trajectory that starts with initial data $u_1(0) = 4.5, u_2(0) = 2$; this solution trajectory is shown in the right panel of Figure 7.11. Based on this curve we can sketch $u_1(t)$ and $u_2(t)$ individually as functions of time t . We see that from $u_1(0) = 4.5$ the function $u_1(t)$ decreases and converges to $u_1 \approx 1.54$. The function u_2 starts at $u_2(0) = 2$, initially decreases, then increases to $u_2 \approx 2.31$. We see that if $u_1(0) = 4.5$ and $u_2(0) = 2$, the u_1 population will decrease steadily to $u_1 \approx 1.54$; the u_2 population will initially decrease a bit, then increase to $u_2 \approx 2.31$. These graphs are shown in Figure 7.12.

Reading Exercise 7.3.7 Based on the direction arrows in the left panel and phase portrait in the right panel of Figure 7.11, sketch graphs of $u_1(t)$ and $u_2(t)$ if the initial populations are

- (a) $u_1(0) = 4, u_2(0) = 8$.
- (b) $u_1(0) = 1, u_2(0) = 8$.
- (c) $u_1(0) = 0.5, u_2(0) = 0.5$.

Changing the Parameters

The parameters in a system of ODEs play a key role in determining the nature of the system's behavior. Consider the phase portrait in Figure 7.13. This is the same system (7.3)-(7.4) of ODEs that led to (7.20)-(7.21), but now the competition parameters a and b have been increased to $a = b = 3$; we retain the values $r_1 = 1, K_1 = 2, r_2 = 2$, and $K_2 = 3$.

The nullclines still divide the first quadrant in the u_1u_2 phase plane into four regions with equilibrium solutions as indicated by the black dots, but comparison to Figure 7.11 shows that the direction arrows in each region have changed, and solutions now behave quite differently. The fixed point with $u_1, u_2 > 0$ doesn't appear to attract nearby solutions, although it seems that the fixed points $(2, 0)$ and $(3, 0)$ (whose locations remain unchanged) now do attract nearby solutions. The origin clearly remains unstable. It appears now that mutual coexistence is not an option, but the existence of one species and the extinction of the other (the points $(2, 0)$ and $(0, 3)$) are stable.

Why did the stability of the fixed points change, especially the fixed point with $u_1, u_2 > 0$ that represents coexistence? Is it because the competition is more fierce? If so, what combinations of a and b doom one species or the other to extinction? What if a and b are both large? How do the

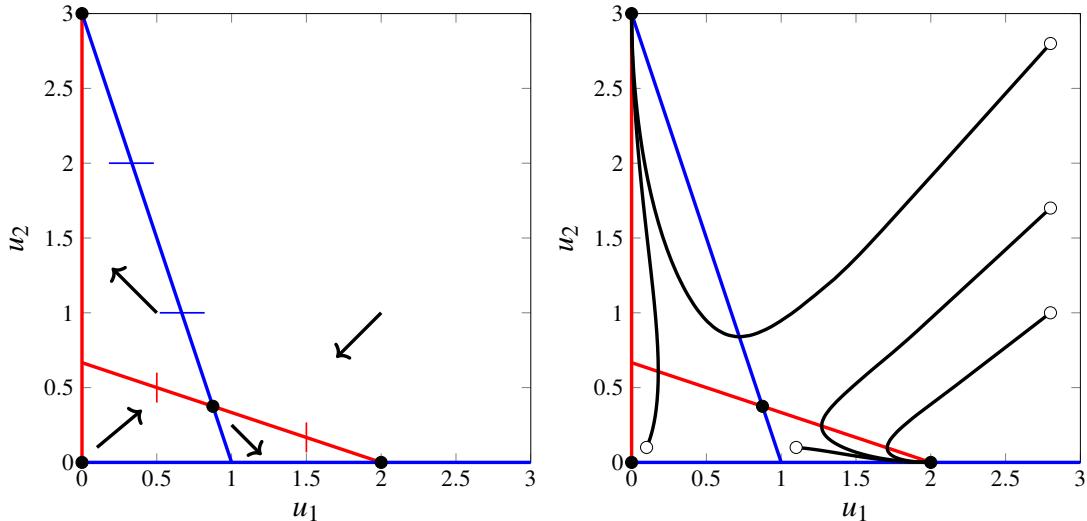


Figure 7.13: Left panel: nullclines and direction arrows for the competing species system (7.3)-(7.4) with $r_1 = 1$, $K_1 = 2$, $r_2 = 2$, $K_2 = 3$, $a = 3$, and $b = 3$; equilibrium solutions shown as black dots, initial conditions as circles. Right panel: typical solution trajectories.

other parameters r_1 , r_2 , K_1 , and K_2 factor into this?

Stability

Before we proceed it will be helpful to clarify various notions of stability for fixed points. These parallel the definitions that were given for scalar ODEs in Section 2.3.4 and Definition 2.3.3.

Definition 7.3.1 — Stability. Suppose \mathbf{x}^* is a fixed point (equilibrium solution) to an autonomous system of ODEs of the form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ with initial data $\mathbf{x}(t_0) = \mathbf{x}_0$.

- The fixed point \mathbf{x}^* is **stable** if all solutions that start sufficiently close to \mathbf{x}^* stay close to \mathbf{x}^* , although these solutions need not approach \mathbf{x}^* . More precisely, the fixed point \mathbf{x}^* is stable if for each $\varepsilon > 0$ there is some real number $\delta > 0$ so that if $|\mathbf{x}_0 - \mathbf{x}^*| < \delta$ then $|\mathbf{x}(t) - \mathbf{x}^*| < \varepsilon$ for all $t > t_0$.
- The fixed point \mathbf{x}^* is **asymptotically stable** if it is stable and all solutions $\mathbf{x}(t)$ that start sufficiently close to \mathbf{x}^* approach \mathbf{x}^* , that is, $\lim_{t \rightarrow \infty} \mathbf{x}(t) = \mathbf{x}^*$. More precisely, the fixed point \mathbf{x}^* is asymptotically stable if it is stable and for some $\delta > 0$, if $|\mathbf{x}_0 - \mathbf{x}^*| < \delta$ then $\lim_{t \rightarrow \infty} \mathbf{x}(t) = \mathbf{x}^*$.
- If the fixed point \mathbf{x}^* is not stable then it is **unstable**.

In Figure 7.11 the fixed point near $(1.54, 2.31)$ is asymptotically stable, while all other fixed points are unstable. In Figure 7.13 it appears that $(2, 0)$ and $(0, 3)$ are asymptotically stable while $(0, 0)$ and the fixed point near $(0.9, 0.4)$ is unstable.

7.3.2 Linearizing ODEs at Equilibrium Points

Any conclusions we come to based on the above analysis won't be ironclad, because these conclusions rest on graphical analysis. We can gain further quantitative insight into the behavior of solutions near fixed points by using linearization.

Linearizing Multivariable Functions

Recall from multivariable calculus that a function $f(x, y)$ of two variables can be well-approximated near a point $x = c, y = d$, as $f(x, y) \approx L(x, y)$, where $L(x, y)$ is the linear function

$$L(x, y) = f(c, d) + \frac{\partial f}{\partial x}(c, d)(x - c) + \frac{\partial f}{\partial y}(c, d)(y - d). \quad (7.22)$$

The function L is called the **linearization of f at the point (c, d)** . We require that the partial derivatives exist, of course, and for L to approximate f well these partial derivatives should be continuous near $x = c, y = d$. The value of linearization is that if L is a good approximation to the (possibly nonlinear) function f near the point $x = c, y = d$, then L can replace f in certain computations. Because L is linear, the computation may be greatly simplified.

■ **Example 7.4** Let $f(x, y) = x(1 - x/2) - xy/10$. To linearize f at the point $x = 2, y = 0$ we compute $\frac{\partial f}{\partial x} = 1 - x - y/10$, $\frac{\partial f}{\partial y} = -x/10$. From (7.22) with $c = 2, d = 0$ we find that $\frac{\partial f}{\partial x}(2, 0) = -1$ and $\frac{\partial f}{\partial y}(2, 0) = -1/5$, so that

$$L(x, y) = 0 - (x - 2) - (y - 0)/5 = -x - y/5 + 2.$$

A graph of the functions $f(x, y)$ and $L(x, y)$ near the point $x = 2, y = 0$ is shown in Figure 7.14. From a geometric perspective, the graph $z = L(x, y)$ is the tangent plane to the surface $z = f(x, y)$ at the point $x = 2, y = 0, z = f(2, 0) = 0$. ■

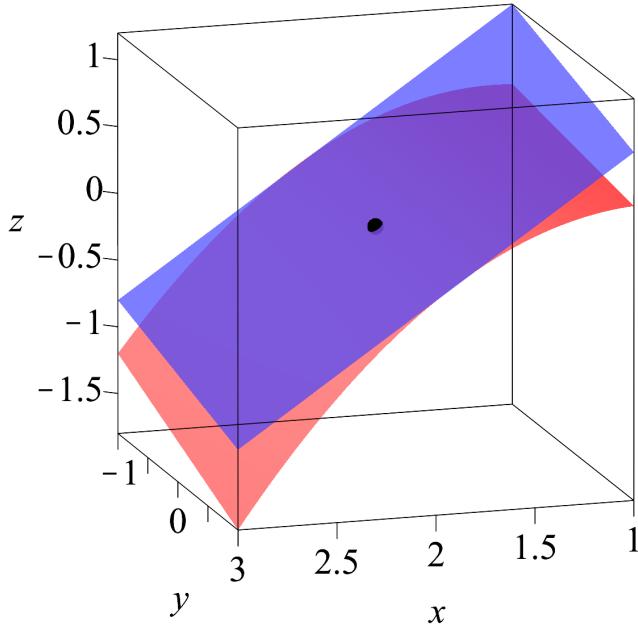


Figure 7.14: Graph of $z = f(x, y) = x(1 - x/2) - xy/10$ in red and graph of linearization $z = L(x, y) = -x - y/5 + 2$ in blue, at $x = 2, y = 0, z = f(2, 0) = 0$ (black dot).

When the partial derivatives $\partial f / \partial x$ and $\partial f / \partial y$ exist and are continuous near $x = c, y = d$, it can be shown that $L(x, y)$ is a good approximation to $f(x, y)$; see [39]. We can linearize a function

of three or more variables in a similar manner. Let $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{c} = (c_1, \dots, c_n)$. A function $f(\mathbf{x})$ can be linearized near the point $\mathbf{x} = \mathbf{c}$ as

$$L(\mathbf{x}) = f(\mathbf{c}) + \sum_{k=1}^n \frac{\partial f}{\partial x_k}(\mathbf{c})(x_k - c_k). \quad (7.23)$$

Application to ODE Stability Analysis

Linearization is an important tool for analyzing how solutions to nonlinear systems of ODEs behave near fixed points. To illustrate, let's return to the system (7.20)-(7.21), whose phase portrait was shown in Figure 7.11, and analyze the behavior of solutions near the fixed point $u_1 = 2, u_2 = 0$. To do this we linearize each of f_1 and f_2 at this point. Let $L_1(u_1, u_2)$ and $L_2(u_1, u_2)$ denote these respective linearizations. We will then use eigenvalue techniques to determine the stability of the linearized system $\dot{u}_1 = L_1(u_1, u_2), \dot{u}_2 = L_2(u_1, u_2)$, and use this to infer the stability of the original nonlinear system near the relevant fixed point. We carry out this computation in Example 7.5. Under the right circumstances the nonlinear and linearized system can be asserted to share the same stability properties at the fixed point, as detailed in the Hartman-Grobman theorem, stated below.

■ Example 7.5 For the system (7.20)-(7.21) the linearization of f_1 at $u_1 = 2, u_2 = 0$ was computed in Example 7.4 (with variables x and y instead of u_1 and u_2) and is $L_1(u_1, u_2) = -u_1 - u_2/5 + 2$. By using $f_2(u_1, u_2) = 2u_2(1 - u_2/3) - 3u_1u_2/10$, we find the linearization $L_2(u_1, u_2) = 7u_2/5$ at $u_1 = 2, u_2 = 0$. The resulting linearized system is thus $\dot{u}_1 = -u_1 - u_2/5 + 2, \dot{u}_2 = 7u_2/5$. In matrix form the linearized system is $\dot{\mathbf{u}} = \mathbf{A}\mathbf{u} + \mathbf{b}$, where $\mathbf{u}(t) = \langle u_1(t), u_2(t) \rangle$,

$$\mathbf{A} = \begin{bmatrix} -1 & -1/5 \\ 0 & 7/5 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = -\begin{bmatrix} 2 \\ 0 \end{bmatrix}. \quad (7.24)$$

Based on the discussion in Section 7.2, the stability of the fixed point $c = 2, d = 0$ for the linearized system is determined by the eigenvalues of \mathbf{A} , which turn out to be $\lambda_1 = -1, \lambda_2 = 7/5$. Since these are real and of mixed sign, the point $u_1 = 2, u_2 = 0$ is a saddle point for the linear system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}$. ■

The Hartman-Grobman theorem will allow us to assert that $u_1 = 2, u_2 = 0$ behaves like a saddle point for the nonlinear system (7.20)-(7.21) as well, so the fixed point is unstable. This computation can be carried out for the other fixed points for (7.20)-(7.21), which we do below.

Linearized Stability Analysis in Two Dimensions

If $u_1 = c, u_2 = d$ is a fixed point for the nonlinear system $\dot{u}_1 = f_1(u_1, u_2), \dot{u}_2 = f_2(u_1, u_2)$, then $f_1(c, d) = f_2(c, d) = 0$, and so from (7.22) the linearizations are

$$\begin{aligned} L_1(u_1, u_2) &= \frac{\partial f_1}{\partial u_1}(c, d)(u_1 - c) + \frac{\partial f_1}{\partial u_2}(c, d)(u_2 - d) \\ L_2(u_1, u_2) &= \frac{\partial f_2}{\partial u_1}(c, d)(u_1 - c) + \frac{\partial f_2}{\partial u_2}(c, d)(u_2 - d). \end{aligned}$$

We also see that $L_1(c, d) = 0$ and $L_2(c, d) = 0$, so that (c, d) is also a fixed point for the linearized system $\dot{u}_1 = L_1(u_1, u_2), \dot{u}_2 = L_2(u_1, u_2)$. Based on this the linearized system can be written in the matrix form $\dot{\mathbf{u}} = \mathbf{A}\mathbf{u} + \mathbf{b}$ where

$$\mathbf{A} = \begin{bmatrix} \frac{\partial f_1}{\partial u_1}(c, d) & \frac{\partial f_1}{\partial u_2}(c, d) \\ \frac{\partial f_2}{\partial u_1}(c, d) & \frac{\partial f_2}{\partial u_2}(c, d) \end{bmatrix} \quad \text{and} \quad \mathbf{b} = -\mathbf{A} \begin{bmatrix} c \\ d \end{bmatrix}.$$

As per the analysis in Subsection 7.2.3, it is the eigenvalues for \mathbf{A} that dictate the stability of the linearized system.

Reading Exercise 7.3.8 Linearize (7.20)-(7.21) at the fixed point $u_1 \approx 1.54, u_2 \approx 2.31$ and write out the matrix \mathbf{A} that governs the linearized system. Compute the eigenvalues for \mathbf{A} . What is the stability of this fixed point for the linearized system? If the corresponding fixed point for the nonlinear system also possesses this stability, what does that say about this model's prediction concerning the mutual coexistence of the species?

The Jacobian Matrix

The process of linearizing a nonlinear system of two ODEs that gives rise to the matrix \mathbf{A} in Example 7.5 works for larger systems. When we linearize each function f_j in a system $\dot{u}_j = f_j(u_1, \dots, u_n)$, $1 \leq j \leq n$, using (7.23) we are led to a linearized system $\dot{\mathbf{u}} = \mathbf{Au} + \mathbf{b}$ where \mathbf{A} is computed using the so-called Jacobian matrix:

Definition 7.3.2 Consider a system of n autonomous ODEs of the form (7.1) in which each function f_i has first partial derivatives with respect to each u_j . The $n \times n$ matrix

$$\mathbf{J}(\mathbf{u}) = \begin{bmatrix} \frac{\partial f_1}{\partial u_1} & \cdots & \frac{\partial f_1}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial u_1} & \cdots & \frac{\partial f_n}{\partial u_n} \end{bmatrix} \quad (7.25)$$

with row i , column j entry $\frac{\partial f_i}{\partial u_j}$ is called the **Jacobian matrix** for the system (7.1).

Of course the partial derivatives in (7.25) are functions of u_1, \dots, u_n and hence so is \mathbf{J} .

■ **Example 7.6** For the system (7.20)-(7.21) we have

$$\frac{\partial f_1}{\partial u_1} = -u_1 - u_2/10, \quad \frac{\partial f_1}{\partial u_2} = -u_1/10, \quad \frac{\partial f_2}{\partial u_1} = -3u_2/10, \quad \frac{\partial f_2}{\partial u_2} = -4u_2/3 - 3u_1/10.$$

The Jacobian matrix is then

$$\mathbf{J}(u_1, u_2) = \begin{bmatrix} -u_1 - u_2/10 & -u_1/10 \\ -3u_2/10 & -4u_2/3 - 3u_1/10 \end{bmatrix}.$$

The Jacobian matrix in Example 7.6, when evaluated at the fixed point $u_1 = 2, u_2 = 0$, becomes exactly the matrix \mathbf{A} in (7.24) of Example 7.5. ■

More generally, for a system of autonomous ODEs of the form $\dot{x}_j = f_j(x_1, \dots, x_n)$, the linearization of the system at a fixed point $\mathbf{x} = \mathbf{p}$ where $\mathbf{p} = \langle p_1, \dots, p_n \rangle$ is of the form $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{b}$ where $\mathbf{A} = \mathbf{J}(\mathbf{p})$ and $\mathbf{b} = -\mathbf{J}(\mathbf{p})\mathbf{p}$. The eigenvalues of $\mathbf{J}(\mathbf{p})$ determine the stability of the linearized system at \mathbf{p} , and under certain circumstances, let us infer the stability of the nonlinear system.

The Hartman-Grobman Theorem

When a fixed point is **hyperbolic**, defined below, the linearized system and the nonlinear system share the same stability properties at that fixed point.

Definition 7.3.3 Suppose \mathbf{p} is an equilibrium (fixed) point for an autonomous system (7.1). Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues for $\mathbf{J}(\mathbf{p})$ (some may be complex). If every eigenvalue λ_k has a nonzero real part then \mathbf{p} is called a **hyperbolic equilibrium point** or **hyperbolic fixed point**.

If an eigenvalue λ has a nonzero real part this means either that λ is real and nonzero or that λ is complex with nonzero real part.

The Hartman-Grobman theorem asserts that if \mathbf{p} is a hyperbolic equilibrium point for an autonomous system of ODEs (7.1) then the local stability of the nonlinear system near \mathbf{p} is the same as that of the linearized system. Of particular interest for us is the fact that if \mathbf{p} is asymptotically

stable for the linearized system, then \mathbf{p} is asymptotically stable for the nonlinear system, which occurs when all eigenvalues in the linearized system have negative real part. More generally, in two dimensions, the intuitive insight is that if \mathbf{p} is a saddle point, sink, source, stable spiral, or unstable spiral for the linearized system, then the nonlinear system will exhibit the same qualitative behavior near \mathbf{p} . But solutions to the nonlinear system that start far from \mathbf{p} may do something quite different, as we will see in some examples that follow.

Be aware that this analysis is only applicable for hyperbolic equilibrium points. If any eigenvalue of the Jacobian matrix at an equilibrium point is 0 or purely imaginary, then the nature of this equilibrium point for the nonlinear and linearized system may differ; some examples of such differences are given in the exercises. Our statement of the Hartman-Grobman Theorem isn't as precise as it could be, but it will suffice for our purposes. For a more careful statement see [57].

Summary of Linearized Stability Analysis

In summary, we can use linearization to determine the stability of any fixed point $\mathbf{u} = \mathbf{p}$ for a nonlinear system $\dot{\mathbf{u}} = \mathbf{f}(\mathbf{u})$ with the following procedure:

1. Compute the Jacobian matrix $\mathbf{J}(\mathbf{u})$ defined by (7.25) and let $\mathbf{A} = \mathbf{J}(\mathbf{p})$. The matrix \mathbf{A} governs the linearized ODE system $\dot{\mathbf{u}} = \mathbf{Au} + \mathbf{b}$.
2. Compute the eigenvalues $\lambda_1, \dots, \lambda_n$ of \mathbf{A} . From these deduce the stability of the fixed point \mathbf{p} for the linearized system.
3. If \mathbf{p} is hyperbolic (Definition 7.3.3), then the nonlinear and linearized system have the same qualitative stability properties at \mathbf{p} .

This procedure of linearizing to determine the stability of fixed points is sometimes called **Lyapunov's first method** (also spelled "Liapunov").

■ **Example 7.7** Let's finish our analysis of the system (7.20)-(7.21). In Example 7.6 we computed the Jacobian matrix $\mathbf{J}(u_1, u_2)$. In Example 7.5 we found that the eigenvalues for the equilibrium point $u_1 = 2, u_2 = 0$ are $\lambda_1 = -1$ and $\lambda_2 = 7/5$. Neither eigenvalue has zero real part, so this equilibrium point is a saddle for the linearized system. From the Hartman-Grobman theorem we conclude this equilibrium point behaves like a saddle point for the nonlinear system as well. In particular, this point is unstable, and solutions will generally not approach it. Physically, the situation in which the first species (with population $u_1(t)$) is near its carrying capacity and there are relatively few of the second species present is an unstable situation; solutions will move away from this fixed point. What the solution trajectory does and where it goes in the long run are not specified by this local analysis.

Let's consider the other three fixed points. At $u_1 = u_2 = 0$ we find

$$\mathbf{J}(0,0) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

with eigenvalues 1 and 2. This is a hyperbolic equilibrium point and both eigenvalues are positive, so this is a source for the linearized system. All solutions radiate away in the linearized case, and we will see the same qualitative behavior for the nonlinear system, as is clear from Figure 7.11. In brief, if there is a positive number of each species present, the solution will move away from mutual extinction (the point $u_1 = u_2 = 0$).

The fixed point $u_1 = 0, u_2 = 3$ is similar to $u_1 = 2, u_2 = 0$. The Jacobian matrix is

$$\mathbf{J}(0,3) = \begin{bmatrix} 0.7 & 0 \\ -0.9 & -2.0 \end{bmatrix}$$

with eigenvalues 0.7 and -2 . This is a hyperbolic equilibrium point, a saddle point, and unstable. Solutions that start nearby will not generally tend toward $u_1 = 0, u_2 = 3$ (the extinction of the first species).

Finally, at the fixed point $u_1 \approx 1.54$, $u_2 \approx 2.31$ we find

$$\mathbf{J}(1.54, 2.31) \approx \begin{bmatrix} -0.769 & -0.154 \\ -0.692 & -1.54 \end{bmatrix}$$

with approximate eigenvalues -0.65 and -1.66 . This is a hyperbolic equilibrium point and asymptotically stable for the linearized system. Solutions to the linearized system approach this point, and so do solutions to the nonlinear system that start sufficiently close to $(1.54, 2.31)$. ■

We can use the analysis of Example 7.7 to confirm what the nullclines and trajectories in Figure 7.11 strongly suggest: for the system (7.20)-(7.21), all initial conditions that start close to the coexistence fixed point approach this point. This strengthens the case made by Figure 7.11: all initial populations with a positive number of each species will converge to stable coexistence. Neither will drive the other to extinction.

7.3.3 Exercises

Exercise 7.3.1 For each system of ODEs $\dot{x}_1 = f_1(x_1, x_2)$, $\dot{x}_2 = f_2(x_1, x_2)$, do the following:

- Find and sketch the x_1 nullcline on the indicated range for x_1 and x_2 . The nullcline will divide the plane into a number of regions; put an arrow in each region to indicate whether solutions are moving left or right in that region (as in the left panel of Figure 7.10).
- Find and sketch the x_2 nullcline on the indicated range for x_1 and x_2 . The nullcline will divide the plane into a number of regions; put an arrow in each region to indicate whether solutions are moving up or down in that region (as in the right panel of Figure 7.10).
- Find the equilibrium solutions by solving $f_1(x_1, x_2) = 0$ and $f_2(x_1, x_2) = 0$ simultaneously for (x_1, x_2) . These points are where the nullclines intersect.
- Linearize the system at each equilibrium solution to determine the stability of the equilibrium solution.
- Use your results to sketch an accurate phase portrait with (at least) four or five representative solutions.
- Sketch on your phase portrait a solution trajectory $(x_1(t), x_2(t))$ with the given initial conditions, and then use this to sketch $x_1(t)$ and $x_2(t)$ individually as functions of t (as shown in Figure 7.12 for the system (7.20)-(7.21).)

- (a) Let $\dot{x}_1 = 2 - x_1^2 - x_2$ and $\dot{x}_2 = x_1 - x_2$ on the range $-5 \leq x_1, x_2 \leq 5$. Sketch a solution trajectory with $x_1(0) = 0$ and $x_2(0) = 4$, as well as $x_1(t)$ versus t and $x_2(t)$ versus t . Repeat for the solution with $x_1(0) = 4$ and $x_2(0) = -2$.
- (b) Let $\dot{x}_1 = -2x_1 - x_2 - 2$ and $\dot{x}_2 = -x_1 x_2$ on the range $-5 \leq x_1, x_2 \leq 5$. Sketch a solution trajectory with $x_1(0) = 2$ and $x_2(0) = -2$, as well as $x_1(t)$ versus t and $x_2(t)$ versus t . Repeat for the solution with $x_1(0) = -1$ and $x_2(0) = 3$.
- (c) Let $\dot{x}_1 = x_1 x_2 + x_2^2$ and $\dot{x}_2 = x_1 - 2x_2 + 3$ on the range $-5 \leq x_1, x_2 \leq 5$. Sketch a solution trajectory with $x_1(0) = -3$ and $x_2(0) = 1$, as well as $x_1(t)$ versus t and $x_2(t)$ versus t . Repeat for the solution with $x_1(0) = -2$ and $x_2(0) = -3$.
- (d) Let $\dot{x}_1 = x_1^3 - 3x_1 - x_2$ and $\dot{x}_2 = x_1 - x_2$ on the range $-5 \leq x_1, x_2 \leq 5$. Sketch a solution trajectory with $x_1(0) = 0$ and $x_2(0) = 3$, as well as $x_1(t)$ versus t and $x_2(t)$ versus t . Repeat for the solution with $x_1(0) = -2$ and $x_2(0) = 3$.

Exercise 7.3.2 Compute the Jacobian matrix for the system (7.3)-(7.4) with parameter choices $r_1 = 1$, $K_1 = 2$, $r_2 = 2$, $K_2 = 3$, and $a = b = 3$. The nullclines and some solution trajectories were shown in Figure 7.13. Show that the fixed points in this case are (u_1, u_2) equal to one of $(0, 0)$, $(2, 0)$, $(0, 3)$, and $(7/8, 3/8)$. Evaluate the Jacobian at each of these fixed points, compute the eigenvalues, and show that in this case $(0, 0)$ is still an unstable source, but now each of $(2, 0)$ and $(0, 3)$ are asymptotically stable, while the coexistence fixed point $(7/8, 3/8)$ behaves like a saddle point.

Convince yourself this is consistent with Figure 7.13. What can you conclude about the long-term behavior of the populations here?

Exercise 7.3.3 The epidemic model (7.6) is a system of three ODEs in three functions $S(t)$, $I(t)$, and $R(t)$, but it can effectively be considered a system of two differential equations

$$\begin{aligned}\dot{S} &= -aSI \\ \dot{I} &= aSI - bI\end{aligned}\tag{7.26}$$

where $a, b > 0$, since the first two equations don't involve R . Let's consider the special case in which $a = 1$ and $b = 2$ and analyze (7.26) to determine the behavior of $S(t)$ and $I(t)$. We will then use this to determine the behavior of $R(t)$ using $\dot{R} = bI$. We can restrict our attention to the first quadrant $S, I \geq 0$, where we will consider S to be the horizontal coordinate and I the vertical coordinate. (We might also think of S and I as quantifying the population in thousands or millions, for a more realistic scale.)

- (a) Show that the S nullcline is given by the coordinate axes $S = 0$ and $I = 0$. Show that the I nullcline is given by the horizontal coordinate axis $I = 0$ and vertical line $S = 2$.
- (b) Sketch the S nullcline and appropriate arrow(s) to indicate the horizontal motion of any solution.
- (c) Sketch the I nullcline and appropriate arrow(s) to indicate the vertical motion of any solution.
- (d) Show that the fixed points for this system are exactly those points for which $I = 0$, which is the entire S axis. Thus for this system, the fixed points are not isolated.
- (e) Linearize the system at a typical fixed point $(S_0, 0)$. Show that the eigenvalues for the Jacobian matrix are $\lambda_1 = 0$ and $\lambda_2 = S_0 - 2$. Does the Hartman-Grobman theorem apply?
- (f) Sketch a phase portrait for this system using the above information. What conclusion can you make about the long-term behavior of the system—how will the number of susceptible and infected people change over time? Will anyone escape getting infected, and what can you say about this number? What will happen to R ?

Exercise 7.3.4 Consider the damped nonlinear pendulum equation (7.7). An equivalent system

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{g}{L} \sin(x_1) - cx_2\end{aligned}$$

was given in (7.8). Here g and L are both positive and denote gravitational acceleration and the length of the pendulum, respectively, c is a nonnegative damping constant, and $x_1(t)$ and $x_2(t)$ are the angular position and angular velocity of the pendulum with respect to the vertical. In what follows take $g = 9.81$, $L = 1$, and $c = 1$.

- (a) Show that the x_1 nullcline is the horizontal axis in the x_1x_2 plane. Sketch this nullcline with appropriate arrows above and below the nullcline to indicate horizontal solution motion.
- (b) Show that the x_2 nullcline is the sine curve with graph $x_2 = -9.81 \sin(x_1)$. Carefully sketch this curve on the range $-4\pi \leq x_1 \leq 4\pi$, with appropriate arrows above and below the nullcline to indicate vertical solution motion.
- (c) Show that the fixed points for this system are all of the form $x_1 = k\pi$, $x_2 = 0$ where k is an integer. Interpret this result physically: What is the configuration of the pendulum if $(x_1, x_2) = (0, 0)$? What is the configuration of the pendulum if $(x_1, x_2) = (\pi, 0)$? What is the configuration of the pendulum if $(x_1, x_2) = (2\pi, 0)$?
- (d) Linearize the system at each point $(x_1, x_2) = (0, 0)$, $(\pi, 0)$, $(2\pi, 0)$ and find the eigenvalues of the Jacobian matrix. Interpret the result. What does this say about the pendulum's stability in each position? What does it tell you about the behavior of the pendulum's motion near these points?
- (e) Use this information to sketch a phase portrait for this system, with a few trajectories. What does this tell you about the motion of the pendulum?

Exercise 7.3.5 Analyze the system

$$\begin{aligned}\dot{x} &= -x + y + 1 \\ \dot{y} &= xz + 1 \\ \dot{z} &= -x - z\end{aligned}$$

by finding the fixed points, linearizing about each, and determining the stability of each fixed point. Of course, you can't really draw nullclines, but can you guess how solutions typically behave? Try solving the system numerically for a variety of initial conditions and plotting the results to confirm your analysis.

7.4 Analyzing Systems with Unspecified Parameters

The phase portrait for the system (7.20)-(7.21) corresponding to the choices $r_1 = 1$, $K_1 = 2$, $r_2 = 2$, $K_2 = 3$, $a = 0.2$, and $b = 0.45$ in (7.3)-(7.4) was shown in Figure 7.11; all solutions tend to the equilibrium point $u_1 \approx 1.54$, $u_2 \approx 2.31$, and stable coexistence will always occur. Changing the competition parameters a and b to $a = b = 3$ leads to the phase portrait of Figure 7.13 and makes a rather dramatic change in the behavior of the system. Stable coexistence is no longer a possibility, and one species or the other must go extinct. These conclusions are supported by the analysis based on linearization as well.

We could have had the computer sketch the direction field for both scenarios above and used that to visualize the solutions. Why trouble ourselves with the nullclines, manually computed direction arrows, and hand-sketched solution curves? The value of this approach is that we can sketch phase portraits without assuming specific values for the critical parameters and then examine how these parameters affect the phase portraits. This is something the computer doesn't do. When applied to the competing species model these techniques will allow us to determine what combinations of the parameters r_1 , K_1 , r_2 , K_2 , a , and b allow coexistence and what combinations doom one species or the other to extinction. Perhaps other outcomes are possible.

7.4.1 Sketching Phase Portraits with Unspecified Parameters

Let's draw a phase portrait for (7.3)-(7.4) without assuming specific parameter values. The techniques used here are of much more general applicability, and opportunities to apply them to other systems of ODEs are given in the exercises at the end of this section and in projects in Section 7.7.

Rescaling the ODEs

Before proceeding it will be helpful to make a change in dependent variables by replacing u_1 and u_2 in (7.3)-(7.4) with rescaled functions

$$v_1 = u_1/K_1 \quad \text{and} \quad v_2 = u_2/K_2, \quad (7.27)$$

in the spirit of Section 4.5. This rescaling changes the units used to measure population, so that v_1 quantifies the population of the first species in units of the carrying capacity K_1 and similarly for v_2 . In this case $u_1 = K_1 v_1$ and $u_2 = K_2 v_2$ so that $\dot{u}_1 = K_1 \dot{v}_1$ and $\dot{u}_2 = K_2 \dot{v}_2$.

Reading Exercise 7.4.1 Verify that with (7.27) (along with the consequences $\dot{u}_1 = K_1 \dot{v}_1$ and $\dot{u}_2 = K_2 \dot{v}_2$) the system (7.3)-(7.4) yields the ODEs

$$\dot{v}_1 = r_1 v_1 (1 - v_1 - \bar{a} v_2) \quad (7.28)$$

$$\dot{v}_2 = r_2 v_2 (1 - v_2 - \bar{b} v_1) \quad (7.29)$$

for the functions $v_1(t)$ and $v_2(t)$, where

$$\bar{a} = K_2 a / K_1 \quad \text{and} \quad \bar{b} = K_1 b / K_2. \quad (7.30)$$

We will analyze the ODE system (7.28)-(7.29), then use (7.27) to make conclusions about the original system (7.3)-(7.4). The advantage of (7.28)-(7.29) is that there are only four explicit parameters to consider, r_1, r_2, \bar{a} , and \bar{b} . As it turns out, r_1 and r_2 have little bearing on the phase portrait we'll draw; it is \bar{a} and \bar{b} that determine the qualitative behavior of the system and the fate of each species.

A Phase Portrait for the Competing Species ODEs (7.28)-(7.29)

The process for sketching the phase portraits in Section 7.3 consisted of computing and sketching the nullclines, drawing direction arrows, linearizing at each fixed point, and then sketching representative solution trajectories. We will now carry out these steps on the system (7.28)-(7.29) without assuming specific values for r_1, r_2, \bar{a} , or \bar{b} , aside from the assumption that all are positive. The only region in the $v_1 v_2$ phase plane of interest is the closed first quadrant, $v_1, v_2 \geq 0$, corresponding to nonnegative populations.

To find the v_1 nullcline, set $f_1(v_1, v_2) = 0$, which leads to

$$r_1 v_1 (1 - v_1 - \bar{a} v_2) = 0.$$

Since $r_1 > 0$ this nullcline consists of the lines $v_1 = 0$ (the v_2 axis) and $1 - v_1 - \bar{a} v_2 = 0$, or $v_2 = 1/\bar{a} - v_1/\bar{a}$. This last line has v_2 -intercept $1/\bar{a}$, v_1 -intercept 1, and slope $-1/\bar{a}$, and is shown in the left panel of Figure 7.15, with a few vertical tick marks to indicate solution directions on the nullcline. This nullcline can be sketched without choosing specific values for the system parameters, as long as the intercepts are labeled appropriately. Straightforward algebra shows that for $v_2 > 1/\bar{a} - v_1/\bar{a}$ (above the diagonal portion of this nullcline), we have $1 - v_1 - \bar{a} v_2 < 0$, and so $\dot{v}_1 < 0$. This means that solutions above the nullcline move generally to the left, in the direction of decreasing v_1 (that is, the v_1 population is declining in this region). Similarly if $v_2 < 1/\bar{a} - v_1/\bar{a}$ then $1 - v_1 - \bar{a} v_2 > 0$, so $\dot{v}_1 > 0$ and solutions move to the right, reflecting that the v_1 population is

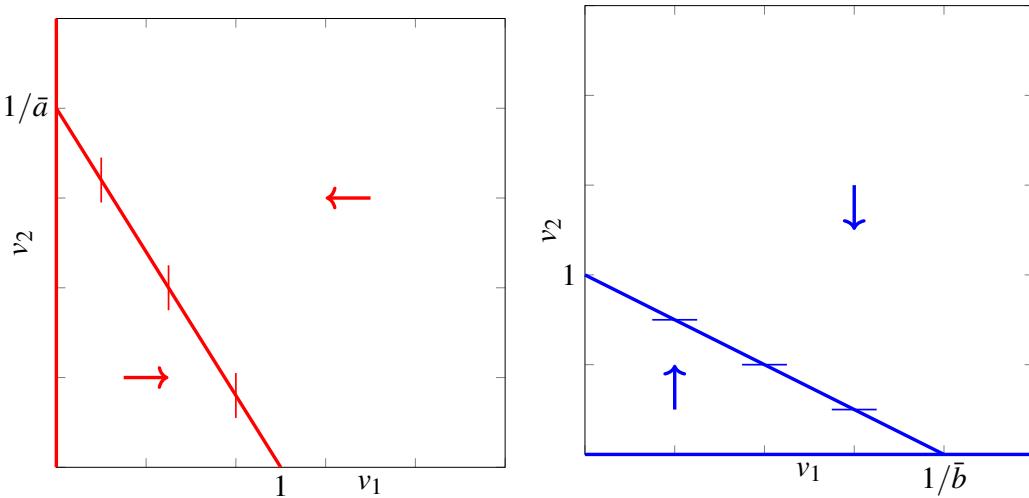


Figure 7.15: Left panel: sketch of v_1 nullcline for the competing species ODE model (7.28)-(7.29). Right panel: sketch of v_2 nullcline for the competing species ODE model (7.28)-(7.29).

increasing. The situation is indicated by the appropriate left and right pointing vectors in the left panel of Figure 7.15.

The v_2 nullcline can be sketched in a similar manner. Set $f_2(v_1, v_2) = 0$ so that from (7.29) the nullcline $\dot{v}_2 = 0$ is defined by the equation $r_2 v_2(1 - v_2 - \bar{b}v_1) = 0$, which yields $v_2 = 0$ (the horizontal v_1 axis) and the line $1 - v_2 - \bar{b}v_1 = 0$. This last line has v_1 intercept $v_1 = 1/\bar{b}$, v_2 intercept $v_2 = 1$, and slope $-1/\bar{b}$. This nullcline is depicted in the right panel of Figure 7.15. The nullcline divides the first quadrant into two distinct pieces, above and below the line $1 - v_2 - \bar{b}v_1 = 0$. Above this line $f_2(v_1, v_2) < 0$ and so $\dot{v}_2 < 0$, while below the line $f_2(v_1, v_2) > 0$ and so $\dot{v}_2 > 0$. This is indicated by the up or down arrows in the right panel.

Reading Exercise 7.4.2 Show that the equilibrium points for (7.28)-(7.29) are

$$(0,0), (1,0), (0,1), ((\bar{a}-1)/(\bar{a}\bar{b}-1), (\bar{b}-1)/(\bar{a}\bar{b}-1)).$$

Interpret the physical meaning of each equilibrium point. When might the last one be physically irrelevant?

The next step is to superimpose the nullcline sketches in the left and right panels of Figure 7.15, draw direction arrows, and sketch solution trajectories (possibly after linearizing as well). However, there is a problem: we don't know the precise relationship between the diagonal lines that comprise a part of each nullcline in Figure 7.15. As drawn in that figure, they represent the case in which $1 < 1/\bar{a}$ and $1 < 1/\bar{b}$ or, equivalently $\bar{a} < 1$ and $\bar{b} < 1$, so the diagonal portions of the nullclines are certain to cross. If this is the case then the nullclines in the left and right panel of Figure 7.15 can be superimposed to produce the left panel in Figure 7.16.

This is qualitatively identical to the left panel in Figure 7.11. As a result, the solutions will be similar to those in the right panel of Figure 7.11.

Reading Exercise 7.4.3 Verify that the parameter choices $r_1 = 1$, $r_2 = 2$, $K_1 = 2$, $K_2 = 3$, $a = 0.2$, and $b = 0.45$ in (7.20)-(7.21) that led to Figures 7.10 and 7.11 yield $\bar{a} = 0.3$ and $\bar{b} = 0.3$ in (7.28)-(7.29) and satisfy $1 < 1/\bar{a}$ and $1 < 1/\bar{b}$.

Now consider the possibility that $1 > 1/\bar{a}$ and $1 > 1/\bar{b}$. This is illustrated in the right panel of Figure 7.16. The situation is similar to the one that led to Figure 7.13 with $r_1 = 1$, $r_2 = 2$, $K_1 = 2$, $K_2 = 3$, $a = 3$, and $b = 3$, corresponding to $\bar{a} = 4.5$, $\bar{b} = 2$.

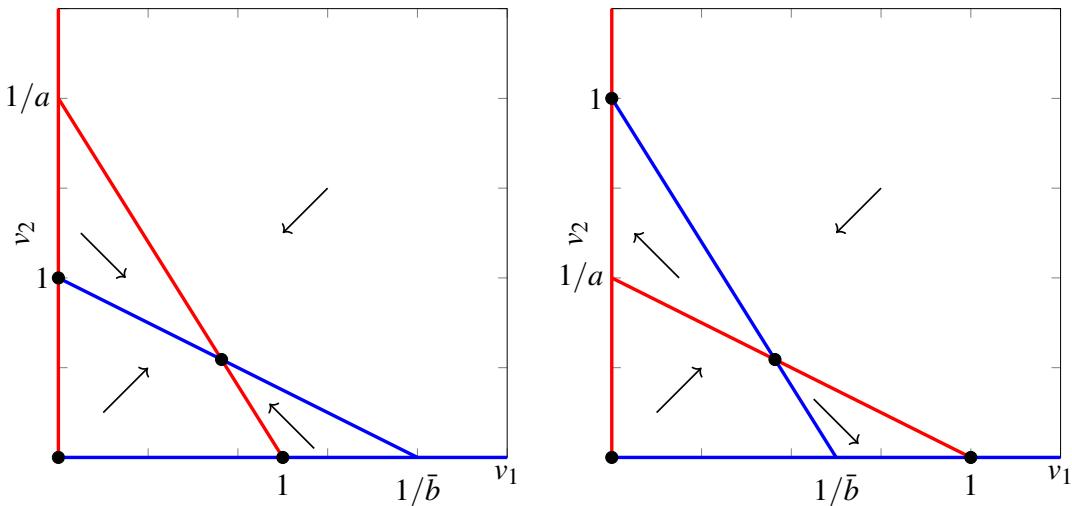


Figure 7.16: Left panel: nullclines and direction arrows for the competing species model (7.28)-(7.29) when $1 < 1/\bar{a}$ and $1 < 1/\bar{b}$. Right panel: nullclines and direction arrows for the competing species model (7.28)-(7.29) when $1 > 1/\bar{a}$ and $1 > 1/\bar{b}$. Equilibrium solutions in each case are shown as black dots.

Reading Exercise 7.4.4 Sketch the nullclines for (7.28)-(7.29) under the assumption that $1 > 1/\bar{a}$ and $1 > 1/\bar{b}$, and verify that the picture in the right panel of Figure 7.16 is correct.

The two cases illustrated in Figure 7.16 aren't the only possibilities. Fortunately the growth rates r_1 and r_2 don't factor in to the analysis, but even with the two parameters \bar{a} and \bar{b} , there are other cases to consider.

Reading Exercise 7.4.5 Argue that even if we exclude the razor's edge cases in which either $\bar{a} = 1$ or $\bar{b} = 1$, there are four total cases to consider:

1. $1 < 1/\bar{a}$ and $1 < 1/\bar{b}$ (examined above in the left panel of Figure 7.16).
2. $1 > 1/\bar{a}$ and $1 > 1/\bar{b}$ (examined above in the right panel of Figure 7.16).
3. $1 < 1/\bar{a}$ and $1 > 1/\bar{b}$.
4. $1 > 1/\bar{a}$ and $1 < 1/\bar{b}$.

We will examine the third and fourth cases in the Exercises, along with the possibilities that at least one of $\bar{a} = 1$ or $\bar{b} = 1$.

Reading Exercise 7.4.6 Based on the analysis above and the left panel of Figure 7.16, finish the phase portrait for the system (7.28)-(7.29) when $1 < 1/\bar{a}$ and $1 < 1/\bar{b}$ by including some representative solution curves. How do solutions behave as t increases to infinity? Translate your conclusions back to the original population variables u_1 and u_2 by using (7.27). What is the fate of each species? Repeat your analysis for the case in which $1 > 1/\bar{a}$ and $1 > 1/\bar{b}$, which is (case (2) in Reading Exercise 7.4.5). The right panel of Figure 7.16 will be helpful.

Although our analysis of this system isn't finished, we can already make some important conclusions: In case (1) above we expect stable coexistence of the competing species, while in case (2) we expect that one of the species will eventually dominate and drive the other to extinction. Which species goes extinct depends on the initial conditions.

7.4.2 Linearizing the Competing Species Model with General Parameters

Let us now illustrate the power of the type of qualitative analysis we've developed in this section by thoroughly analyzing the competing species model (7.28)-(7.29) with unspecified positive

parameters r_1, r_2, \bar{a} , and \bar{b} . We've already sketched the nullclines and a simple phase portrait in Figure 7.16. We will now perform linearization of the system at the equilibrium points and determine precisely how the various parameters affect the fate of each species. The material in Appendix B, and in particular Section B.4, will be a great help in understanding how the eigenvalues of the linearized system and the fixed point stability depend on the parameters.

Equations (7.28)-(7.29) are reproduced here for convenience:

$$\dot{v}_1 = r_1 v_1 (1 - v_1 - \bar{a} v_2) \quad (7.31)$$

$$\dot{v}_2 = r_2 v_2 (1 - v_2 - \bar{b} v_1). \quad (7.32)$$

Recall that $\bar{a} = K_2 a / K_1$ and $\bar{b} = K_1 b / K_2$, and both \bar{a} and \bar{b} are positive.

The equilibrium points for this system were given in Reading Exercise 7.4.2 and are

$$(0,0), (1,0), (0,1), ((\bar{a}-1)/(\bar{a}\bar{b}-1), (\bar{b}-1)/(\bar{a}\bar{b}-1)). \quad (7.33)$$

The first three points above are always physically relevant; the fourth fixed point is relevant only when it lies in the first quadrant. Some typical possibilities for the nullclines were shown in Figure 7.16 and it is clear that the stability of the fixed points changes depending on the values of the parameters. In particular, in the left panel of Figure 7.16 it appears that the fixed point at $((\bar{a}-1)/(\bar{a}\bar{b}-1), (\bar{b}-1)/(\bar{a}\bar{b}-1))$ is stable and those at $(1,0)$ and $(0,1)$ are unstable, while in the right panel the situation is reversed. The origin looks unstable in both cases. Linearization can provide additional insight into this phenomena.

The Jacobian matrix for (7.31)-(7.32) is

$$\mathbf{J}(v_1, v_2) = \begin{bmatrix} r_1(1 - 2v_1 - \bar{a}v_2) & -r_1\bar{a}v_1 \\ -r_2\bar{b}v_2 & r_2(1 - 2v_2 - \bar{b}v_1) \end{bmatrix}.$$

We will evaluate $\mathbf{J}(v_1, v_2)$ at each equilibrium point, compute the relevant eigenvalues, and use this to determine the nature of the equilibrium point.

Stability of the Origin

At $v_1 = v_2 = 0$ (corresponding to $u_1 = u_2 = 0$ in the original system when both species are extinct) we find

$$\mathbf{J}(0,0) = \begin{bmatrix} r_1 & 0 \\ 0 & r_2 \end{bmatrix}.$$

The eigenvalues are r_1 and r_2 , and both are positive. The origin is always an unstable source for any positive values of the parameters r_1 and r_2 . If any positive amount of either species is present, the system will move away from this point of mutual extinction.

Stability When One Species Is Extinct

The fixed point $v_1 = 1, v_2 = 0$ corresponds to $u_1 = K_1, u_2 = 0$ in the original system, so the second species is extinct and the first species is at its carrying capacity K_1 . In this case

$$\mathbf{J}(1,0) = \begin{bmatrix} -r_1 & -r_1\bar{a} \\ 0 & r_2(1-\bar{b}) \end{bmatrix}.$$

This matrix is upper triangular as discussed in Appendix B, so the eigenvalues are exactly the diagonal elements $-r_1$ and $r_2(1-\bar{b})$; see Reading Exercise B.3.7. Of course $-r_1$ is always negative, while the sign of the second eigenvalue is determined by the sign of $1-\bar{b}$. If $\bar{b} < 1$ then $1-\bar{b} > 0$ and this is a saddle point, while if $\bar{b} > 1$ then $1-\bar{b} < 0$ and this is an asymptotically stable fixed

point. When $\bar{b} = 1$ this is not a hyperbolic equilibrium and we cannot ascertain the stability of this fixed point using these methods.

At $v_1 = 0, v_2 = 1$ we find

$$\mathbf{J}(0, 1) = \begin{bmatrix} r_1(1 - \bar{a}) & 0 \\ -r_2\bar{b} & -r_2 \end{bmatrix}.$$

This matrix is lower triangular, so the eigenvalues are exactly the diagonal elements $-r_2$ and $r_1(1 - \bar{a})$; again, see Reading Exercise B.3.7. This is similar to the previous fixed point. If $\bar{a} < 1$ this is a saddle point, and if $\bar{a} > 1$ this is an asymptotically stable fixed point. If $\bar{a} = 1$ this equilibrium solution is not hyperbolic and linearization will not reveal its nature.

Reading Exercise 7.4.7 Summarize the stability analysis above for the fixed points $(1, 0)$ and $(0, 1)$ in terms of \bar{a} and \bar{b} . What does each case imply for the population of each species? Reconcile your conclusions with the phase portraits on the left ($\bar{a} < 1$ and $\bar{b} < 1$) and on the right ($\bar{a} > 1$ and $\bar{b} > 1$) of Figure 7.16, and your work in Reading Exercise 7.4.6. Given the interpretation of \bar{a} and \bar{b} as competition parameters, why does this make sense?

Stability for Mutual Coexistence

The last equilibrium solution in the list (7.33) corresponds to mutual coexistence. To examine the stability here let $v_1^* = (\bar{a} - 1)/(\bar{a}\bar{b} - 1)$ and $v_2^* = (\bar{b} - 1)/(\bar{a}\bar{b} - 1)$ denote the coordinates of this equilibrium point, assuming $\bar{a}\bar{b} - 1 \neq 0$. This equilibrium solution is of interest only when the point (v_1^*, v_2^*) lies in the first quadrant with $v_1^* > 0$ and $v_2^* > 0$. The Jacobian at this point is

$$\begin{aligned} \mathbf{J}(v_1^*, v_2^*) &= \frac{1}{\bar{a}\bar{b} - 1} \begin{bmatrix} -r_1(\bar{a} - 1) & -r_1\bar{a}(\bar{a} - 1) \\ -r_2\bar{b}(\bar{b} - 1) & -r_2(\bar{b} - 1) \end{bmatrix} \\ &= \begin{bmatrix} -r_1v_1^* & -r_1\bar{a}v_1^* \\ -r_2\bar{b}v_2^* & -r_2v_2^* \end{bmatrix}. \end{aligned} \tag{7.34}$$

The matrix in (7.34) is not upper or lower triangular and the eigenvalues are a bit messy. This is where the trace-determinant analysis of Theorem B.4.1 in Section B.4 is a big help.

To determine the nature of the eigenvalues for $\mathbf{J}(v_1^*, v_2^*)$ in (7.34) we compute the trace and determinant of $\mathbf{J}(v_1^*, v_2^*)$,

$$\begin{aligned} T &= \text{tr}(\mathbf{J}(v_1^*, v_2^*)) = -r_1v_1^* - r_2v_2^* \\ D &= \det(\mathbf{J}(v_1^*, v_2^*)) = (1 - \bar{a}\bar{b})r_1r_2v_1^*v_2^*. \end{aligned} \tag{7.35}$$

Note that $T < 0$ in all cases. If $\bar{a}\bar{b} > 1$ then $(1 - \bar{a}\bar{b}) < 0$ and so $D < 0$. Based on Theorem B.4.1, the eigenvalues of $\mathbf{J}(v_1^*, v_2^*)$ are real and of mixed sign. In this case (v_1^*, v_2^*) behaves as a saddle point, and is unstable. If $\bar{a}\bar{b} < 1$ then $(1 - \bar{a}\bar{b}) > 0$ and so $D > 0$. Based on the results in Section B.4, both eigenvalues are either real and negative, or both eigenvalues are complex with negative real part. In either case, this fixed point will be asymptotically stable and mutual coexistence is a stable condition.

Reading Exercise 7.4.8 Consider the case when $D > 0$. Use (7.35) to show that

$$T^2 - 4D = (r_1v_1^* - r_2v_2^*)^2 + 4\bar{a}\bar{b}r_1r_2v_1^*v_2^*$$

and explain why this shows that $T^2 - 4D > 0$ is always true. Use the analysis of Section B.4 to conclude that both eigenvalues of $\mathbf{J}(v_1^*, v_2^*)$ are in fact real, so this is a stable node.

7.4.3 Conclusions for Competing Species

We can now make some firm conclusions concerning the fate of each species in (7.3)-(7.4), and how their fates depend on the values of the relevant parameters. In all cases we assume $\bar{a}\bar{b} - 1 \neq 0$, and recall that $\bar{a}, \bar{b} > 0$. Since $\bar{a} = K_2 a / K_1$ and $\bar{b} = K_1 b / K_2$ we have

$$\bar{a}\bar{b} = ab$$

and therefore conclusions that rest on the value of $\bar{a}\bar{b}$ hold for the product ab as well. Moreover, conditions like $\bar{a} < 1$ can be translated by using (7.30) into, for example, $K_2 a / K_1 < 1$ or $a < K_1 / K_2$.

For any parameter choices in the system (7.28)-(7.29), the origin $(v_1, v_2) = (0, 0)$ is an unstable source. From (7.27) we have $u_1 = K_1 v_1$ and $u_2 = K_2 v_2$, so we can make the same conclusion concerning $(u_1, u_2) = (0, 0)$ for (7.3)-(7.4). But for the other equilibrium solutions the possibilities are as follows.

1. Consider the low competition case in which $0 < \bar{a} < 1$ and $0 < \bar{b} < 1$. Then $\bar{a}\bar{b} < 1$ and the coexistence equilibrium point

$$(v_1^*, v_2^*) = \left(\frac{\bar{a} - 1}{\bar{a}\bar{b} - 1}, \frac{\bar{b} - 1}{\bar{a}\bar{b} - 1} \right) \quad (7.36)$$

lies in the first quadrant. From the analysis above we see that both equilibrium points $(v_1, v_2) = (1, 0)$ and $(v_1, v_2) = (0, 1)$ for (7.28)-(7.29) are saddle points, and hence so are the points $(K_1, 0)$ and $(0, K_2)$ for (7.3)-(7.4). But the discussion concerning mutual coexistence then shows that since $\bar{a}\bar{b} < 1$, the point (v_1^*, v_2^*) is asymptotically stable. In the original system (7.3)-(7.4) the coexistence equilibrium point defined by (7.5) is also asymptotically stable.

2. Consider the high competition case in which $\bar{a} > 1$ and $\bar{b} > 1$. Then $\bar{a}\bar{b} > 1$ and again the coexistence equilibrium point (v_1^*, v_2^*) defined by (7.36) lies in the first quadrant. From the analysis above we see that both equilibrium points $(v_1, v_2) = (1, 0)$ and $(v_1, v_2) = (0, 1)$ for (7.28)-(7.29) are asymptotically stable and hence so are the points $(K_1, 0)$ and $(0, K_2)$ for (7.3)-(7.4). But the discussion concerning mutual coexistence shows that since $\bar{a}\bar{b} > 1$, the point (v_1^*, v_2^*) is a saddle point. The same conclusion applies to the coexistence equilibrium point (7.5) for (7.3)-(7.4).
3. Suppose $\bar{a} > 1$ while $\bar{b} < 1$. In this case exactly one of v_1^* or v_2^* in (7.36) is negative and the equilibrium solution (v_1^*, v_2^*) is not physically relevant; the same holds for the solution (7.5) for (7.3)-(7.4). In this case since $\bar{a} > 1$, the fixed point $(0, 1)$ is asymptotically stable, and since $\bar{b} < 1$ the point $(1, 0)$ is a saddle point. This means that $(0, K_2)$ is asymptotically stable and $(K_1, 0)$ is a saddle point for the original system.
4. Suppose $\bar{a} < 1$ while $\bar{b} > 1$. This is entirely analogous to the last case, but with the roles of $(K_1, 0)$ and $(0, K_2)$ reversed.

In summary, we can conclude that if the competition is light ($a < K_1 / K_2$ and $b < K_2 / K_1$ in (7.3)-(7.4)), then the species populations will approach mutual coexistence as defined by (7.5). If the competition is mutually intense ($a > K_1 / K_2$ and $b > K_2 / K_1$) then one population must be driven to extinction, but it could be either population, depending on the initial populations. If the competition is lopsided ($a < K_1 / K_2$ and $b > K_2 / K_1$, or $a > K_1 / K_2$ and $b < K_2 / K_1$) then one species is driven to extinction and the other is destined to dominate for any nonzero starting populations.

7.4.4 Higher-Dimensional Systems

The techniques we've employed to analyze the competing species model (7.3)-(7.4) can in principle be extended to higher dimensions. The difficulty is that each nullcline for a system of n ODEs is typically an $(n - 1)$ -dimensional surface in \mathbb{R}^n and implicitly defined by an equation of the form

$f_j(x_1, \dots, x_n) = 0$. Even in the case that $n = 3$ the results are difficult to interpret geometrically. But the principle itself is sound: the nullcline for $\dot{x}_j = 0$ divides \mathbf{R}^n into a number of distinct regions; in each region $\dot{x}_j > 0$ or $\dot{x}_j < 0$, so for any solution trajectory in this region x_j is increasing or decreasing.

The stability analysis for fixed points based on linearization, and in particular the Hartman-Grobman theorem, remains valid and useful for local stability analysis (as in Exercise 7.3.5 of the previous section). The trace-determinant analysis and Theorem B.4.1 for determining the nature of the eigenvalues for the linearized system at fixed points can also be generalized to some extent by using the **Routh-Hurwitz Theorem**; see Section 7.6.3 and the reference [100]. Section 7.6 also presents some additional techniques for analyzing systems of ODEs. References [106] and [56] contain even more methods that are useful for analyzing higher-dimensional systems.

In the exercises that follow, keep in mind that the phase portraits—in particular, the nullclines, direction arrows, and typical solution trajectories—can usually be drawn accurately by hand, and there is great value in doing this. Many of these are variations on exercises from Section 7.3.

7.4.5 Exercises

Exercise 7.4.1 Consider the system

$$\begin{aligned}\dot{x}_1 &= -ax_2 + x_2^2 \\ \dot{x}_2 &= x_1 - x_2\end{aligned}$$

where $a > 0$ is some unspecified parameter.

- Show that $(0, 0)$ and (a, a) are the equilibrium solutions for this system.
- Compute the x_1 nullcline and sketch it on pair of x_1x_2 axes that includes a region around the origin. This nullcline should divide the plane into three regions. Determine solution directions (left or right) in each region and sketch appropriate arrows.
- Compute the x_2 nullcline and sketch it on pair of x_1x_2 axes that includes a region around the origin. This nullcline should divide the plane into two regions. Determine solution directions (up or down) in each region and sketch appropriate arrows.
- Compute the Jacobian matrix at each fixed point. Use the trace-determinant techniques of Theorem B.4.1 to show that the fixed point at (a, a) always behaves as a saddle point. Show that for $0 < a < 1/4$ the fixed point at $(0, 0)$ is an asymptotically stable node and that for $a > 1/4$ the fixed point $(0, 0)$ is an asymptotically stable spiral point.
- Use the information from parts (a)-(d) to sketch representative phase portraits for this system in each case that $0 < a < 1/4$ and $a > 1/4$.

Exercise 7.4.2 Sketch a phase portrait for the system (7.28)-(7.29) under the assumption that $\bar{a} < 1$ and $\bar{b} > 1$ (and $\bar{a}\bar{b} \neq 1$) by following these steps:

- Sketch the nullclines for the system with appropriate direction arrows.
- Show that the fixed points are as listed in Reading Exercise 7.4.2, and that the last point there is not physically relevant.
- Compute the Jacobian for the system at each of the three remaining fixed points and determine the stability of each.
- Use the information from parts (a)-(c) to sketch a phase portrait with representative solution trajectories.
- Use your phase portrait to argue that the second species goes extinct for any initial data.

- (f) Repeat (a)-(e) for $\bar{a} > 1$ and $\bar{b} < 1$. What conclusion can you draw in this case?

Exercise 7.4.3 Follow the analysis of Exercise 7.4.2 to sketch a phase portrait for the system (7.28)-(7.29) in each of the cases

- (a) $\bar{a} = 1$ and $\bar{b} < 1$
- (b) $\bar{a} < 1$ and $\bar{b} = 1$
- (c) $\bar{a} = 1$ and $\bar{b} = 1$.

In some cases linearization will fail to determine the stability of some equilibrium solutions. Can you still make a reasonable conclusion about the behavior of the system and the fate of each species?

Exercise 7.4.4 The epidemic model (7.6) is a system of three ODEs in three functions $S(t)$, $I(t)$, and $R(t)$, but it can effectively be considered a system of two differential equations

$$\begin{aligned}\dot{S} &= -aSI \\ \dot{I} &= aSI - bI\end{aligned}\tag{7.37}$$

since the first two equations don't involve R . The parameters a and b are positive. We can analyze (7.37) to determine the behavior of $S(t)$ and $I(t)$, then use this to determine the behavior of $R(t)$ using $\dot{R} = bI$. This can be done without assuming specific values for a and b . We can concern ourselves only with the first quadrant where S and I are nonnegative.

- (a) Show that the S nullcline is given by the coordinate axes $S = 0$ and $I = 0$. Show that the I nullcline is given by the horizontal coordinate axis $I = 0$ and vertical line $S = b/a$.
- (b) Sketch the S nullcline and appropriate arrows to indicate the horizontal motion of any solution.
- (c) Sketch the I nullcline and appropriate arrows to indicate the horizontal motion of any solution.
- (d) Show that the fixed points for this system are exactly those points for which $I = 0$. Thus for this system, the fixed points are not isolated, but form a continuous line, the entire S axis.
- (e) Linearize the system at a typical fixed point $(S_0, 0)$. Show that the eigenvalues for the Jacobian matrix are $\lambda = 0$ and $\lambda = aS_0 - b$. What conclusion does the linearization allow you to make about the stability of each fixed point?
- (f) Sketch a phase portrait for this system using the above information. What conclusion can you make about the long-term behavior of the system—how will the number of susceptible and infected people change over time? What will happen to R ?

Exercise 7.4.5 Consider the damped nonlinear pendulum equation (7.7). An equivalent system

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{g}{L} \sin(x_1) - cx_2\end{aligned}$$

was given in (7.8). Here g and L are both positive and denote gravitational acceleration and the length of the pendulum, respectively, c is a positive damping constant, and $x_1(t)$ and $x_2(t)$ are the angular position and angular velocity of the pendulum with respect to the vertical.

- (a) Show that the x_1 nullcline is the horizontal axis in the x_1x_2 plane. Sketch this nullcline with appropriate arrows above and below the nullcline, to indicate solution motion.
- (b) Assume $c > 0$; show that the x_2 nullcline is the sine curve with graph $x_2 = -\frac{g}{cL} \sin(x_1)$. Carefully sketch this curve on the range $-4\pi \leq x_1 \leq 4\pi$, taking note of the fact that c, g , and L are all positive.
- (c) Show that the fixed points for this system are all of the form $x_1 = m\pi, x_2 = 0$ where m is an integer. Interpret this result physically: What configuration is the pendulum in if $(x_1, x_2) = (0, 0)$? What configuration is the pendulum in if $(x_1, x_2) = (\pi, 0)$? What configuration is the pendulum in if $(x_1, x_2) = (2\pi, 0)$?
- (d) Linearize the system (compute the Jacobian matrix) at the equilibrium solution at $x_1 = x_2 = 0$ (when the pendulum is hanging straight down and is motionless). Compute the trace and determinant of $\mathbf{J}(0, 0)$ and use Theorem B.4.1 to show that $\det(\mathbf{J}(0, 0)) > 0$. Use the analysis of Section B.4 to conclude that $(0, 0)$ is always an asymptotically stable fixed point. Under what conditions on g, L , and c will it be an asymptotically stable node? Under what conditions on g, L , and c will it be an asymptotically stable spiral point? Argue that your analysis holds for any fixed point of the form $x_1 = 2k\pi$ and $x_2 = 0$, where k is an integer.
- (e) Linearize the system at the equilibrium solution at $x_1 = \pi, x_2 = 0$ (when the pendulum is perfectly balanced upside down and is motionless). Compute the trace and determinant of $\mathbf{J}(0, 0)$ and use Theorem B.4.1 to show that $\det(\mathbf{J}(0, 0)) < 0$. Use the analysis of Section B.4 to conclude that $(0, 0)$ is always a saddle point and hence unstable. Argue that your analysis holds for any fixed point of the form $x_1 = (2k+1)\pi$ and $x_2 = 0$, where k is an integer.
- (f) Use the information from parts (a)-(e) to sketch representative phase portraits in each case $(c/2)^2 - g/L \geq 0$ and $(c/2)^2 - g/L < 0$. Include the range $-4\pi \leq x_1 \leq 4\pi$ in your phase portraits.

Exercise 7.4.6 Consider the three-dimensional system

$$\begin{aligned}\dot{x} &= xz + az \\ \dot{y} &= 1 - y + z \\ \dot{z} &= -x - z\end{aligned}$$

where a is a real constant. Note that the trace-determinant analysis of Theorem B.4.1 for fixed point stability will not work here, since the system is three-dimensional. For stability analysis we will have to examine the eigenvalues of the Jacobian matrix at the relevant fixed point directly.

- (a) Show that the fixed points for this system are $(0, 1, 0)$ and $(-a, a+1, a)$.
- (b) Compute the Jacobian for the system and show that at the fixed point $(0, 1, 0)$ the relevant eigenvalues are

$$\lambda_1 = -1, \quad \lambda_2 = -1/2 + \sqrt{1-4a}/2, \quad \lambda_3 = -1/2 - \sqrt{1-4a}/2.$$

Argue that this fixed point is unstable for $a < 0$, asymptotically stable for $a > 0$ and some kind of three dimensional spiral point for $a > 1/4$.

- (c) Use the Jacobian to show that at the fixed point $(-a, a+1, a)$ the relevant eigenvalues are

$$\lambda_1 = a, \quad \lambda_2 = -1, \quad \lambda_3 = -1.$$

Argue that this fixed point is asymptotically stable for $a < 0$ and unstable for $a > 0$.

- (d) Solve the system numerically for a variety of initial data in the case that $a > 0$ and $a < 0$ and plot the results. Do the graphs support your conclusions in parts (b) and (c)?

7.5 Numerical Methods for Systems of First Order ODE's

7.5.1 Extending Basic Numerical Methods to Systems

Motivation

In Sections 7.1 to 7.4 we developed qualitative techniques for analyzing systems of ODEs, for example, the competing species equations (7.3)-(7.4), or an SIR model as in (7.6), or the nonlinear pendulum governed by (7.7) (after converting to a first-order system (7.8)). In each case the methods that were developed let us make conclusions about the qualitative and long-term behavior of the relevant systems. But what if we need quantitative information about the systems of interest? Since nonlinear systems of ODEs are rarely solvable in any analytical form, we must use approximate numerical methods. These methods are the focus of this section.

Vector-Valued Formulation for Systems

Consider a system of ODEs of the form (6.3) for functions $x_1(t), \dots, x_n(t)$. As noted in Section 7.1, this system can be expressed in the form

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t)) \quad (7.38)$$

where $\mathbf{x}(t) = \langle x_1(t), x_2(t), \dots, x_n(t) \rangle$ and

$$\mathbf{f}(t, \mathbf{x}) = \langle f_1(t, \mathbf{x}), f_2(t, \mathbf{x}), \dots, f_n(t, \mathbf{x}) \rangle \quad (7.39)$$

for functions f_1, \dots, f_n . Note that \mathbf{f} accepts the scalar t and vector \mathbf{x} as arguments and returns an n -dimensional vector.

As discussed in Section 6.1, higher-order ODEs and systems of higher-order ODEs can be converted to first-order systems of ODEs. As a result, any numerical method for first-order systems is applicable to higher-order ODEs. This is a common approach to solving higher-order ODEs numerically.

Example 7.8 The damped nonlinear pendulum equation (4.132) that governs the angular position $\theta = \theta(t)$ with respect to the vertical of a pendulum of length L swinging in a gravitational field with positive downward acceleration g is

$$\ddot{\theta} + c\dot{\theta} + g \sin(\theta)/L = 0$$

where c is a positive damping constant. If $x_1 = \theta$ and $x_2 = \dot{\theta}$ then this second-order ODE is equivalent to the first-order system

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -g \sin(x_1)/L - cx_2.\end{aligned}$$

This system can be written in the form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ where

$$\mathbf{f}(\mathbf{x}) = \langle x_2, -g \sin(x_1)/L - cx_2 \rangle.$$

Euler's Method for Systems

In the remainder of this section it will be convenient to use superscripts rather than subscripts for the iterates produced by a numerical ODE solver. Thus we will write $\mathbf{x}^1, \dots, \mathbf{x}^k, \dots$ for the estimates of the solution to an ODE $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$ at times t_1, \dots, t_k, \dots (we continue to use subscripts on t). This avoids confusing an iterate \mathbf{x}^k with a component x_k of the vector-valued function $\mathbf{x}(t)$. Each iterate \mathbf{x}^k is an n -dimensional vector. We suppose the system (7.38) has initial data $\mathbf{x}(t_0) = \mathbf{x}^0$ for some initial time t_0 and n -dimensional vector \mathbf{x}^0 . Assume for the moment that the step size is fixed, so that $t_{k+1} = t_k + h$ for some step size h .

With this notation in place, the extension of Euler's method, the improved Euler's method, and the Runge-Kutta fourth-order method (RK4) to this setting is straightforward. Suppose we have produced estimates $\mathbf{x}^1, \dots, \mathbf{x}^k$ for $\mathbf{x}(t_1), \dots, \mathbf{x}(t_k)$. Euler's method produces \mathbf{x}^{k+1} (an estimate of $\mathbf{x}(t_{k+1})$, where $t_{k+1} = t_k + h$) according to

$$\mathbf{x}^{k+1} = \mathbf{x}^k + h\mathbf{f}(t_k, \mathbf{x}^k). \quad (7.40)$$

The step from $t = t_k$ to $t = t_{k+1}$ in (7.40) is tangent line extrapolation in the form $x_j(t_k + h) \approx x_j(t_k) + h\dot{x}_j(t_k)$, but applied to each component x_j of \mathbf{x} , for $j = 1$ to $j = n$, with $f_j(t, \mathbf{x}^k)$ in (7.39) used to estimate $\dot{x}_j(t_k)$.

■ **Example 7.9** Consider the nonlinear system

$$\begin{aligned}\dot{x}_1 &= x_2 - x_1 x_2 + \sin(t) \cos(t) \\ \dot{x}_2 &= 2x_1 - x_2 + \cos(t) - 3 \sin(t)\end{aligned} \quad (7.41)$$

with $x_1(0) = 0$ and $x_2(0) = 1$. In this example the equations have been carefully rigged so that the true solution is $x_1(t) = \sin(t)$ and $x_2(t) = \cos(t)$, so we can examine how Euler's method performs, especially as the step size gets smaller.

The system (7.41) can be expressed in the form (7.38) by taking (in column-vector notation)

$$\mathbf{f}(t, \mathbf{x}) = \begin{bmatrix} x_2 - x_1 x_2 + \sin(t) \cos(t) \\ 2x_1 - x_2 + \cos(t) - 3 \sin(t) \end{bmatrix}$$

where $\mathbf{x} = \langle x_1, x_2 \rangle$. The initial data is $\mathbf{x}^0 = \langle 0, 1 \rangle$ at time $t_0 = 0$ and

$$\mathbf{f}(t_0, \mathbf{x}^0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

We begin with step size $h = 0.5$. From (7.40) it follows that $\mathbf{x}^1 = \mathbf{x}^0 + h\mathbf{f}(t_0, \mathbf{x}^0)$ or

$$\mathbf{x}^1 = \langle 0, 1 \rangle + 0.5 \langle 1, 0 \rangle = \langle 0.5, 1 \rangle.$$

A second iteration to compute \mathbf{x}^2 (noting $t_1 = t_0 + h = 0.5$) yields

$$\mathbf{x}^2 = \langle 0.5, 1 \rangle + 0.5\mathbf{f}(t_1, \mathbf{x}^1) \approx \langle 0.960, 0.720 \rangle.$$

A graph of the components of these iterates using step size $h = 0.5$ out to time $t = 5$ is shown in Figure 7.17, along with the graph of the true solution components.

As in the scalar case, a smaller step size yields better results. Figure 7.18 shows the results obtained by using a step size of $h = 0.01$. ■

Reading Exercise 7.5.1 Continue with step size $h = 0.5$ to compute \mathbf{x}^3 (an approximation to $\mathbf{x}(1.5)$) and \mathbf{x}^4 (an approximation to $\mathbf{x}(2)$) for the system (7.41). Compare your answers to the true values (recall $\mathbf{x}(t) = \langle \sin(t), \cos(t) \rangle$).

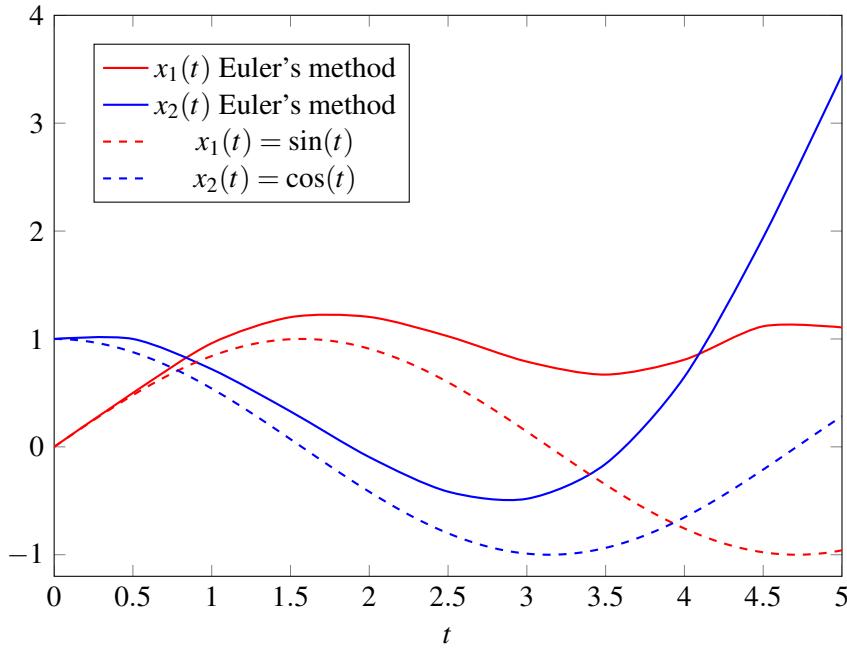


Figure 7.17: Euler's method estimates and true solution components to system (7.41) using step size $h = 0.5$.

The Improved Euler's and RK4 Method for Systems; Error Control

The formulas (3.16) to (3.18) for the improved Euler's method generalize to systems. We march the solution estimate from time $t = t_k$ to time t_{k+1} according to

$$\mathbf{x}^{k+1} = \mathbf{x}^k + h \left(\frac{\mathbf{f}(t_k, \mathbf{x}^k) + \mathbf{f}(t_k + h, \mathbf{w})}{2} \right) \quad (7.42)$$

where $\mathbf{w} = \mathbf{x}^k + h\mathbf{f}(t_k, \mathbf{x}^k)$; note that \mathbf{w} is the Euler estimate of $\mathbf{x}(t_{k+1})$, just as in the scalar case.

The Runge-Kutta formulas (3.22) also generalize as

$$\begin{aligned} \mathbf{m}^1 &= \mathbf{f}(t_k, \mathbf{x}^k), \\ \mathbf{m}^2 &= \mathbf{f}(t_k + h/2, \mathbf{x}^k + h\mathbf{m}^1/2), \\ \mathbf{m}^3 &= \mathbf{f}(t_k + h/2, \mathbf{x}^k + h\mathbf{m}^2/2), \\ \mathbf{m}^4 &= \mathbf{f}(t_k + h, \mathbf{x}^k + h\mathbf{m}^3), \\ \mathbf{m} &= \frac{1}{6}(\mathbf{m}^1 + 2\mathbf{m}^2 + 2\mathbf{m}^3 + \mathbf{m}^4), \\ \mathbf{x}^{k+1} &= \mathbf{x}^k + h\mathbf{m}. \end{aligned} \quad (7.43)$$

Notice how naturally the scalar quantities are replaced by appropriate vector quantities.

■ **Example 7.10** Let's consider the system (7.41) of Example 7.9, and use each of Euler's method, the improved Euler's method, and the Runge-Kutta fourth-order method to estimate $\mathbf{x}(5)$. In particular, let's examine how the error in each method depends on the step size h . The error will be measured as $\|\mathbf{x}^n - \mathbf{x}(5)\|$ where \mathbf{x}^n is the numerical iterate corresponding to $t = 5$ for each method, $\mathbf{x}(5) = \langle \sin(5), \cos(5) \rangle$ is the exact solution, and the notation $\|\mathbf{y}\|$ means the usual Euclidean norm of a vector \mathbf{y} (the square root of the sum of the squares of the components of \mathbf{y}). The results are shown in Table 7.2 for a variety of step sizes.

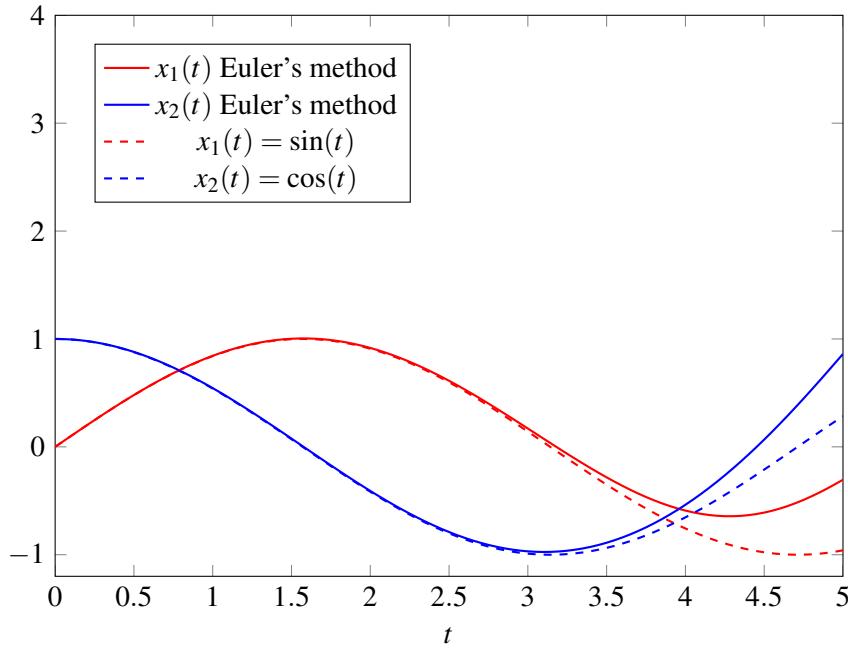


Figure 7.18: Euler's method estimates and true solution components to system (7.41) using step size $h = 0.01$.

Step size h	Euler Error	Improved Euler Error	RK4 Error
1.0	4.0425	3.3222	0.1485
10^{-1}	2.8125	0.1248	7.092×10^{-5}
10^{-2}	0.8726	1.223×10^{-3}	8.549×10^{-9}
10^{-3}	0.1118	1.220×10^{-5}	8.700×10^{-13}
10^{-4}	0.0115	1.217×10^{-7}	8.716×10^{-17}

Table 7.2: Error for system (7.41) with Euler, improved Euler, and RK4 methods with indicated step sizes at time $t = 5$, error measured as $\|\mathbf{x}^n - \mathbf{x}(5)\|$. (Computations done to 20 digit accuracy in Maple.)

As in the scalar case, Euler's method is first order, that is, the error is proportional to the step size h , at least once the step size is sufficiently small. The improved Euler's method is second order, error proportional to h^2 , and the RK4 method is fourth-order, error proportional to h^4 . ■

The same general philosophy and procedures of Section 3.3.2 are also applicable to the numerical solution of systems of ODEs. An estimate of the local truncation error at each step can be made and the method's step size adjusted to control this error. See [76] for more on this topic.

Reading Exercise 7.5.2 Consider the linear system of ODEs $\dot{x}_1 = x_1 - x_2$, $\dot{x}_2 = 6x_1 - 4x_2$ with initial data $x_1(0) = 1$ and $x_2(0) = 0$.

- Formulate the system as $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$. Find the analytical solution $\mathbf{x}(t)$ to the system and compute $\mathbf{x}(2)$.
- Estimate $\mathbf{x}(2)$ by using Euler's method with step size $h = 0.1$. Compute the error $\|\mathbf{x}(2) - \mathbf{x}^n\|$, where \mathbf{x}^n is your estimate of $\mathbf{x}(2)$.
- Repeat part (b) using the improved Euler's method with step size $h = 0.1$.
- Repeat part (b) using the RK4 method with step size $h = 0.1$.
- Repeat (b)-(d) using $h = 0.01$. Are the results in accordance with the order of each method?

7.5.2 Stiff Systems of ODEs

Stiffness is a particular property that some ODEs and systems of ODEs possess that makes them challenging to solve numerically. Stiffness often reflects an essential disparity of scale in a physical system; one variable in a physical system may evolve in time on a relatively slow scale, for example, seconds, while another aspect of the same system evolves much more rapidly, perhaps on a scale of microseconds. It might be the slower scale behavior of the system that interests us, but the rapidly changing aspect of the system's behavior demands the attention of the numerical method and makes long-term simulations computationally expensive. Stiff systems are surprisingly common, but rarely talked about in introductory ODE courses.

In this section we consider a variety of straightforward examples that illustrate the essential idea behind stiffness. The goal here is not a complete introduction to the theory of numerical methods for stiff systems of ODEs, but rather an intuitive and practical guide to what stiff ODEs are and some ideas on how to handle them. There are also some physical examples and exercises to illustrate how stiff systems come about.

A Motivational Example

Let us revisit the double spring-mass system of Example 6.2 as illustrated in Figure 6.2. A standard model based on Hooke's law and viscous damping led to equations (6.8)-(6.9) or the equivalent first-order system $\dot{\mathbf{w}} = \mathbf{Aw}$ where $\mathbf{w}(t)$ is a vector-valued function with four components and

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -(k_1 + k_2)/m_1 & -c_1/m_1 & k_2/m_1 & 0 \\ 0 & 0 & 0 & 1 \\ k_2/m_2 & 0 & -k_2/m_2 & -c_2/m_2 \end{bmatrix}. \quad (7.44)$$

Here m_1 and m_2 quantify the masses in the system, k_1 and k_2 are the stiffness constants for the springs, and c_1 and c_2 are the damping constants; refer to Figure 6.2. The function $w_1(t)$ is the position of the first mass, $w_2(t)$ is the velocity of the first mass, $w_3(t)$ is the position of the second mass, and $w_4(t)$ is the velocity of the second mass.

Consider the case in which $m_1 = m_2 = 1$ kg, $k_1 = 1$ newton per meter, $k_2 = 10^4$ newton per meter, and $c_1 = c_2 = 0.1$ newtons per meter per second. Note that spring 2 is very stiff compared to spring 1. Let us substitute these values into the matrix \mathbf{A} in (7.44) and solve the resulting linear system of four ODEs analytically using the eigenvalue/eigenvector techniques of Section 6.2.2, with initial data $w_1(0) = -1.51060655$, $w_2(0) = 0.9999450017$, $w_3(0) = -1.510689634$, and $w_4(0) = 1$ (these initial conditions are chosen to illustrate a point, explained below). The solution is, to ten significant figures,

$$\mathbf{w}(t) = e^{-t/20} \begin{bmatrix} -1.51060655 \cos(\alpha t) + 1.31061355 \sin(\alpha t) \\ 0.99994500 \cos(\alpha t) + 0.99994500 \sin(\alpha t) \\ -1.51068963 \cos(\alpha t) + 1.31068563 \sin(\alpha t) \\ \cos(\alpha t) + \sin(\alpha t) \end{bmatrix}$$

where $\alpha \approx 0.70532971$. Each component of $\mathbf{w}(t)$ is an exponentially tapered periodic function with a period of approximately $2\pi/\alpha \approx 8.908$ seconds. The components $w_1(t)$ and $w_2(t)$ of $\mathbf{w}(t)$ are graphed in Figure 7.19 on the interval $0 \leq t \leq 50$. The graph of $w_3(t)$ is almost identical to that of $w_1(t)$ and the graph of $w_4(t)$ to that of $w_2(t)$, so these aren't shown.

Although this system is analytically solvable, let's construct a numerical solution using an RK4 solver. Given the time scale of the solutions shown in Figure 7.19, a step size of 0.1 or smaller should certainly track the solution well. The left panel of Figure 7.20 shows the estimate of the first component $w_1(t)$ obtained using step size $h = 0.02$ for the system, on the interval $0 \leq t \leq 2$. The solution is accurate to within about 2.25×10^{-6} over this time interval, so to visual accuracy

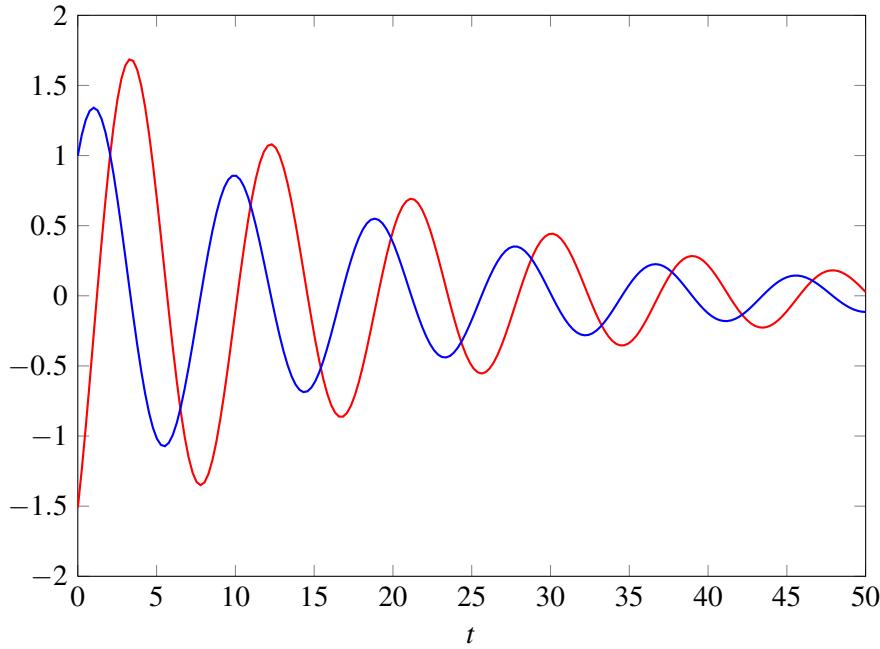


Figure 7.19: Solution components $w_1(t)$ (red) and $w_2(t)$ (blue) to system $\dot{\mathbf{w}} = \mathbf{Aw}$ with $m_1 = m_2 = 1$, $c_1 = c_2 = 0.1$, $k_1 = 1$, $k_2 = 10^4$.

this may as well be the exact solution. But the right panel in Figure 7.20 shows the numerical RK4 solution with the slightly larger time step $h = 0.0205$.

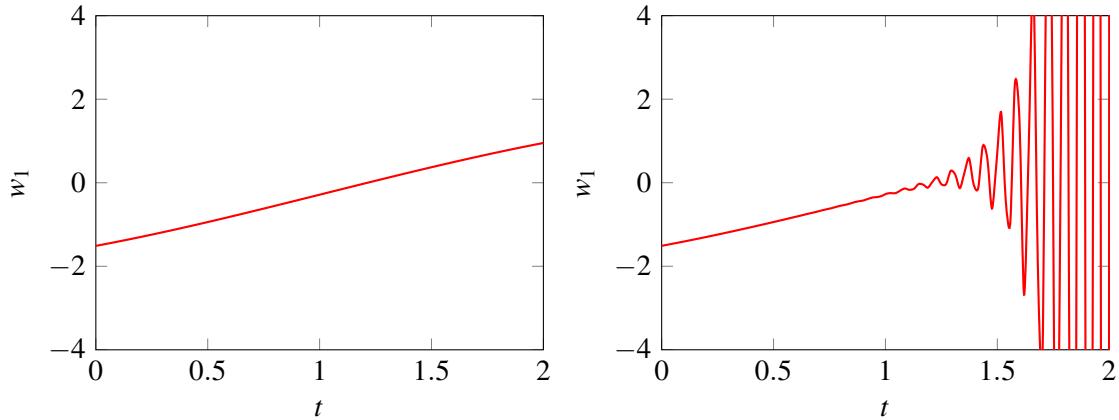


Figure 7.20: Left panel: Solution component $w_1(t)$, step size $h = 0.02$. Right panel: same, but with step size $h = 0.0205$.

Things start off well in the right panel, but ultimately the solution with step size $h = 0.0205$ is disastrous. If a step size of $h = 0.02$ tracks the solution so accurately, why does the tiny increase to $h = 0.0205$ go so wrong? The heart of the problem is that this system of differential equations is *stiff*, a concept we will now explore. We will return to this double spring-mass system in Exercise 7.5.10.

A Scalar Example

The concept of stiffness is most fully illustrated using a system of two or more ODEs, but a principle insight can be gained from a scalar example. Consider the classic exponential decay ODE

$$\dot{x}(t) = -\lambda x(t) \quad (7.45)$$

with initial condition $x(0) = 1$, where λ is a positive constant. The analytical solution is $x(t) = e^{-\lambda t}$. Let us examine how Euler's method with step size h behaves when applied to this system. You may have already carried out much of this analysis if you did Exercise 3.1.7; see also Exercises 3.2.6 and 3.3.6. With initial iterate $x^0 = 1$ and $t_k = kh$, Euler's method with step size h marches the iterates out in time as $x^{k+1} = x^k - h\lambda x^k$ or

$$x^{k+1} = (1 - h\lambda)x^k. \quad (7.46)$$

Thus $x^0 = 1, x^1 = 1 - h\lambda, x^2 = (1 - h\lambda)^2$, and more generally $x^k = (1 - h\lambda)^k$ (the superscripts on $1 - h\lambda$ are actually exponents, but as previously noted, x^k is the k th iterate in this procedure).

As we have seen, Euler's method tracks the true solution more accurately when h is close to 0. In each panel of Figure 7.21 we illustrate the case in which $\lambda = 1$ so that (7.45) becomes $\dot{x}(t) = -x(t)$ with initial condition $x(0) = 1$ and solution $x(t) = e^{-t}$. With a small step size the iterates defined by (7.46) approximate the solution well, as shown in the upper left panel with $h = 0.25$. The Euler iterates are shown as a solid red curve and the true solution as a blue dashed curve, both superimposed on a direction field for $\dot{x}(t) = -x(t)$. The iterates and true solution are almost indistinguishable. The upper right panel shows the result when $h = 0.8$. It's clear that this is a poor approximation to the true solution, though at least the Euler iterates decay to zero, just as the true solution does.

The lower left panel of Figure 7.21 shows the result with step size $h = 1.5$. In this case the Euler iterates don't even remain positive like the true solution. Take careful note of how the Euler iterates are constructed in each case, and in particular the lower left panel: at each point $t = t_k, x = x^k$ the direction field is computed and used to linearly extrapolate the solution forward in time t by a distance $h = 1.5$. The result is that when extrapolating from $x^0 = 1$ to x^1 , the tangent line actually heads into $x < 0$ territory, then back to $x > 0$, then back to $x < 0$, etc.

The bottom right panel of Figure 7.21 illustrates what happens with $h = 2.2$. In this case the iterates don't even decay to zero as they should, but rather grow geometrically. We'll analyze the situation more carefully below.

Reading Exercise 7.5.3 Use (7.46) with $\lambda = 1$ to show that the Euler iterates for $\dot{x}(t) = -x(t)$ with $x(0) = 1$ with step size $h = 1.5$ are given by $x^k = (-1/2)^k$. Show that if $h = 2.2$ then $x^k = (-1.2)^k$.

Some Stability Bounds

For the ODE (7.45) with $x(0) = 1$ the Euler iterates are given by $x^k = (1 - h\lambda)^k$. Under what condition on the step size h is it (at least) guaranteed that these iterates decay to zero like the true solution? This requires that $|1 - h\lambda| < 1$ or equivalently, $-1 < 1 - h\lambda < 1$. A bit of algebra shows that this is equivalent to $-2 < -h\lambda < 0$ or, if we divide through by $-\lambda$ and assume $\lambda > 0$ (so the inequalities are reversed),

$$0 < h < 2/\lambda \quad (7.47)$$

as a condition for the Euler iterates to converge to zero. It's easy to see that this analysis holds for any initial condition, not just $x(0) = 1$.

Reading Exercise 7.5.4 If the Euler iterates $x^k = (1 - h\lambda)^k$ are to decay to zero and also remain positive then we need $0 < 1 - h\lambda < 1$. Show this requires h to satisfy the more stringent condition $0 < h < 1/\lambda$ (assume λ is positive). Is this in accordance with Figure 7.21?

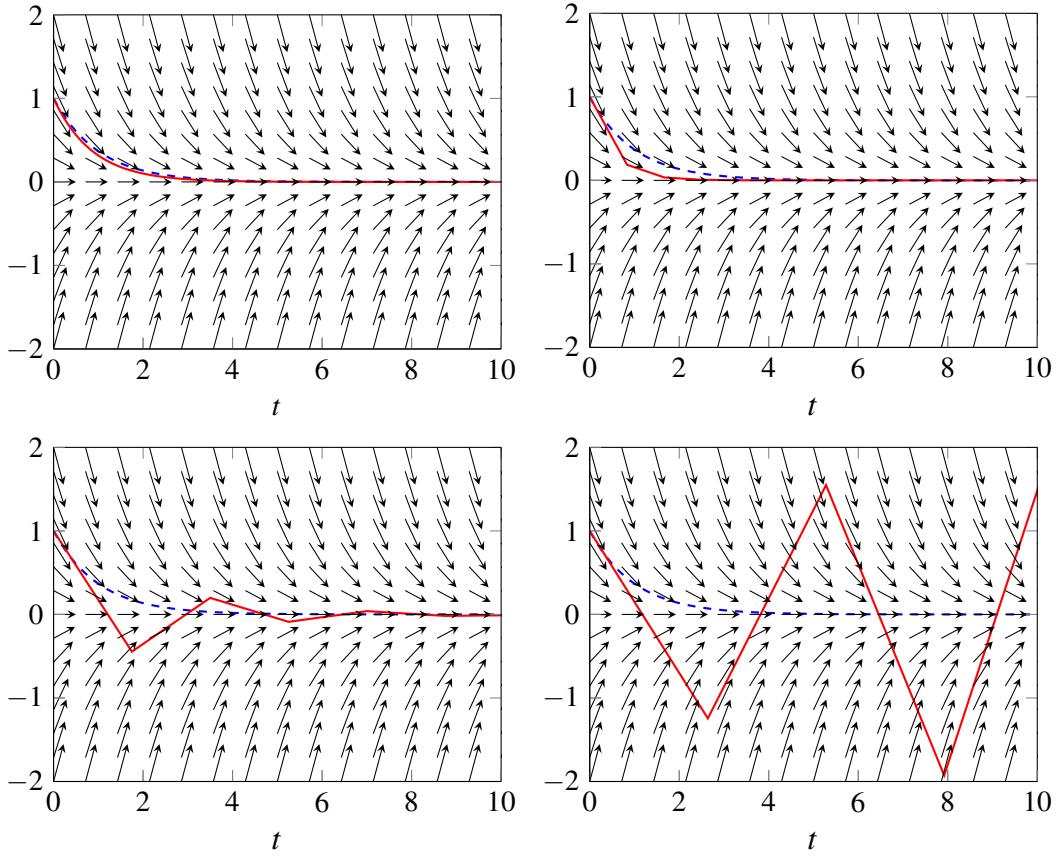


Figure 7.21: Top left panel: Estimated solution to $x'(t) = -x(t)$, step size $h = 0.25$; Euler iterates as solid red lines, true solution $x(t) = e^{-t}$ as dashed blue graph. Top right panel: same, but with step size $h = 0.8$. Bottom left panel: same, but with step size $h = 1.5$. Bottom right panel: same, but with step size $h = 2.2$.

The moral of the story is this: For an ODE like (7.45), Euler's method will perform poorly with a large step size and the iterates may oscillate or even grow without bound if the step size is too large in relation to λ . Similar conclusions hold for the improved Euler's and RK4 methods of Chapter 3, and many other similar methods that extrapolate the solution from t_k to t_{k+1} .

A Stiff System of ODEs I

Let's start by considering a very simple autonomous linear system of ODEs of the form $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ where $\mathbf{x}(t) = \langle x_1(t), x_2(t) \rangle$ and \mathbf{A} is a 2×2 diagonal matrix

$$\mathbf{A} = \begin{bmatrix} -\lambda_1 & 0 \\ 0 & -\lambda_2 \end{bmatrix}$$

with diagonal entries $-\lambda_1$ and $-\lambda_2$ for some positive constants λ_1 and λ_2 . Written out explicitly this system is

$$\begin{aligned} \dot{x}_1 &= -\lambda_1 x_1 \\ \dot{x}_2 &= -\lambda_2 x_2. \end{aligned} \tag{7.48}$$

This system consists of two decoupled scalar ODEs and each equation can be solved independently of the other. Nonetheless, handling (7.48) as a system will serve to illustrate an essential facet

of what constitutes a stiff system of ODEs. The analytical solution is $x_1(t) = x_1(0)e^{-\lambda_1 t}$ and $x_2(t) = x_2(0)e^{-\lambda_2 t}$.

Let's use Euler's method to solve the system (7.48) with initial time $t_0 = 0$ and initial conditions $x_1(0) = 1$ and $x_2(0) = 1$. Let h be the step size, define $t_k = kh$, and use x_1^k and x_2^k to denote the iterates produced by Euler's method, as approximations to $x_1(t_k)$ and $x_2(t_k)$, so $\mathbf{x}^k = \langle x_1^k, x_2^k \rangle$. Euler's method (7.40) for the system (7.48) takes the form

$$\begin{aligned} x_1^{k+1} &= (1 - \lambda_1 h)x_1^k \\ x_2^{k+1} &= (1 - \lambda_2 h)x_2^k \end{aligned} \quad (7.49)$$

which is merely Euler's method with step size h applied to each ODE $\dot{x}_1 = -\lambda_1 x_1$ and $\dot{x}_2 = -\lambda_2 x_2$ separately. From (7.49) it's easy to see that $x_1^k = (1 - \lambda_1 h)^k$ and $x_2^k = (1 - \lambda_2 h)^k$. In brief, the iterate components x_1^k and x_2^k for Euler's method applied to the system (7.48) are the iterates obtained from Euler's method applied to the scalar equations $\dot{x}_1 = -\lambda_1 x_1$ and $\dot{x}_2 = -\lambda_2 x_2$ separately.

Based on the discussion for the scalar equation (7.45), the Euler iteration for the x_1 equation will decay to zero (as $x_1(t)$ itself does) if and only if $0 < h < 2/\lambda_1$. The Euler iteration for the x_2 equation will decay to zero (as does $x_2(t)$) if and only if $0 < h < 2/\lambda_2$. As a result, the iterates \mathbf{x}^k will decay to $(0, 0)$ like the true solution $\mathbf{x}(t) = \langle e^{-\lambda_1 t}, e^{-\lambda_2 t} \rangle$ precisely when $0 < h < 2/\max(\lambda_1, \lambda_2)$. It is thus the larger of λ_1 and λ_2 that dictates the largest step size we can take and still have the iterates \mathbf{x}^k decay to zero. This does not guarantee that the numerical solution is accurate, but it is a necessary condition if we hope to track the true solution with any precision. If $h > 2/\max(\lambda_1, \lambda_2)$ then the iterates actually grow geometrically, rather than decay. It's worth noting that the eigenvalues of the matrix \mathbf{A} are exactly $-\lambda_1$ and $-\lambda_2$.

A geometric interpretation of the situation is useful. In the left panel of Figure 7.22 we show the trajectory in the $x_1 x_2$ phase plane for the solution to (7.48) with $\lambda_1 = 1$ and $\lambda_2 = 10$, along with a numerically approximated trajectory obtained using Euler's method with step size $h = 0.05$. This numerically computed trajectory, shown as the solid red curve, is a reasonable approximation to the true trajectory, shown as the dashed blue curve (and parameterized by $x_1(t) = e^{-t}, x_2(t) = e^{-10t}$); note that the initial point for this trajectory is $x_1 = 1, x_2 = 1$ on the right side of the figure. Both curves are shown superimposed on the direction field for the system, with the vectors scaled to a shorter length for visual appeal. The right panel shows the situation when $h = 0.205$, and the solution here is clearly unacceptable.

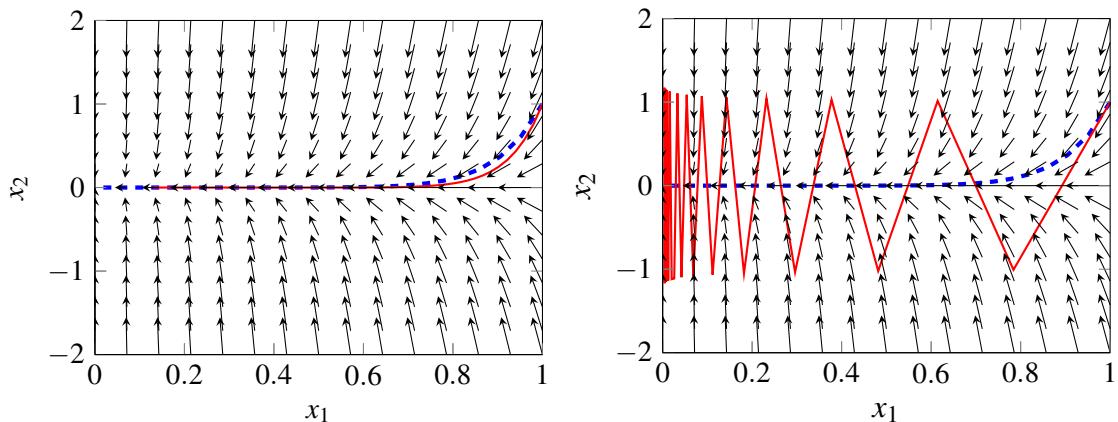


Figure 7.22: Left panel: True solution trajectory for the system (7.48) (dashed blue curve) and Euler's method approximation (solid red curve), step size $h = 0.05$. Right panel: same, but with step size $h = 0.205$.

Note that at each iteration Euler's method extrapolates along a line tangent to the direction field vector $\langle -x_1, -10x_2 \rangle$ at the relevant point, but in the right panel of Figure 7.22 the step size is such that the extrapolation greatly overshoots the true solution. The numerical iterates get farther and farther away from the true trajectory.

Reading Exercise 7.5.5 Explain the behavior of the numerical solutions to the system (7.48) in the left and right panels of Figure 7.22 (step sizes $h = 0.05$ and $h = 0.205$, respectively) in light of the bound (7.47).

A Stiff System of ODEs II

The system (7.48) is a bit trivial with regard to the application of Euler's method. Because the system is decoupled, it would be easier to apply Euler's method (or any solution method, analytical or numerical) to each equation in the system (7.48) separately. Thus it would make more sense to solve $\dot{x}_1 = -\lambda_1 x_1$ with an appropriate step size based on λ_1 and $\dot{x}_2 = -\lambda_2 x_2$ with an appropriate step size based on λ_2 .

Let's consider a more typical example in which the ODEs are coupled. The ODEs of interest are

$$\begin{aligned}\dot{x}_1 &= -56x_1 + 55x_2 \\ \dot{x}_2 &= 44x_1 - 45x_2.\end{aligned}\tag{7.50}$$

This system can be expressed as $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ where

$$\mathbf{A} = \begin{bmatrix} -56 & 55 \\ 44 & -45 \end{bmatrix}.$$

For later reference, the eigenvalues and eigenvectors of \mathbf{A} are

$$\lambda_1 = -1, \quad \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \text{and} \quad \lambda_2 = -100, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ -4/5 \end{bmatrix}.\tag{7.51}$$

A general solution to the system is thus given by

$$\begin{aligned}\mathbf{x}(t) &= c_1 e^{\lambda_1 t} \mathbf{v}_1 + c_2 e^{\lambda_2 t} \mathbf{v}_2 \\ &= c_1 e^{-t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 e^{-100t} \begin{bmatrix} 1 \\ -4/5 \end{bmatrix}.\end{aligned}\tag{7.52}$$

The general solution makes it clear that a typical solution to $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ consists of a sum of a relatively slowly decaying piece corresponding to e^{-t} and a rapidly decaying piece stemming from the e^{-100t} term. All solutions decay to the origin as t increases.

Consider the solution with initial conditions $x_1(0) = 1$ and $x_2(0) = 1.09$. The exact solution is easily obtained from (7.52). The initial conditions dictate $c_1 = 1.05$ and $c_2 = 0.05$ and the solution is

$$\mathbf{x}(t) = 1.05e^{-t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 0.05e^{-100t} \begin{bmatrix} 1 \\ -4/5 \end{bmatrix}.$$

Notice that the e^{-100t} , which is much smaller in magnitude than the term involving e^{-t} even at time $t = 0$, decays very rapidly.

Let's apply Euler's method to solve this system. Given that the solution changes on a time scale dictated primarily by e^{-t} , we might hope that a step size of $h = 0.1$ or smaller should do a good job of tracking the solution. Figure 7.23 shows a direction field for this ODE system, along with a plot of the trajectory of the true solution (shown as a dashed red curve) for $0.8 \leq x_1 \leq 1.1$,

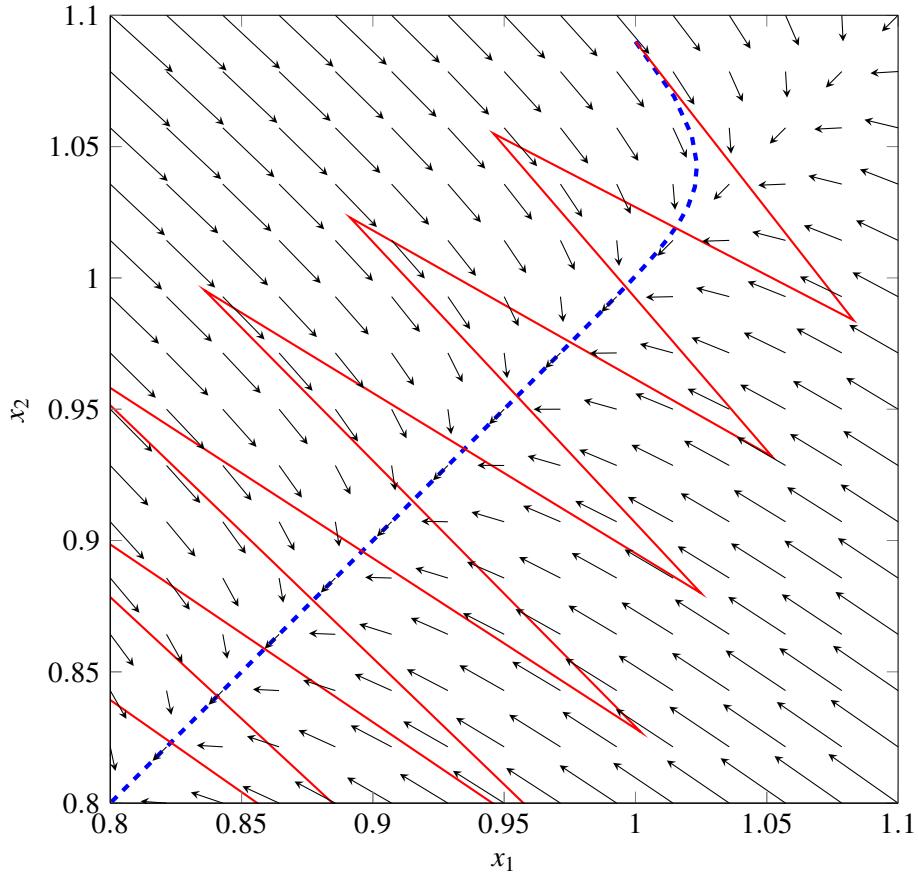


Figure 7.23: Direction field and solution trajectory for the system (7.50) with step size $h = 0.021$ (solution trajectory in dashed blue) and Euler iterates with step size $h = 0.021$ (shown in solid red).

$0.8 \leq x_2 \leq 1.1$, and the resulting trajectory for the numerical solution with step size $h = 0.021$. Clearly this step size is a disaster, even though the dominant term e^{-t} in the solution suggests that this step size should work well. But each iteration of Euler's method with $h = 0.021$ puts us *further* from the actual solution trajectory. The situation is similar to that of the right panel of Figure 7.22.

To understand this phenomena, consider a typical point $\mathbf{x} = \langle x_1, x_2 \rangle$ in the phase plane; refer again to Figure 7.23. The direction field vector at \mathbf{x} is given by the product \mathbf{Ax} . The matrix \mathbf{A} has two linearly independent eigenvectors given in (7.51), so we can express \mathbf{x} as $\mathbf{x} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2$ for some constants c_1 and c_2 . The direction field vector \mathbf{Ax} is given by

$$\begin{aligned}\mathbf{Ax} &= \mathbf{A}(c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2) \\ &= c_1 \mathbf{Av}_1 + c_2 \mathbf{Av}_2 \\ &= c_1 \lambda_1 \mathbf{v}_1 + c_2 \lambda_2 \mathbf{v}_2.\end{aligned}\tag{7.53}$$

If \mathbf{x} happens to lie on the line $x_2 = x_1$ then \mathbf{x} is a multiple of \mathbf{v}_1 and so $c_2 = 0$, but if \mathbf{x} is not on the line $x_2 = x_1$ then $c_2 \neq 0$. Since λ_2 is much larger in magnitude than λ_1 the term $c_2 \lambda_2 \mathbf{v}_2$ in (7.53) likely dominates \mathbf{Ax} . That is, $\mathbf{Ax} \approx c_2 \lambda_2 \mathbf{v}_2$ and \mathbf{Ax} has a large magnitude unless c_2 is close to zero (\mathbf{x} is close to $x_2 = x_1$). This large component of the direction field vector \mathbf{Ax} in the direction of the eigenvector \mathbf{v}_2 causes the numerical solver to overshoot the true trajectory, and in fact in Figure 7.23 the iterates oscillate farther and farther from the correct trajectory, which is asymptotic to the line $x_2 = x_1$.

The heart of the problem in this example is the large disparity in the magnitude of the eigenvalues

of \mathbf{A} . The portion of the solution (7.52) corresponding to the -1 eigenvalue or e^{-t} term dominates the actual solution in most cases, at least after a very short time, and the solution decays smoothly and gradually to zero along the line $x_2 = x_1$. But the portion of the solution corresponding to the eigenvalue of -100 is what limits the size of the steps we can take, no matter how negligible that portion of the solution might be. A bound like (7.47) still applies, where here $\lambda = 100$ is the magnitude of the largest magnitude (most negative) eigenvalue of \mathbf{A} .

In certain systems of ODEs one might encounter eigenvalues that vary over a much larger range, five or six orders of magnitude, or more. The solver is then forced to take time steps that are orders of magnitude smaller than the actual rate of change of the solution. This is the essence of stiffness, which does not have a precise or universally accepted definition. The general idea is that a system of ODE's is considered to be **stiff** if the step size required to maintain accuracy in a numerical ODE solver is small in relation to the scale on which the solution changes with respect to the independent variable. Although the examples have been confined to linear systems with real eigenvalues, the same phenomena arises in linear systems with complex eigenvalues (as in the double spring-mass system at the start of this section) and in nonlinear systems of ODEs. A nonlinear system of ODEs may even be stiff in one time interval and not another.

Adaptive step sizing algorithms as described in Section 3.3.2 can overcome stiffness to some extent, but at the expense of the solver taking very small steps, which might be inefficient. A better approach is to use numerical algorithms tailored to stiff systems. We describe some in the next section.

Reading Exercise 7.5.6 Apply the improved Euler's method and RK4 method to the system (7.50) with initial condition $x_1(0) = 1, x_2(0) = 1.09$, using a variety of step sizes; start with $h = 0.001$ and work your way up. How large can the step size be for each algorithm before the numerical solver exhibits oscillatory or obviously unstable behavior?

7.5.3 Implicit Numerical ODE Solvers

The numerical ODE algorithms introduced in Chapter 3 and Section 7.5.1 were examples of **explicit one-step** methods. They are one-step because these algorithms advance an estimate \mathbf{x}^k of the solution to $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$ from time $t = t_k$ to an estimate \mathbf{x}^{k+1} at time t_{k+1} using only the value of \mathbf{x}^k and evaluations of \mathbf{f} on the interval $t_k \leq t \leq t_{k+1}$. More precisely, \mathbf{x}^{k+1} is computed as

$$\mathbf{x}^{k+1} = \mathbf{x}^k + h\phi(t_k, \mathbf{x}^k, h) \quad (7.54)$$

where $\phi(t, \mathbf{x}, h)$ is a function of the indicated variables. For example, when solving an ODE $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$ with Euler's method we use $\phi(t, \mathbf{x}, h) = \mathbf{f}(t, \mathbf{x})$. For the improved Euler's method we use $\phi(t, \mathbf{x}, h) = \frac{1}{2}(\mathbf{f}(t, \mathbf{x}) + \mathbf{f}(t + h, \mathbf{x} + h\mathbf{f}(t, \mathbf{x})))$ (easily obtained from (7.42)). In contrast, other types of numerical algorithms might also use information concerning previous iterates $\mathbf{x}^{k-1}, \mathbf{x}^{k-2}$, etc., and are called **multistep** methods. The algorithms of Chapter 3 are explicit because ϕ , and so \mathbf{x}^{k+1} , can be expressed explicitly using (7.54). Explicit methods are convenient because the iterates can be marched forward in time with a minimum of fuss. Their drawback, for stiff systems, is that the step size h has to be small to obtain good accuracy, even if the solution doesn't change rapidly in relation to the time step h .

An alternative for stiff systems is to use an **implicit** method. These frequently remain well-behaved for large step sizes. With a large step size the solver may become inaccurate, but will not exhibit the kind of oscillatory and exponentially growing errors typical of explicit methods. An implicit method typically involves solving an equation to obtain \mathbf{x}^{k+1} from \mathbf{x}^k and other variables. Let's illustrate by introducing the **implicit Euler's method**, also called the **backward Euler's method**.

The Implicit Euler's Method

Let's consider the scalar case first, and seek an approximate solution to an ODE $\dot{x} = f(t, x)$ with initial condition $x(t_0) = x^0$. This solution will consist of estimates x^k of $x(t_k)$ at times t_1, t_2, \dots . The usual version of Euler's method with step size h approximates $\dot{x}(t_k) \approx (x^{k+1} - x^k)/h$, which leads to the standard (explicit) Euler's method $x^{k+1} = x^k + hf(t_k, x^k)$. In the implicit or backward version of Euler's method we instead approximate $\dot{x}(t_k) \approx (x^k - x^{k-1})/h$. Using this approximation in conjunction with the ODE $\dot{x} = f(t, x)$ yields $(x^k - x^{k-1})/h = f(t_k, x^k)$ or equivalently, $x^k = x^{k-1} + hf(t_k, x^k)$. Shift indices by replacing k by $k+1$ to see that this can be expressed equivalently as

$$x^{k+1} = x^k + hf(t_{k+1}, x^{k+1}). \quad (7.55)$$

If x^0, x^1, \dots, x^k have been computed then (7.55) provides an equation that determines x^{k+1} implicitly. That is, we have to solve (7.55) for x^{k+1} , typically with a numerical root-finding method unless f is very simple.

■ **Example 7.11** Consider the equation $\dot{x}(t) = -\lambda x(t)$ (this is (7.45)) where λ is a positive constant and initial condition $x(0) = 1$. In this case $f(t, x) = -\lambda x$ and (7.55) becomes

$$x^{k+1} = x^k - h\lambda x^{k+1}.$$

This defines x^{k+1} implicitly, and in this particular case we can solve analytically to find that $x^{k+1} = x^k/(1 + \lambda h)$. With $x^0 = 1$ it follows that $x^1 = 1/(1 + \lambda h)$, $x^2 = 1/(1 + \lambda h)^2$, and more generally

$$x^k = 1/(1 + \lambda h)^k. \quad (7.56)$$

Compare this to the iteration $x^k = (1 - \lambda h)^k$ obtained from (7.46). When h is sufficiently small, both methods yield good results (that converge to the true solution $x(t) = e^{-\lambda t}$; see Exercise 7.5.3). But according to (7.47) when $h > 2/\lambda$ the explicit Euler's method produces iterates that grow without bound, even as the solution decays. The implicit Euler's method, however, decays for any positive choice of h . In the left panel of Figure 7.24 we illustrate the case $\lambda = 1$ with step size $h = 0.5$, with a graph of the iterates from the standard Euler's method and the implicit Euler's method, along with a plot of the analytical solution. This is a fairly large step size, but both numerical methods produce decaying iterates. However in the right panel we show the results when $h = 2.1$. The iterates from the standard Euler's method now grow; with this large step size the implicit Euler iterates are inaccurate, but at least they decay. ■

The Implicit Euler's Method for Systems

The implicit Euler's method extends to systems in a straightforward manner. For a system $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$ with initial condition $\mathbf{x}(t_0) = \mathbf{x}^0$ the iterate \mathbf{x}^{k+1} is computed as the (hopefully unique) solution to $(\mathbf{x}^{k+1} - \mathbf{x}^k)/h = \mathbf{f}(t_{k+1}, \mathbf{x}^{k+1})$ or

$$\mathbf{x}^{k+1} = \mathbf{x}^k + h\mathbf{f}(t_{k+1}, \mathbf{x}^{k+1}). \quad (7.57)$$

Note that \mathbf{x}^{k+1} is an n -dimensional vector, so (7.57) is a system of n equations (possibly nonlinear) in the n unknown components of \mathbf{x}^{k+1} .

■ **Example 7.12** Let's apply the implicit Euler's method to the system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ where \mathbf{A} is given by (7.50) and the initial data is $\mathbf{x}(0) = \langle 1, 1.09 \rangle$. This means we must solve (7.57) with $\mathbf{f}(t, \mathbf{x}) = \mathbf{A}\mathbf{x}$ and this leads to

$$(\mathbf{I} - h\mathbf{A})\mathbf{x}^{k+1} = \mathbf{x}^k \quad (7.58)$$

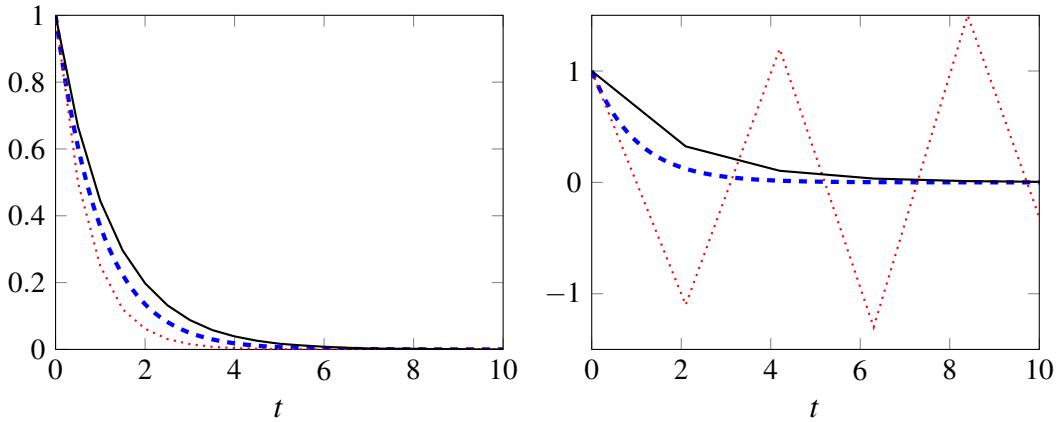


Figure 7.24: Left panel: True solution $x(t) = e^{-t}$ to $\dot{x} = -x$ (dashed blue), standard Euler's method approximation (dotted red), and implicit Euler approximation (solid black), step size $h = 0.5$. Right panel: same, but with step size $h = 2.1$.

with $\mathbf{x}^0 = \langle 1, 1.09 \rangle$. The equations defined by (7.58) are linear and can be handled by Gaussian elimination or any linear solver. This is a fortunate consequence of the fact that our ODE's here are linear. In general the resulting equations for \mathbf{x}^{k+1} could be nonlinear and a root-finding technique like Newton's method might be needed.

With $\mathbf{x}^0 = \langle 1, 1.09 \rangle$ and step size $h = 0.021$, Figure 7.25 shows the trajectories in the $x_1 x_2$ phase plane obtained by applying both Euler's method and the implicit Euler's method to this system, centered on the region $0.8 \leq x_1 \leq 1.1$, $0.8 \leq x_2 \leq 1.1$; compare to Figure 7.23, though here the direction field is omitted. The implicit method is a bit inaccurate at first (when the fast-decaying transient terms involving e^{-100t} are present) but eventually tracks the true solution well, and does not exhibit the unbounded behavior of the standard explicit Euler's method. As the step size h is decreased, the implicit Euler's method exhibits error that is proportional to h , just as the standard Euler's method does. That is, the implicit version of Euler's method is first order. ■

Other Implicit Methods

There are many other implicit methods for systems of ODE's. These methods often take the form

$$\psi(\mathbf{x}^{k+1}, \mathbf{x}^k, t_k, t_{k+1}, h) = 0, \quad (7.59)$$

for some function ψ , at least in a one-step ODE numerical method. Here \mathbf{x}^{k+1} must be solved for, usually in some non-trivial way. An example is the *trapezoidal method*, which takes the form

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \frac{h}{2}(\mathbf{f}(t_k, \mathbf{x}^k) + \mathbf{f}(t_{k+1}, \mathbf{x}^{k+1})). \quad (7.60)$$

The use of an implicit solver is often advisable when the system is known to be stiff. Although such a system can sometimes be beat to death with computational power and explicit solvers that use adaptive time stepping, this can be extremely inefficient. In some cases the algorithm may simply grind to a halt as the step size shrinks to zero. Most software packages for scientific computing (e.g., Maple, Mathematica, Matlab) have built-in solvers specifically designed for stiff systems of ODEs. See [34] for much more information on numerical methods for ODEs, stiff and otherwise.

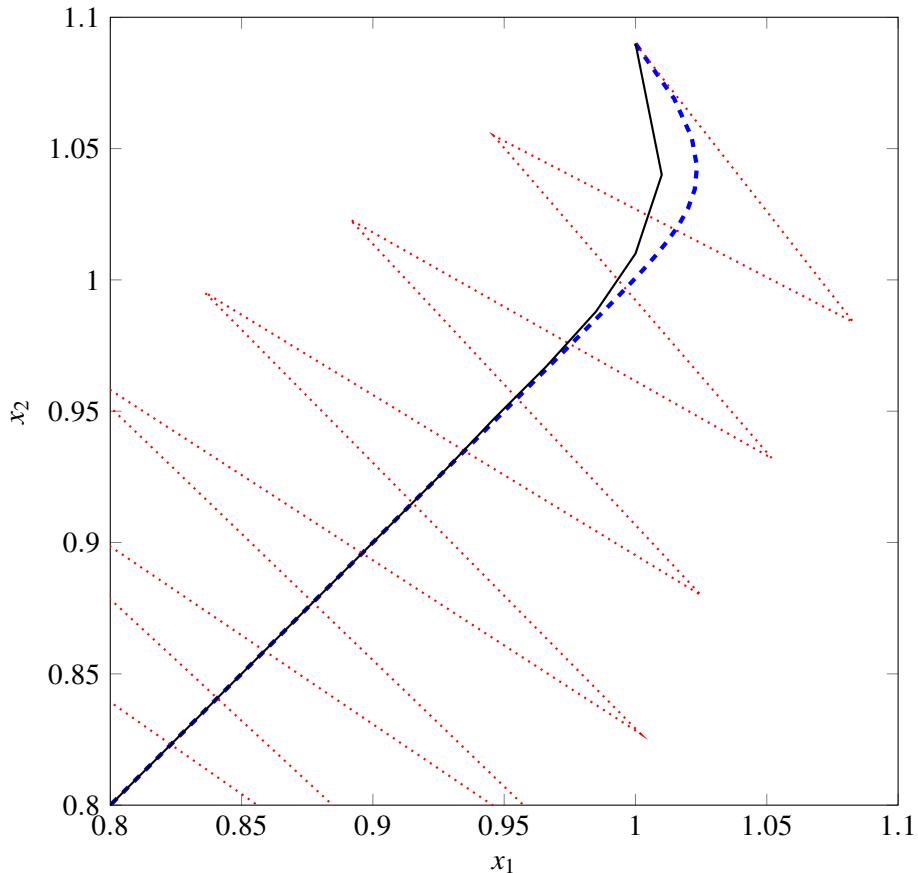


Figure 7.25: True solution trajectory to system (7.50) (dashed blue), standard Euler's method approximation (dotted red) with step size $h = 0.021$, and implicit Euler approximation (solid black) with step size $h = 0.021$.

7.5.4 Exercises

Exercise 7.5.1

- Use two steps of Euler's method with step size $h = 0.5$ to estimate $\mathbf{x}(1.0)$ where $\mathbf{x}(t) = \langle x_1(t), x_2(t) \rangle$ satisfies the equations $\dot{x}_1 = x_1 - x_2$, $\dot{x}_2 = x_1 + x_2$ with $x_1(0) = 1, x_2(0) = 2$. Compare the estimate of $\mathbf{x}(1.0)$ to the true solution value (these equations are linear, so easily solvable).
- Use two steps of Euler's method with step size $h = 0.5$ to estimate $\mathbf{x}(1.0)$ where $\mathbf{x}(t) = \langle x_1(t), x_2(t) \rangle$ satisfies the equations $\dot{x}_1 = x_1 + x_2$, $\dot{x}_2 = x_1 + x_2$ with $x_1(0) = 1, x_2(0) = 2$. Compare the estimate of $\mathbf{x}(1.0)$ to the true solution value (these equations are linear, so easily solvable).
- Use two steps of Euler's method with step size $h = 0.5$ to estimate $\mathbf{x}(1.0)$ where $\mathbf{x}(t) = \langle x_1(t), x_2(t), x_3(t) \rangle$ satisfies the equations $\dot{x}_1 = x_1 x_2 + 1 - t^3$, $\dot{x}_2 = x_1 + x_2 + t - t^2$, and $\dot{x}_3 = x_2 x_3 - 1 - t^2 + t^3$ with $x_1(0) = 0, x_2(0) = 0$, and $x_3(0) = 1$. Verify that the analytical solution is $\mathbf{x}(t) = \langle t, t^2, 1-t \rangle$. Compare the estimate of $\mathbf{x}(1.0)$ to the true solution value.

- (d) Using whatever technology you have available, redo part (b) with step sizes $h = 0.1, 0.01$, and 0.001 . Compute the error in the Euler estimate in each case. Does the error appear to be proportional to h ?
- (e) Using whatever technology you have available, redo part (c) with step sizes $h = 0.1, 0.01$, and 0.001 . Compute the error in the Euler estimate in each case. Does the error appear to be proportional to h ?

Exercise 7.5.2 In the project “The Pendulum 2” of Section 4.6.6 we derived the equation that governs the angle $\theta(t)$ of a pendulum as it swings. The relevant equation

$$\ddot{\theta}(t) + c\dot{\theta}(t) + \frac{g}{L} \sin(\theta(t)) = 0 \quad (7.61)$$

is reproduced here. The constant c quantifies the damping in the system, assumed proportional to the angular rate $\dot{\theta}$ at which the pendulum is pivoting. Let $x_1 = \theta$ and $x_2 = \dot{\theta}$.

- (a) Show that equation (7.61) is equivalent to the first-order system

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{g}{L} \sin(x_1) - cx_2.\end{aligned}$$

- (b) Use $L = 1, c = 0.1$, and $g = 9.8$ in the first-order system of part (a). Take initial conditions $x_1(0) = \pi, x_2(0) = 0$, corresponding to the pendulum perfectly upside down and motionless. What is the analytical solution? Hint: use your physical intuition and common sense.
- (c) Solve the system using Euler’s method with step size $h = 0.01$; express the value of π in the initial condition to ten significant figures. Plot the solution for $0 \leq t \leq 50$. What happens? Can you explain it?
- (d) Repeat part (c) using the RK4 method and $h = 0.01$. Try varying the step size. Can you keep the pendulum balanced for 50 seconds?
- (e) Repeat part (d) but with $c = 0$. What happens? Can you explain it?

Exercise 7.5.3 Consider the ODE $\dot{x} = -\lambda x(t)$ where λ is a positive constant and the initial condition is $x(0) = 1$. The analytical solution is $x(t) = e^{-\lambda t}$.

- (a) The explicit Euler method with step size h applied to this ODE yields iterates defined by $x^k = (1 - \lambda h)^k$, as follows from (7.46). Suppose we estimate the solution $x(T)$ by using n steps of Euler’s method, so the step size is $h = T/n$. Show that as n approaches infinity (and so h approaches 0) we have

$$\lim_{n \rightarrow \infty} x^n = x(T). \quad (7.62)$$

That is, the estimate from the standard Euler’s method converges to the correct value. Hint: use the fact that $\lim_{n \rightarrow \infty} (1 - A/n)^n = e^{-A}$ for any real number A . (Can you prove this fact as well?)

- (b) The implicit Euler’s method with step size h applied to this ODE yields iterate $x^k = 1/(1 + \lambda h)^k$ as shown in (7.56). Suppose we estimate the solution $x(T)$ by using n steps of the implicit Euler’s method, so the step size is $h = T/n$. Show that as n approaches infinity (and so h approaches 0) (7.62) also holds, so the implicit Euler estimate converges

to the correct value.

Exercise 7.5.4

- Perform two iterations of the implicit (backward) Euler's method on the ODE $\dot{x} = -0.25x$ with initial condition $x(0) = 1$; use step size $h = 0.5$. Compare the value obtained to the true value of the solution at $t = 1$.
- Perform two iterations of the implicit (backward) Euler's method on the ODE $\dot{x} = 0.5x(2-x)$ with initial condition $x(0) = 1$; use step size $h = 1$. Compare the value obtained to the true value of the solution at $t = 5$. (At each stage equation (7.55) may have two roots for x^k ; choose x^k as the root closest to x^{k-1}).
- Perform three iterations of the implicit (backward) Euler's method on the linear ODE system $\dot{\mathbf{x}} = \mathbf{Ax}$ where

$$\mathbf{A} = \begin{bmatrix} 3 & -4 \\ 8 & -9 \end{bmatrix}$$

and with initial condition $\mathbf{x}(0) = \langle 1, 3 \rangle$; use step size $h = 1$. Compare the value obtained to the true value of the solution at $t = 3$.

- Perform five iterations of the implicit (backward) Euler's method on the nonlinear ODE system

$$\begin{aligned}\dot{x}_1 &= x_1^2 - x_2 \\ \dot{x}_2 &= -x_1 - x_2\end{aligned}$$

and with initial condition $\mathbf{x}(0) = \langle 1, 3 \rangle$; use step size $h = 0.2$. If when solving for \mathbf{x}^{k+1} there are multiple possible solutions, choose that one that is closest to \mathbf{x}^k as measured in the Euclidean norm.

Exercise 7.5.5

- Run each of Euler's method and the implicit Euler's method on the ODE $\dot{x} = -10x$ with $x(0) = 1$ on the interval $0 \leq t \leq 1$, with each step size $h = 0.05$, $h = 0.125$, and $h = 0.25$. For each step size plot the iterates for both methods, and the true solution. Explain what you see in light of (7.47) and Reading Exercise 7.5.4.
- Run each of Euler's method and the implicit Euler's method on the ODE system $\dot{\mathbf{x}} = \mathbf{Ax}$ with

$$\mathbf{A} = \begin{bmatrix} 3 & -4 \\ 8 & -9 \end{bmatrix}$$

and initial condition $\mathbf{x}(0) = \langle 1, -1 \rangle$, with each step size $h = 0.1$, $h = 0.25$, and $h = 1.0$ out to final time $t = 5.0$; equation (7.58) may be helpful. For each step size produce a parametric plot of the iterates in the x_1x_2 plane, overlayed on a parametric plot of the true solution trajectory.

Exercise 7.5.6

- (a) Use the implicit Euler's method to solve the ODE $\dot{x} = -x + t$ with $x(0) = 1$ for step sizes $h = 0.1, 0.01, 0.001$, and 0.0001 , on the interval $0 \leq t \leq 1$. Find the analytical solution to this ODE and compute the error $|x(1) - x^n|$ where x_n is the implicit Euler estimate for $x(1)$. Verify that the implicit Euler's method appears to be first-order.
- (b) Use the implicit Euler's method to solve the ODE system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}(t)$ with

$$\mathbf{A} = \begin{bmatrix} 3 & -4 \\ 8 & -9 \end{bmatrix} \text{ and } \mathbf{b}(t) = \begin{bmatrix} t \\ -t \end{bmatrix}$$

with initial condition $\mathbf{x}(t) = \langle 1, -1 \rangle$. Use step sizes $h = 0.1, 0.01$, and 0.001 , out to final time $t = 1.0$. Find the analytical solution to this system and compute the error $\|\mathbf{x}(1) - \mathbf{x}^n\|$ where x_n is the implicit Euler estimate for $x(1)$. Verify that the implicit Euler's method appears to be first-order.

Exercise 7.5.7 Repeat both parts of Exercise 7.5.6 using the trapezoidal method (7.60). In each part verify that the method appears to be second-order accurate.

Exercise 7.5.8 Consider the very simple scalar ODE $\dot{x} = f(t)$ with initial condition $x(t_0) = 0$. This ODE is of the type we considered in Section 1.4.3 and can be solved by integrating both sides of the ODE with respect to t .

Suppose we wish to use a numerical ODE solver to compute $u(T)$ for some $T > t_0$ and we use step size $h = (T - t_0)/n$ for some integer $n \geq 1$. We then have $t_k = t_0 + kh$, and $t_n = T$.

- (a) Integrate both sides of $\dot{x} = f(t)$ with respect to t from $t = t_0$ to $t = T$ and use the initial condition $x(t_0) = 0$ to argue that true value for $x(T)$ is given by the integral

$$x(T) = \int_{t_0}^T f(t) dt.$$

- (b) The standard Euler's method for the ODE $\dot{x} = f(t)$ takes the form $x_{k+1} = x_k + hf(t_k)$. Apply this to $\dot{x} = f(t)$ and argue that this leads to the approximation

$$x(T) \approx x_n = \frac{T - t_0}{n} (f(t_0) + f(t_1) + \cdots + f(t_{n-1})).$$

Note the right side is just the familiar left Riemann sum approximation to $\int_{t_0}^T f(t) dt$ from integral calculus.

- (c) Repeat part (b) using the implicit Euler's method and show that the resulting approximation to $x(T)$ is the right Riemann sum approximation to $\int_{t_0}^T f(t) dt$.

Exercise 7.5.9 Consider the second-order scalar ODE $x''(t) + 2x'(t) + 101x(t) = 0$ with initial data $x(0) = 1, x'(0) = 0$.

- (a) Convert this second-order linear ODE into a first-order system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$; use $x_1 = x$ and $x_2 = \dot{x}$. Give an appropriate initial vector \mathbf{x}^0 for the system.
- (b) Find the eigenvalues and eigenvectors of \mathbf{A} and write out a general solution to the system.

Use this to find the solution with the given initial data.

- (c) Solve the system numerically for $0 \leq t \leq 5$ using both Euler's method and the implicit Euler's method with step size $h = 0.1$; Euler's method should behave badly. Plot the numerical estimates and the true solution trajectory for each method in the x_1x_2 plane. Use a range $-1 \leq x_1 \leq 1, -10 \leq x_2 \leq 10$.
- (d) Repeat part (b) using a sufficiently small step size h such that Euler's method does a reasonable job (to visual approximation). How small must h be?

Exercise 7.5.10

- (a) Take $k_1 = 1, k_2 = 10^4, m_1 = 1, m_2 = 0.001, c_1 = c_2 = 0.1$ in the double spring-mass ODE $\ddot{\mathbf{w}} = \mathbf{Aw}$ with \mathbf{A} given by (7.44) of Section 7.5.2. Write out \mathbf{A} and compute its eigenvalues (they are complex). What exponential functions govern the decay rates in this problem? What natural frequencies are present?
- (b) With the parameters as above and initial conditions $w_1(0) = 0.1$ and $w_2(0) = w_3(0) = w_4(0) = 0$ it turns out that the analytical solution for $w_1(t)$ that governs the motion of the first mass is (to a few significant figures)

$$\begin{aligned} w_1(t) \approx & e^{-49.95t}(10^{-4}\cos(3163t) + (4.73 \times 10^{-6})\sin(3163t)) \\ & + e^{-0.01t}(0.1\cos(0.994t) + 10^{-5}\sin(0.994t)). \end{aligned}$$

Note that the fast oscillating term (radial frequency $\omega \approx 3163$) is much smaller and decays much faster than the more slowly oscillating term (radial frequency $\omega \approx 0.994$).

Plot $w_1(t)$ for $0 \leq t \leq 50$. Repeat on smaller intervals, e.g., $0 \leq t \leq 1$ and $0 \leq t \leq 0.1$. The graphs of $w_2(t), w_3(t)$, and $w_4(t)$ look fairly similar. It seems a step size of $h = 0.01$ would track the solution well, at least once t is very large (so the fast oscillating part is negligible).

- (c) Solve the system with Euler's method and $h = 0.01$ out to time $t = 5$, or however far the solver will go. What happens?
- (d) Repeat part (c) with the implicit Euler's method. Is the numerical solution procedure stable? Accurate?

Exercise 7.5.11

Consider the double-loop RC circuit in Figure 7.26, components and current directions as labeled. Let $q_1(t)$ denote the charge on capacitor C_1 and $q_2(t)$ the charge on capacitor C_2 . Suppose that at time $t = 0$ the circuit is completed and the capacitors have some initial charge.

- (a) Use Kirchhoff's laws to show that \dot{q}_1 and \dot{q}_2 satisfy the first order ODEs

$$\dot{q}_1(t) = -\left(\frac{1}{R_2C_1} + \frac{1}{R_1C_1}\right)q_1(t) + \frac{1}{R_2C_2}q_2(t) \quad (7.63)$$

$$\dot{q}_2(t) = \frac{1}{R_2C_1}q_1(t) - \frac{1}{R_2C_2}q_2(t). \quad (7.64)$$

- (b) Take $C_1 = C_2 = 0.001$ farad, $R_1 = 1000$ ohms, and $R_2 = 0.1$ ohm. The ODE's (7.63)-(7.64)

become

$$\dot{q}_1(t) = -10001q_1(t) + 10000q_2(t) \quad (7.65)$$

$$\dot{q}_2(t) = 10000q_1(t) - 10000q_2(t). \quad (7.66)$$

Show that with initial condition $q_1(0) = 0.000001$ and $q_2(0) = 0.000001$ the solution to (7.65)-(7.66) is (to good approximation)

$$q_1(t) = (9.99975 \times 10^{-7})e^{-0.49999t} + (2.5 \times 10^{-11})e^{-20000.5t}$$

$$q_2(t) = (1.0 \times 10^{-6})e^{-0.49999t} - (2.5 \times 10^{-11})e^{-20000.5t}.$$

Note that both $q_1(t)$ and $q_2(t)$ contain a comparatively large amount of $e^{-0.49999t}$, a decaying exponential, and a very small multiple of $e^{-20000.5t}$, a much faster decaying exponential (this term decays 40,000 times faster).

- (c) Plot $q_1(t)$ and $q_2(t)$ on the range $0 \leq t \leq 10$; they should look pretty similar.
- (d) Based on a plot of $q_1(t)$ and $q_2(t)$ it seems that a step size of $h = 0.1$ would do a good job of tracking both of $q_1(t)$ and $q_2(t)$. Solve the system (7.65)-(7.66) numerically with Euler's method and $h = 0.1$ on the interval $0 \leq t \leq 10$ (or as far as the numerics will go). What happens?
Decrease h until the Euler iteration is stable over $0 \leq t \leq 10$. How small does h have to be?
- (e) Repeat the last problem but with the implicit Euler's method and step size $h = 0.1$. Compare to the true solution.

Remark: Note the disparity of time scales in the circuit's physical behavior. The $e^{-20000.5t}$ terms stem from the capacitors discharging through the small R_2 resistor, and they quickly assume an equal charge; this occurs on a time scale of 10^{-4} seconds. But both capacitors then discharge through the much larger R_1 resistor at a comparatively slow rate—this is the $e^{-0.49999t}$ term in the solutions, and this discharge occurs on a time scale of seconds. It may be this slower scale behavior that interests us, while the very rapid equalization of the capacitor charges that occurs when the switch is closed is of no interest. Nonetheless, it is this short time-scale phenomena that dictates the very small step size Euler's method must take to remain stable. We are thus forced to take many very small time steps in order to march the solution out in time. This problem is not confined to Euler's method, but shared by any explicit method.

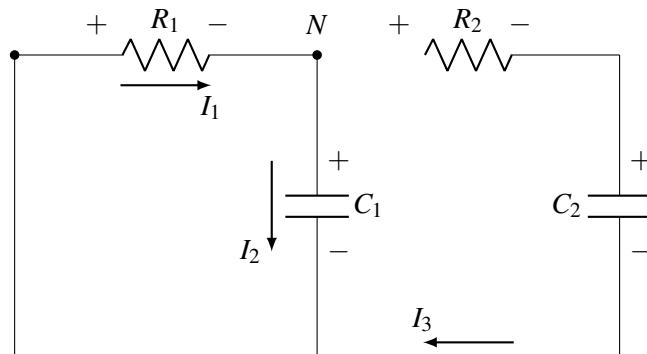


Figure 7.26: Double-loop RC circuit for Exercise 7.5.11.

7.6 Additional Techniques for Systems of First Order ODEs

The technique of drawing nullclines and direction arrows introduced in Sections 7.2 to 7.4 is largely confined to the analysis of planar systems of ODEs. Linearization for the local stability analysis of fixed points is applicable in any dimension, although its application to systems with unspecified parameters can be difficult in higher dimensions, and it fails when the real part of any eigenvalue of the Jacobian matrix is zero. In this section we introduce a few new methods that are potentially applicable to three- or higher-dimensional systems (they work in two dimensions as well) and that may aid analysis when early techniques fail.

7.6.1 First Integrals and Conservative Systems

The Pendulum Revisited

Consider a pendulum of length L as depicted in Figure 4.32. When set in motion the pendulum swings back and forth in the presence of gravitational acceleration g and makes an angle $\theta(t)$ with the vertical at time t . In the project “The Pendulum 2” of Section 4.6.6 some basic physics was used to derive a nonlinear second-order ODE (4.130) that $\theta(t)$ obeys, under the assumption that the pendulum experiences no frictional forces. That ODE is reproduced here,

$$\ddot{\theta}(t) + \frac{g}{L} \sin(\theta(t)) = 0. \quad (7.67)$$

With a prescribed initial position for $\theta(0)$ and initial velocity for $\dot{\theta}(0)$, the ODE (7.67) possesses a unique solution. Intuition tells us that since the system has no friction, the pendulum should swing back and forth forever with a constant amplitude of motion, centered on the straight-down position (at least if $\dot{\theta}(0)$ isn’t too big, so the pendulum doesn’t spin in full circles). But (7.67) can’t be solved in any simple analytical way. How can we affirm what intuition tells us?

Conservation of Energy

To answer this we reformulate (7.67) as a coupled pair of first-order ODEs (you may have done this in Reading Exercise 6.1.2 or Exercise 7.5.2). Let $x_1 = \theta$ and $x_2 = \dot{\theta}$ so that

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{g}{L} \sin(x_1).\end{aligned} \quad (7.68)$$

The coupled system (7.68) is entirely equivalent to (7.67) and we will work with (7.68) rather than (7.67). Next define a function $E(u, v)$ as

$$E(u, v) = mgL(1 - \cos(u)) + \frac{1}{2}mL^2v^2. \quad (7.69)$$

With $u = x_1(t)$ and $v = x_2(t)$ the function $E(x_1(t), x_2(t))$ is simply the total energy, kinetic plus potential, of the pendulum, scaled so that $E = 0$ when the pendulum is hanging motionless straight down. More precisely, $E(\theta(t), \dot{\theta}(t))$ is the total energy of the pendulum at time t . This was shown in detail in the project “The Pendulum” in Section 4.6.5.

The quantity $E(x_1(t), x_2(t))$ is a function of t and a straightforward computation shows that $E(x_1(t), x_2(t))$ is constant with respect to t on any solution trajectory since

$$\begin{aligned}\frac{d}{dt}(E(x_1, x_2)) &= \frac{\partial E}{\partial x_1} \frac{dx_1}{dt} + \frac{\partial E}{\partial x_2} \frac{dx_2}{dt} \\ &= mgL \sin(x_1) \dot{x}_1 + mL^2 x_2 \dot{x}_2 \\ &= mgL \sin(x_1) \dot{x}_1 + mL^2 \dot{x}_1 \dot{x}_2 \\ &= mL \dot{x}_1 \underbrace{(g \sin(x_1) + L \dot{x}_2)}_{=0} \\ &= 0,\end{aligned} \quad (7.70)$$

where we write x_1 for $x_1(t)$ and x_2 for $x_2(t)$. We used $x_2 = \dot{x}_1$ in the transition from the second to the third line in (7.70). We also used the second ODE $\dot{x}_2 = -g \sin(x_1)/L$ from (7.68) to conclude in the fourth line of (7.70) that $g \sin(x_1) + Lx_2 = 0$. Since $d(E(x_1(t), x_2(t)))/dt = 0$ it follows that $E(x_1(t), x_2(t))$ is constant with respect to t along any specific solution trajectory. Equivalently, $E(\theta, \dot{\theta})$ is constant for any solution $\theta(t)$ to (7.67). What conclusions can we draw from this?

Interpretation

If $(x_1(t), x_2(t))$ is any particular solution to the system (7.68), then we have shown that the curve parameterized by $u = x_1(t), v = x_2(t)$ in the uv plane is a level curve for the function $E(u, v)$, that is, $E(x_1(t), x_2(t)) = c$ for some constant c . By considering the graph of $E(u, v)$, some geometric conclusions about the nature of solutions to (7.68) can be made. To illustrate, the left panel of Figure 7.27 shows the graph of $E(u, v)$ in the case that $g = 9.8$, $m = 1$, and $L = 1$, with a few level curves sketched as well. The right panel shows a contour plot of $E(u, v)$, essentially a top down view of the situation in the left panel, with the direction field for the system (7.68) superimposed. The key observation is this: *each trajectory $u = x_1(t), v = x_2(t)$ for the system (7.68) lies on a level curve for E .* We can see these level curves in the right panel of Figure 7.27. The minima for E , shown as red dots in the right panel of Figure 7.27, are also level “curves.” As you can show in Reading Exercise 7.6.1, these minima occur at points of the form $u = 2k\pi, v = 0$ where k is any integer. These points correspond to equilibrium solutions of the form $x_1(t) = 2k\pi, x_2(t) = 0$ in which the pendulum hangs motionlessly, straight down.

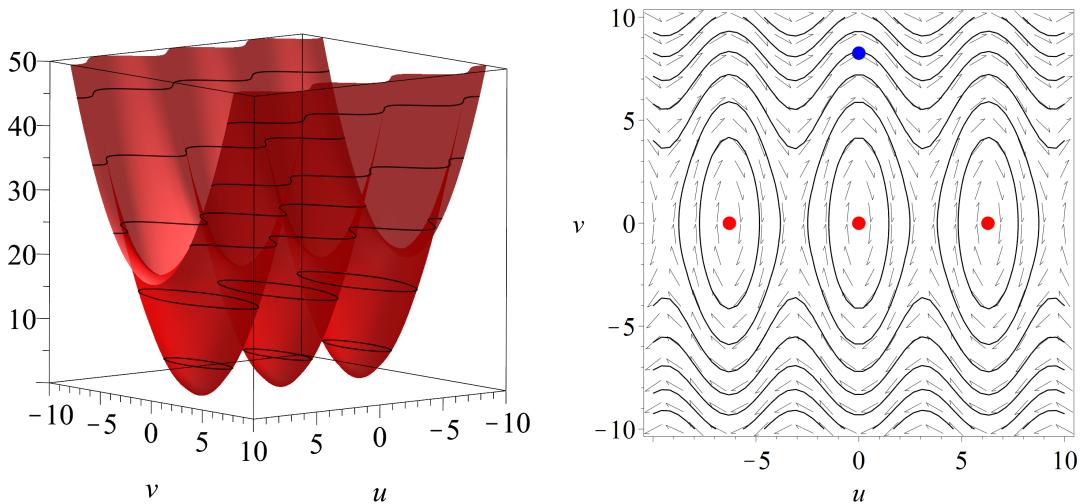


Figure 7.27: Left panel: graph of $E(u, v)$ defined by (7.69), with level curves. Right panel: Contour plot of $E(u, v)$ and direction field for (7.68); the blue dot marks a particular solution of interest in Reading Exercise 7.6.2.

The direction field indicates that the solution trajectories spiral clockwise. Because the solution trajectories remain on the level curves for E , this direction field is tangent to the level curves at all points. Figure 7.27 makes it clear that solutions that start sufficiently near these equilibrium points have trajectories that are simple closed curves that spiral clockwise around the equilibrium solutions. These solutions correspond to a perpetual back and forth swinging of the pendulum about its equilibrium. For example, Figure 7.28 shows a graph of $x_1(t)$ and $x_2(t)$ versus t (or $\theta(t)$ and $\dot{\theta}(t)$), which are the angular position and velocity of the pendulum, for the innermost level curve of E around $(0, 0)$ in the right panel of Figure 7.27. The point $x_1 = 1.4, x_2 = 0$ is used as the position and velocity of the pendulum at time $t = 0$, respectively. As the trajectory spirals around

this closed curve in the x_1x_2 phase plane, $x_1(t)$ oscillates (not quite a sinusoid) with amplitude 1.4, and $x_2(t)$ oscillates out of phase with approximate amplitude 4. At all times $E(x_1(t), x_2(t))$ remains constant.

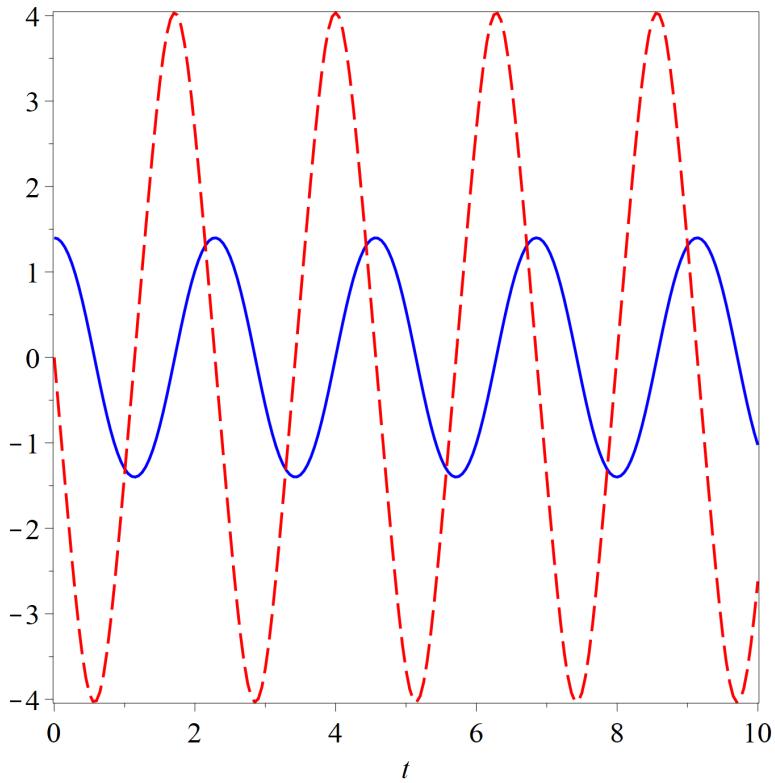


Figure 7.28: Position $x_1(t)$ (solid blue curve) and velocity (dashed red curve) $x_2(t)$ of pendulum equation solution (7.68) with $x_1(0) = 1.4$, $x_2(0) = 0$.

Reading Exercise 7.6.1 Verify that the minima of $E(u, v)$ defined by (7.69) occur at points $u = 2k\pi$ and $v = 0$, where k is any integer.

Reading Exercise 7.6.2 Some trajectories in the right panel of Figure 7.27 do not form closed curves, for example, the trajectory through the blue dot near $u = 0, v = 8.3$. What physical interpretation can you give to these types of trajectories (what is the pendulum doing)? Hint: the trajectory through the blue dot has $\theta = 0$ and $\dot{\theta} \approx 8.3$ radians per second; that is, at that instant the pendulum is in a straight-down position with a large angular velocity. Sketch $x_1(t)$ versus t and $x_2(t)$ versus t for this trajectory.

The First Integral

Since the system has no friction, it makes sense that the total energy remains constant. This observation, in conjunction with the analysis in the last section, allows us to conclude that the pendulum swings forever back and forth, unless the initial velocity is large (as in the trajectory through the blue dot in the right panel of Figure 7.27, in which case the pendulum will “wind up” by repeatedly swinging through vertical).

The function E on which solutions to (7.68) remain constant in time is called a **first integral** of this ODE system. More generally, let us make the following definition.

Definition 7.6.1 Let $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ be an autonomous system of n first-order ODEs and $E(\mathbf{u})$ a nonconstant differentiable function of an n -dimensional vector \mathbf{u} . If $E(\mathbf{x}(t))$ is constant with respect to t for all solutions to $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ then E is called a **first integral** of the system.

A first integral might also be called an **integral of motion**, a **constant of motion**, or a **conserved quantity**, especially in the context of physics. Systems for which a first integral exists are said to be **conservative**.

Reading Exercise 7.6.3 Why do you think that constant functions are excluded in the definition of first integral?

Reading Exercise 7.6.4 Consider a spring-mass system governed by $m\ddot{x} + kx = 0$.

- Formulate this second-order ODE as a pair of first-order ODEs; take $x_1 = x$ and $x_2 = \dot{x}$.
- Show that the function $E(u_1, u_2) = \frac{k}{2}u_1^2 + \frac{m}{2}u_2^2$ is a first integral of this system.
- Use the result of part (b) to conclude that the trajectories $u_1 = x_1(t)$, $u_2 = x_2(t)$ of this system are ellipses centered at $(0, 0)$ in the u_1u_2 plane. What does this say about the behavior of solutions to $m\ddot{x} + kx = 0$?

Finding a first integral for a system of ODEs may not be easy, but in many physical applications involving mechanics or electrical phenomena the total energy of the system provides a first integral. See the exercises at the end of this section for further examples.

If E is a first integral for an autonomous system of n ODEs $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ then $E(\mathbf{x})$ is constant with respect to t for any solution $\mathbf{x}(t)$, and so from the chain rule it follows that

$$\begin{aligned} 0 &= \frac{d}{dt}(E(\mathbf{x}(t))) \\ &= \nabla E(\mathbf{x}(t)) \cdot \dot{\mathbf{x}}(t) \\ &= \nabla E(\mathbf{x}(t)) \cdot \mathbf{f}(\mathbf{x}(t)) \end{aligned} \tag{7.71}$$

where ∇E denotes the gradient the function E . Equation (7.71) holds for each solution $\mathbf{x}(t)$ though any point in the common domain of E and \mathbf{f} , so we conclude that if

$$\nabla E \cdot \mathbf{f} = 0 \tag{7.72}$$

then E is a first integral.

■ **Example 7.13** For the nonlinear pendulum it follows from (7.69) that

$$\nabla E(\mathbf{x}) = \langle mgL\sin(x_1), mL^2x_2 \rangle$$

and from (7.68) that

$$\mathbf{f}(\mathbf{x}) = \langle x_2, -g\sin(x_1)/L \rangle.$$

It's easy to check that $\nabla E \cdot \mathbf{f} = 0$ in this case. ■

Stability

If we can find a first integral for a system, this may allow us to say something about the stability of the system's fixed point(s) when the technique of linearization fails. Suppose that a continuous function $E(\mathbf{u})$ is a first integral for a system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ and that $\mathbf{u} = \mathbf{x}^*$ is a strict local minimum for $E(\mathbf{u})$ —that is, $E(\mathbf{x}^*) < E(\mathbf{u})$ for all \mathbf{u} sufficiently near \mathbf{x}^* . Then it can be shown that \mathbf{x}^* is a stable equilibrium point for $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ (recall Definition 7.3.1). However, \mathbf{x}^* cannot be asymptotically stable; see Exercise 7.6.14. The situation is illustrated by the stable equilibrium points of the undamped nonlinear pendulum.

■ **Example 7.14** The function E defined in (7.69) is a first integral for the nonlinear pendulum system (7.68). It's easy to check that $E(0, 0) = 0$. A second derivative test for this function of two variables also shows that $(u, v) = (0, 0)$ is a strict local minimum. As a result the equilibrium solution $x_1 = x_2 = 0$ (corresponding to $\theta = \dot{\theta} = 0$, the pendulum hanging straight down and motionless) is a stable equilibrium point. However it is not asymptotically stable since if the pendulum is moved a bit off of vertical or given a small angular velocity, or both, the pendulum will not approach the equilibrium point, but oscillate back and forth near equilibrium; recall Definition 7.3.1. A similar conclusion holds at the equilibrium solutions $x_1 = 2k\pi, x_2 = 0$, which are physically identical to $x_1 = x_2 = 0$.

Note that trying to determine the stability of these equilibrium solutions by linearization yields a Jacobian matrix

$$\mathbf{J} = \begin{bmatrix} 0 & 1 \\ -g/L & 0 \end{bmatrix}$$

with purely imaginary eigenvalues $\pm i\sqrt{g/L}$. The Hartmann-Grobman theorem does not apply here and linearization fails to determine the stability of these fixed points. ■

7.6.2 Lyapunov Functions

The Damped Nonlinear Pendulum

The method of Lyapunov functions is related to the framework of first integrals and allows us to establish the stability, even the asymptotic stability, of fixed points for systems of ODEs. To build intuition let us consider a concrete example, by returning to the nonlinear pendulum, but this time with damping. In particular, let us consider the stability of the fixed point at which $\theta = \dot{\theta} = 0$.

■ **Example 7.15** In the project “The Pendulum 2” of Section 4.6.6 we derived a second-order ODE that is obeyed (up to modeling assumptions) by $\theta(t)$, the angle the damped pendulum makes with the vertical. The relevant equation

$$\ddot{\theta}(t) + c\dot{\theta}(t) + \frac{g}{L} \sin(\theta(t)) = 0$$

is reproduced here. The constant c quantifies the damping in the system, assumed proportional to the angular rate $\dot{\theta}$ at which the pendulum is pivoting. Let $x_1 = \theta$ and $x_2 = \dot{\theta}$ to obtain an equivalent system

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{g}{L} \sin(x_1) - cx_2. \end{aligned} \tag{7.73}$$

of first-order differential equations.

Let's consider how we might show that the fixed point $x_1 = x_2 = 0$ (pendulum motionless, hanging straight down) is asymptotically stable. The idea is to show that the friction in the system dissipates the total energy and asymptotically brings the pendulum to a halt. To this end let $E(u, v)$ again be the energy function defined by (7.69). The quantity $E(x_1, x_2)$ with $x_1 = x_1(t), x_2 = x_2(t)$ as a solution to (7.73) quantifies the total energy of the system, kinetic plus potential. The computation (7.70) showed that $d(E(x_1, x_2))/dt = 0$ in the undamped ($c = 0$) case, so that in the frictionless system, energy is conserved. This is no longer true when damping is present.

Let's redo that computation under the assumption that $c > 0$. The quantity dE/dt along a

trajectory is given by

$$\begin{aligned}
 \frac{d}{dt}(E(x_1(t), x_2(t))) &= \nabla E(x_1, x_2) \cdot \langle dx_1/dt, dx_2/dt \rangle \\
 &= mgL\sin(x_1)\dot{x}_1 + mL^2x_2\dot{x}_2 \\
 &= mgL\sin(x_1)\dot{x}_1 + mL^2\dot{x}_1\dot{x}_2 \\
 &= mL\dot{x}_1(g\sin(x_1) + L\dot{x}_2) \\
 &= mL\dot{x}_1(g\sin(x_1) - g\sin(x_1) - cLx_2) \\
 &= -cmL^2x_2^2,
 \end{aligned} \tag{7.74}$$

where we have made repeated use of (7.73). Thus $d(E(x_1, x_2))/dt = -cmL^2x_2^2$ and since c, m , and L are all positive it follows that $dE/dt \leq 0$ at all times, with strict inequality $dE/dt < 0$ whenever $x_2 \neq 0$. Since $x_2 = \dot{\theta}$ this means that if the pendulum is moving (which it always is, except at the extremes of its swing), it is losing energy. The fact that $dE/dt < 0$ at almost all points strongly suggests, but doesn't quite prove, that the pendulum will asymptotically approach a position in which $E = 0$, that is, the equilibrium point in which it hangs straight down.

Figure 7.29 provides a geometric view of the situation in the x_1x_2 phase plane. A typical solution trajectory is shown as a solid blue curve, spiraling toward the fixed point at $x_1 = x_2 = 0$, along with a direction field for the system. The dashed black ovals are level curves for the function $E(x_1, x_2)$, with the origin satisfying $E(0, 0) = 0$ and $E(x_1, x_2) > 0$ away from the origin, so the origin is a strict minimizer for E . The important geometric observation here is that the solution trajectory almost always cuts across the level curves at an acute angle, moving from higher values of E to lower values. The only exception is at points where $x_2 = 0$ (the horizontal axis), when the trajectory is briefly tangent to the corresponding level curve. This reflects the fact that $d(E(x_1(t), x_2(t))/dt < 0$ for all t , except at those isolated times when $x_2 = 0$ where $d(E(x_1(t), x_2(t))/dt = 0$, as shown by (7.74).

That this fixed point for the pendulum is asymptotically stable can be shown by linearizing the system (this was done in Exercise 7.4.5). ■

The Method of Lyapunov Functions

Example 7.15 might seem to add little to our analytical techniques for ODEs, since it merely suggests that the damped nonlinear pendulum has asymptotically stable equilibrium points, something we can already prove with linearization. But the technique of Example 7.15 can also be used to show this rigorously and works in some cases where linearization fails.

Let's begin with a definition.

Definition 7.6.2 Let $V(\mathbf{u})$ be a function of n variables defined on some region D containing a point $\mathbf{u} = \mathbf{x}^*$. Suppose that $V(\mathbf{x}^*) = 0$. If $V(\mathbf{u}) > 0$ for $\mathbf{u} \neq \mathbf{x}^*$ then V is said to be **positive definite with respect to \mathbf{x}^*** on D , while if $V(\mathbf{u}) \geq 0$ for $\mathbf{u} \neq \mathbf{x}^*$ then V is said to be **positive semidefinite with respect to \mathbf{x}^*** . In the cases $V(\mathbf{u}) < 0$ or $V(\mathbf{u}) \leq 0$ we say V is **negative definite with respect to \mathbf{x}^*** or **negative semidefinite with respect to \mathbf{x}^*** , respectively.

■ **Example 7.16** The function $V(\mathbf{u}) = mgL(1 - \cos(u_1)) + \frac{1}{2}mL^2u_2^2$ where $\mathbf{u} = \langle u_1, u_2 \rangle$ (this is essentially E as defined in (7.69)) is positive definite with respect to the origin $\mathbf{u} = \mathbf{0}$ on the region D defined by $-1 < u_1, u_2 < 1$, which contains the origin $\mathbf{x}^* = \langle 0, 0 \rangle$. This follows since $1 - \cos(u_1) > 0$ if $-1 < u_1 < 1$ (notice that all points \mathbf{u} in D satisfy $-1 < u_1 < 1$). So the first term $mgL(1 - \cos(u_1))$ in V is always positive, and the second term $mL^2u_2^2/2$ is always nonnegative. ■

Reading Exercise 7.6.5 Let $m = L = 1$ and $g = 9.8$ in the function $V(\mathbf{u})$ in Example 7.16.

- (a) Plot the function $V(\mathbf{u})$ on the region D in that example and confirm that V is positive definite with respect to the origin on D .

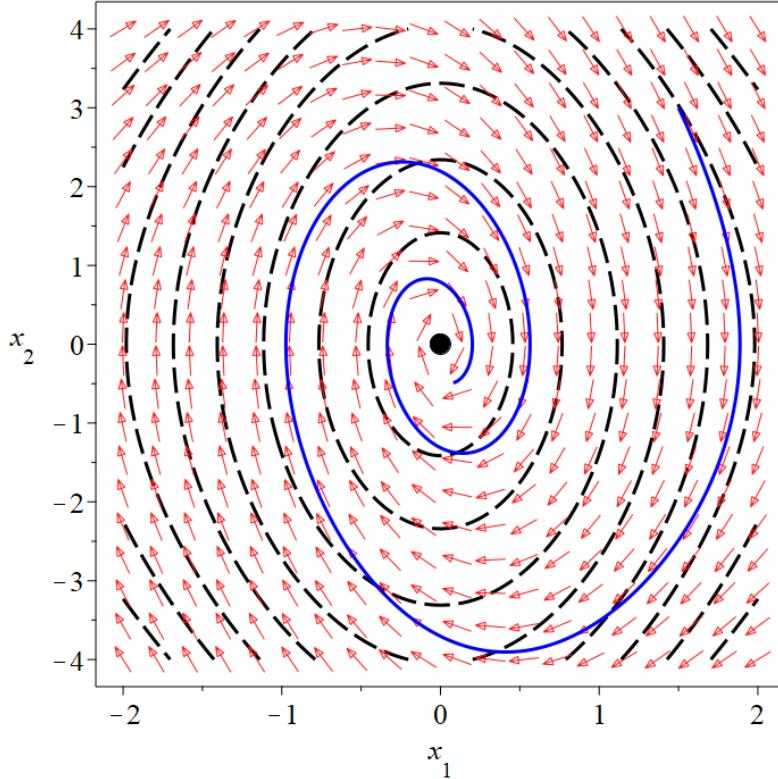


Figure 7.29: Direction field and typical solution curve (solid blue) for the nonlinear pendulum system (7.73), with level curves (dashed black) for the system energy function $E(x_1, x_2)$ defined by (7.69).

- (b) Suppose D is the region defined by $-3\pi < u_1 < 3\pi$, $-1 < u_2 < 1$. Show that V is only positive semidefinite with respect to the origin on this rectangle.

The general approach used in Example 7.15 was to show that the system energy decreases along trajectories. This makes it plausible that the system motion decays to a minimum energy state, which is an equilibrium. To flesh this out, note that the same computation that led to (7.71) shows that if $V(\mathbf{u})$ is any differentiable function of n variables $\langle u_1, \dots, u_n \rangle$ and $\mathbf{x}(t)$ is a solution to a system of n ODEs $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ then

$$\frac{d(V(\mathbf{x}(t)))}{dt} = \nabla V(\mathbf{x}(t)) \cdot \mathbf{f}(\mathbf{x}(t)). \quad (7.75)$$

If $\nabla V \cdot \mathbf{f} < 0$ in some region D then from (7.75) it follows that $\frac{d(V(\mathbf{x}(t)))}{dt} < 0$ and so V is strictly decreasing along the solution trajectories of the ODE system that lie in D . Suppose that \mathbf{x}^* lies in D with $V(\mathbf{x}^*) = 0$ and $V(\mathbf{u}) > 0$ for $\mathbf{u} \neq \mathbf{x}^*$. Then the inequality $\frac{d(V(\mathbf{x}(t)))}{dt} < 0$ suggests that $V(\mathbf{x}(t))$ decreases to 0 as t increases and that $\mathbf{x}(t)$ converges to \mathbf{x}^* , so that \mathbf{x}^* is asymptotically stable. This can be shown to be true, at least if \mathbf{x}^* is an **isolated fixed point**. An isolated fixed point \mathbf{x}^* is one in which there is a ball $B_r = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}^*\|\}$ of some radius $r > 0$ centered at \mathbf{x}^* that contains no other fixed points.

This is summarized in the following theorem.

Theorem 7.6.1 Suppose $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ is a system of n ODEs with \mathbf{f} defined on some region D and \mathbf{x}^* is a fixed point for this system that lies in D . Suppose $V(\mathbf{u})$ is a differentiable function

of $\mathbf{u} = \langle u_1, \dots, u_n \rangle$ where $V(\mathbf{x}^*) = 0$ and V is positive definite with respect to \mathbf{x}^* on D . If $\nabla V(\mathbf{u}) \cdot \mathbf{f} \leq 0$ for all points $\mathbf{u} \neq \mathbf{x}^*$ in D then \mathbf{x}^* is stable. If \mathbf{x}^* is isolated and $\nabla V(\mathbf{u}) \cdot \mathbf{f} < 0$ for all points $\mathbf{u} \neq \mathbf{x}^*$ in D then \mathbf{x}^* is asymptotically stable.

For a proof of this, see [100]. Also note that if a fixed point \mathbf{x}^* is not isolated, it cannot be asymptotically stable since there are other fixed points \mathbf{w}^* arbitrarily close to \mathbf{x}^* , and the solution that starts at \mathbf{w}^* does not converge to \mathbf{x}^* (it stays at \mathbf{w}^*).

When weak inequality holds, so $\nabla V \cdot \mathbf{f} \leq 0$, the function V in Theorem 7.6.1 is called a **Lyapunov function** for the system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$. If the inequality is strict, so $\nabla V \cdot \mathbf{f} < 0$, then V is a **strict Lyapunov function**. For a given system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ there may be many choices for V that can be used to show the stability of a fixed point \mathbf{x}^* (if \mathbf{x}^* is in fact stable). Coming up with a suitable choice for V to prove stability is a bit of an art form or guessing game, as the following examples will illustrate. Establishing the stability of a fixed point by finding an appropriate Lyapunov function is called **Lyapunov's second method** or **Lyapunov's direct method**.

■ **Example 7.17** Linear systems of ODEs can be solved explicitly and the stability of fixed points determined by examining the eigenvalues of the relevant Jacobian matrix, so Lyapunov's second method is most useful for nonlinear systems, and in particular systems where linearization fails. Nonetheless, let us illustrate the method first with a straightforward linear example.

Consider the system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ where

$$\mathbf{A} = \begin{bmatrix} -3 & 1 \\ 1 & -3 \end{bmatrix}.$$

The system is of the form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ with

$$\mathbf{f}(\mathbf{x}) = \langle -3x_1 + x_2, x_1 - 3x_2 \rangle.$$

The only fixed point for this system is the origin, $\mathbf{x}^* = \langle 0, 0 \rangle$. The eigenvalues of \mathbf{A} are -2 and -4 , so the origin is asymptotically stable, but let's use Lyapunov's second method to establish this fact.

We claim that the function $V(\mathbf{u}) = u_1^2 + u_2^2$ acts as a strict Lyapunov function for this system at fixed point $\mathbf{x}^* = \langle 0, 0 \rangle$. To see this, first compute $\nabla V = \langle 2u_1, 2u_2 \rangle$. Note that $V(\mathbf{x}^*) = V(\mathbf{0}) = 0$ and clearly $V(\mathbf{u}) > 0$ if $\mathbf{u} \neq \langle 0, 0 \rangle$. A simple computation shows that

$$\nabla V(\mathbf{u}) \cdot \mathbf{f}(\mathbf{u}) = -6u_1^2 + 4u_1u_2 - 6u_2^2.$$

The expression $-6u_1^2 + 4u_1u_2 - 6u_2^2$ on the right is quadratic in u_1 and u_2 and is called a **quadratic form**. Let's define $Q(\mathbf{u}) = -6u_1^2 + 4u_1u_2 - 6u_2^2$, where $\mathbf{u} = \langle u_1, u_2 \rangle$. To establish the asymptotic stability of \mathbf{x}^* we need to show that $Q(\mathbf{u}) < 0$ for all points \mathbf{u} close to \mathbf{x}^* .

A plot of $Q(\mathbf{u})$ as a function of u_1 and u_2 is shown in Figure 7.30 and makes it clear that $Q(\mathbf{u}) < 0$ at all points near the origin. This isn't a proof, but it is a fact that a quadratic form $Q(x, y) = ax^2 + bxy + cy^2$ in two variables x and y satisfies $Q(x, y) < 0$ for all x and y if $a < 0$ and $ac - b^2/4 > 0$; see Exercise 7.6.8. In the present case since $-6 < 0$ and $(-6)^2 - 4^2/4 = 32 > 0$ it follows that $Q(\mathbf{u}) < 0$ for all u_1 and u_2 (although this only needs to be true for all u_1, u_2 sufficiently close to the origin). According to Theorem 7.6.1 the fixed point at the origin is asymptotically stable. ■

In Example 7.17 we used $V(\mathbf{u}) = u_1^2 + u_2^2$ as a Lyapunov function. As mentioned above, coming up with an appropriate Lyapunov function is an art form. There are no strategies that are guaranteed to work, but quadratic functions with a minimum value of 0 at the equilibrium point \mathbf{x}^* of interest are often a good start; see Exercises 7.6.3 or 7.6.4.

Let's consider another example, this time nonlinear.

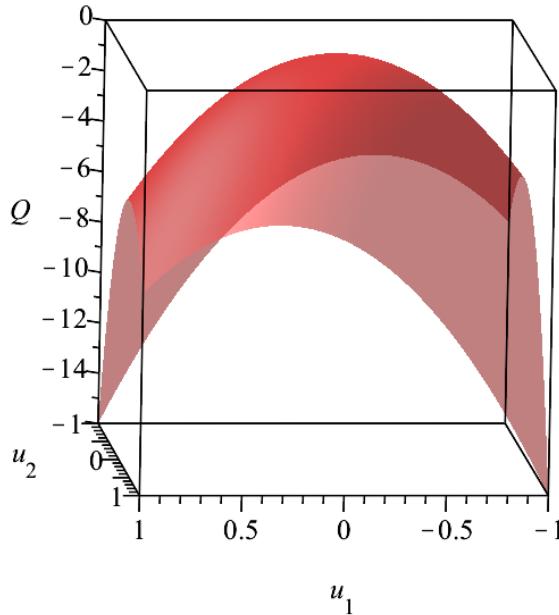


Figure 7.30: A plot of the quadratic form $Q(\mathbf{u}) = -6u_1^2 + 4u_1u_2 - 6u_2^2$.

■ **Example 7.18** Consider the nonlinear system

$$\begin{aligned}\dot{x}_1 &= -x_1^3 + x_2 \\ \dot{x}_2 &= -x_1 - x_2^3.\end{aligned}$$

Straightforward algebra shows this system has one fixed point, at $x_1 = x_2 = 0$. However, linearizing the system using the method of Section 7.3.2 yields a Jacobian matrix with eigenvalues $\pm i$, both of which have real part equal to zero, so the Hartman-Grobman Theorem yields no conclusion concerning the stability of the fixed point $(0,0)$.

Consider the function $V(\mathbf{u}) = u_1^2 + u_2^2$ as in the previous example. Then $V(0,0) = 0$ and V is positive definite with respect to $(0,0)$ on any region containing $(0,0)$. With $\mathbf{f}(x_1, x_2) = \langle -x_1^3 + x_2, -x_1 - x_2^3 \rangle$ we find that

$$\nabla V(u_1, u_2) \cdot \mathbf{f}(\mathbf{u}) = -2(u_1^4 + u_2^4) < 0$$

for $(u_1, u_2) \neq (0,0)$. According to Theorem 7.6.1 the fixed point $(0,0)$ is asymptotically stable. ■

Basin of Attraction

The analysis of Example 7.18 allows us to conclude that solutions that start close to the fixed point $(0,0)$ approach that fixed point. But what about solutions that start far away? In some cases the method of Lyapunov functions lets us determine that solutions that start in a given region approach a specific fixed point of interest. Suppose \mathbf{x}^* is a fixed point for $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ where \mathbf{f} is defined on a set D . We define the set

$$B(\mathbf{x}^*) = \{\mathbf{x}_0 \in D : \lim_{t \rightarrow \infty} \mathbf{x}(t) = \mathbf{x}^* \text{ where } \mathbf{x}(0) = \mathbf{x}_0\}. \quad (7.76)$$

The set $B(\mathbf{x}^*)$ is called the **basin of attraction** for \mathbf{x}^* and consists of all points \mathbf{x}_0 such that a solution to $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ that starts at \mathbf{x}_0 approaches \mathbf{x}^* at $t \rightarrow \infty$.

■ **Example 7.19** Consider a simple linear system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ where \mathbf{A} has all eigenvalues with negative real part. Examples are the pharmacokinetic model (6.1)-(6.2) (with $g(t) = 0$) for LSD, the system (6.5)-(6.6) for a damped spring-mass system, or the system (6.10) to (6.11) for the damped double spring-mass system. In cases such as these we can solve the system explicitly and see that all solutions decay to the equilibrium solution $\mathbf{x}^* = \mathbf{0}$. Thus the basin of attraction $B(\mathbf{0})$ for the fixed point at the origin in each case is the entire space of initial conditions. If this occurs for a fixed point \mathbf{x}^* then we say that \mathbf{x}^* is **globally attractive** or **globally asymptotically stable**. ■

■ **Example 7.20** Consider the competing species model (7.3)-(7.4) with parameters $r_1 = 1$, $K_1 = 2$, $r_2 = 2$, $K_2 = 3$, $a = 3$, and $b = 3$. A phase portrait was sketched in Figure 7.13. In Exercise 7.3.2 you were asked to use linearization to show that the fixed points $(2, 0)$ and $(0, 3)$ are asymptotically stable, as is strongly suggested by the right panel in Figure 7.13. A visual examination of that picture suggests that the basin of attraction for the fixed point at $(2, 0)$ consists of all points below a curve through the points $(0, 0)$ and the unstable fixed point at $(7/8, 3/8)$, and then continuing diagonally upward. Points above this curve form the basin of attraction for the fixed point at $(0, 3)$. On the other hand, with lower competition parameters $a = 0.2$ and $b = 0.45$, an examination of Figure 7.11 suggests that the fixed point near $(1.54, 2.31)$ attracts all solutions that start with both initial populations strictly positive. Thus $(1.54, 2.31)$ is globally asymptotically stable, if we confine our analysis to positive initial data. ■

It is sometimes helpful to be able to estimate the basin of attraction of a fixed point. Here is a minor modification of Theorem 7.6.1 that addresses this issue.

Theorem 7.6.2 Let $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ be a system of ODEs defined on some region D , \mathbf{x}^* an isolated equilibrium solution in D , $V(\mathbf{x})$ a continuously differentiable strict Lyapunov function on D (so $V(\mathbf{x}^*) = 0$ and $V(\mathbf{u}) > 0$ for $\mathbf{u} \neq \mathbf{x}^*$). For some $r > 0$ let D_r be the set of all \mathbf{u} in D such that $V(\mathbf{u}) < r$ (note \mathbf{x}^* is in D_r). If $\nabla V \cdot \mathbf{f} < 0$ in D_r then $D_r \subseteq B(\mathbf{x}^*)$ as defined by (7.76).

For a proof see [117].

■ **Example 7.21** For the system of Example 7.18 (with $D = \mathbb{R}^2$) we found that the origin is the only fixed point, and taking $V(u_1, u_2) = u_1^2 + u_2^2$ yielded $\nabla V \cdot \mathbf{f} = -2(u_1^4 + u_2^4) < 0$ for all u_1, u_2 . The set D_r is a disk of radius \sqrt{r} centered at the origin, and since $\nabla V \cdot \mathbf{f} < 0$ on D_r it follows that any solution starting within a distance \sqrt{r} of the origin approaches the origin. Since r is arbitrary it follows that all solutions to this system approach the origin, so the basin of attraction for this fixed point is the whole plane. ■

■ **Example 7.22** Consider the competing species model (7.3)-(7.4) with parameters $r_1 = 1$, $K_1 = 2$, $r_2 = 2$, $K_2 = 3$, $a = 0.2$, and $b = 0.45$ (the second case examined in Example 7.20). This system has a fixed point at $u_1 = 20/13 \approx 1.54$, $u_2 = 30/13 \approx 2.33$. Let $V(u_1, u_2) = (u_1 - 20/13)^2 + (u_2 - 30/13)^2$, which is positive definite with respect to this fixed point on any region containing the fixed point. Then

$$\nabla V(\mathbf{u}) \cdot \mathbf{f}(\mathbf{u}) = -u_1^3 + \frac{46}{13}u_1^2 - \frac{1}{5}u_1^2u_2 - \frac{40}{13}u_1 + \frac{22}{13}u_1u_2 - \frac{92}{13}u_2^2 - \frac{3}{5}u_1u_2^2 - \frac{120}{13}u_2 - \frac{4}{3}u_2^3. \quad (7.77)$$

In Figure 7.31 the right side of (7.77) is graphed on a ball of radius $r = 1/2$ around the fixed point (this region is exactly the set $V(\mathbf{u}) < 1/4$). It is clear that $\nabla V \cdot \mathbf{f} < 0$ here, except at the fixed point $\mathbf{u} = \mathbf{x}^* = (20/13, 30/13)$. According to Theorem 7.6.2, the ball of radius $1/2$ around \mathbf{x}^* is contained in the basin of attraction $B(\mathbf{x}^*)$, though we will not stop to prove that $\nabla V \cdot \mathbf{f} < 0$. As noted in Example 7.20 and based on Figure 7.11 it seems that the basin of attraction for this fixed point is the entire first quadrant $u_1 > 0, u_2 > 0$. ■

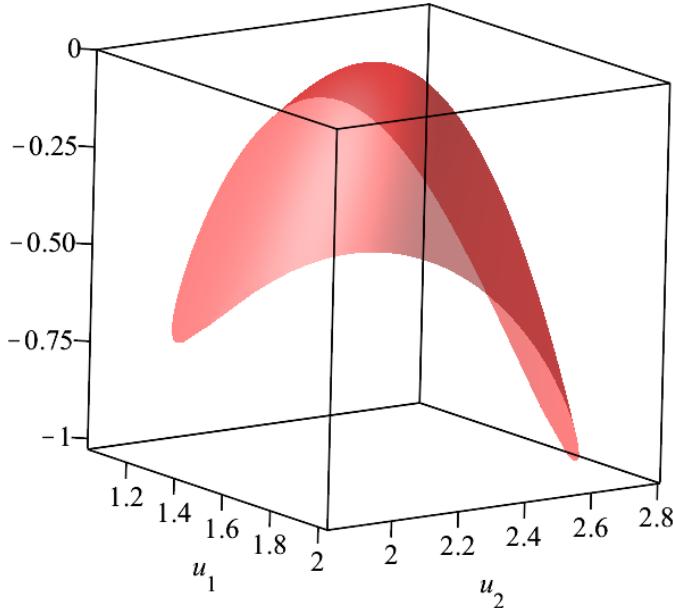


Figure 7.31: A plot of the function $\nabla V \cdot \mathbf{f}$ defined in (7.77).

Back to the Nonlinear Pendulum

Let's return to Example 7.15 and show that the equilibrium solution $(x_1, x_2) = (0, 0)$ (or $\theta = \dot{\theta} = 0$) is asymptotically stable for the damped pendulum. For our Lyapunov function we will use the system energy $V(x_1, x_2) = mgL(1 - \cos(x_1)) + \frac{1}{2}mL^2x_2^2$ (previously defined in (7.69) as $E(u, v)$). The computation (7.74) shows that

$$\nabla V(\mathbf{u}) \cdot \mathbf{f}(\mathbf{u}) = -cmL^2u_2^2. \quad (7.78)$$

using $\mathbf{f}(\mathbf{u}) = \langle u_1, -\frac{g}{L} \sin(u_1) - cu_2 \rangle$. According to Theorem 7.6.1 the fixed point $\mathbf{x}^* = \mathbf{0}$ is stable since $\nabla V \cdot \mathbf{f} \leq 0$ on any region containing \mathbf{x}^* (in fact, this Lyapunov function works at any fixed point $x_1 = 2k\pi, x_2 = 0$). But $\nabla V \cdot \mathbf{f}$ is not strictly negative in any neighborhood of \mathbf{x}^* (take $u_2 = 0$ and u_1 to be anything nonzero), so we cannot conclude that \mathbf{x}^* is asymptotically stable.

However, the following useful theorem, known as **LaSalle's invariance principle**, is a refinement of Theorem 7.6.1 that can address this situation.

Theorem 7.6.3 Suppose $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ is a system of n ODEs with \mathbf{f} defined on some region D . Let \mathbf{x}^* be a point in D that is an isolated fixed point for this system. Let $V(\mathbf{u})$ be a differentiable function of $\mathbf{u} = \langle u_1, \dots, u_n \rangle$ where $V(\mathbf{x}^*) = 0$ and V is positive definite with respect to \mathbf{x}^* on D . Suppose $\nabla V(\mathbf{u}) \cdot \mathbf{f} \leq 0$ for all points $\mathbf{u} \neq \mathbf{x}^*$ in D and that the set $S = \{\mathbf{u} \in D : \nabla V(\mathbf{u}) \cdot \mathbf{f}(\mathbf{u}) = 0\}$ contains no solution trajectories for the ODE system except $\mathbf{x}(t) = \mathbf{x}^*$. Then \mathbf{x}^* is asymptotically stable.

For a discussion of this theorem see [114].

In application to the damped pendulum problem with V as above, consider a ball D of some small radius, say radius $1/2$, centered on $\mathbf{x}^* = \mathbf{0}$. From (7.78) we conclude that $\nabla V \cdot \mathbf{f} \leq 0$ in D . The set S here consists of the horizontal axis $u_2 = 0$, intersected with D . It's not hard to see that S contains

no solution trajectories for the pendulum motion (they would be of the form $x_2(t) = 0, x_1(t) = h(t)$ for some $h(t)$, but then $\dot{x}_1 = x_2$ forces $h(t)$ to be constant, and from $\dot{x}_2 = -\frac{g}{L} \sin(x_1) - cx_2$ it follows that $0 = \sin(h(t))$, so only $h(t) = 0$ works, which yields $\mathbf{x}(t) = \mathbf{x}^*$). From Theorem 7.6.3 the fixed point $\mathbf{x}^* = \mathbf{0}$ is asymptotically stable.

7.6.3 Linearization and the Routh-Hurwitz Theorem

In Section 7.3.2 we saw how to use linearization to determine the stability of fixed points to nonlinear ODEs. In order to establish the asymptotic stability of a fixed point we must show that the eigenvalues of the Jacobian at that point all have negative real part. When the system does not contain any unspecified parameters this is straightforward, since the eigenvalues are just numbers and easy to compute with standard software. But when the system contains unspecified parameters the situation becomes a more difficult symbolic computation. Section 7.4 contained an example centered on the competing species model. The relevant eigenvalues, and fixed point stability, depend on the value of various parameters. For two-dimensional systems the trace-determinant analysis of Theorem B.4.1 in Section B.4 is useful. The alternative is a direct examination of the eigenvalues of the Jacobian and their dependence on the unspecified parameters, which can be quite messy.

However, the trace-determinant analysis is specific to two dimensions. Is there anything analogous for three- or higher-dimensional systems? In particular, it would be useful to have a convenient method to determine when the eigenvalues of the Jacobian at a fixed point all have negative real part, so that we can infer that the fixed point is asymptotically stable. This is where the Routh-Hurwitz theorem can be helpful. We begin with an example.

A Three Species Food Chain

Consider an ecosystem consisting of three species, a prey species with population $x_1(t)$, a second species with population $x_2(t)$ that preys upon the first species, and a third species with population $x_3(t)$ that preys upon the second species (but not the first). Such a scenario is called a **three species food chain**. Assume that the first species, in the absence of predation by the second species, would grow logistically with intrinsic growth rate r and carrying capacity K . The second species gains sustenance from eating the first, but otherwise would die out exponentially. Similarly the third species gains sustenance from eating the first, but otherwise would also die out exponentially. One model for this situation is

$$\begin{aligned}\dot{x}_1 &= rx_1(1 - x_1/K - ax_2/K) \\ \dot{x}_2 &= bx_1x_2 - cx_2x_3 - dx_2 \\ \dot{x}_3 &= ex_2x_3 - fx_3\end{aligned}\tag{7.79}$$

where r, K, a, b, c, d, e , and f are certain positive parameters. Of course x_1, x_2 , and x_3 are all nonnegative. See, for example, [59].

Reading Exercise 7.6.6 Justify each term on the right in (7.79) in light of the assumptions made. What does each coefficient a, b, c, d, e , and f represent?

A bit of algebra shows that there are five fixed points for this system, namely (x_1, x_2, x_3) equal to one of

$$(0, 0, 0), \quad (K, 0, 0), \quad (0, f/e, -d/c), \quad (d/b, (Kb - d)/(ab), 0), \quad \text{and } (K - af/e, f/e, (Kbe - fac - de)/(ec)).\tag{7.80}$$

Reading Exercise 7.6.7 Interpret each fixed point in (7.80) in plain English—what is the population of each species doing? The last three fixed points may or may not be physically relevant depending on the value of the parameters. What conditions on the parameters are necessary for

each of these fixed points to have physical meaning? Which fixed point in the list (7.80) is never physically relevant?

For simplicity, let us initially take $a = b = c = d = e = f = 1$, while leaving r and K undefined; there's nothing special about these parameter choices. For the moment let us focus on the last fixed point in the list (7.80), which represents the mutual coexistence of all three species. we seek conditions on the parameters r and K under which such three-way coexistence is stable. With these parameters this fixed point is given by $(K - 1, 1, K - 2)$ (this is the last entry in the list (7.80) when $a = b = c = d = e = f = 1$). Note that this fixed point is physically relevant only if $K > 2$, which we now assume is the case. Intuitively, since the first species supports the entire food chain, the environment must support enough of this species or the other species cannot survive. But even if $K > 2$, is this fixed point stable? What role if any does r play in the stability?

The Jacobian matrix for the system is

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} r(K - 2x_1 - x_2)/K & -rx_1/K & 0 \\ x_2 & x_1 - x_3 - 1 & -x_2 \\ 0 & x_3 & x_2 - 1 \end{bmatrix}. \quad (7.81)$$

At the fixed point $\mathbf{x}^* = \langle K - 1, 1, K - 2 \rangle$ the Jacobian becomes

$$\mathbf{J}(\mathbf{x}^*) = \begin{bmatrix} r(1 - K)/K & r(1 - K)/K & 0 \\ 1 & 0 & -1 \\ 0 & K - 2 & 0 \end{bmatrix} \quad (7.82)$$

The stability of this fixed point is determined by the eigenvalues of $\mathbf{J}(\mathbf{x}^*)$, but despite the relative simplicity of this matrix the eigenvalues are extremely complicated as functions of r and K . sorting out precisely when all have negative real part looks hopeless.

We will return to the analysis of the three species food chain after developing an additional tool for stability analysis.

The Routh-Hurwitz Theorem

The eigenvalues of a matrix \mathbf{A} are the roots of the characteristic polynomial of \mathbf{A} . The characteristic polynomial is straightforward to compute, at least for modestly-sized matrices, even if the matrix contains symbolic entries (as does $\mathbf{J}(\mathbf{x}^*)$ in (7.81)). We need a criterion for deciding when these roots, the eigenvalues, have negative real part. This is addressed by the following theorem.

Theorem 7.6.4 — Routh-Hurwitz. Let $p(z) = z^n + a_1z^{n-1} + a_2z^{n-2} + \dots + a_{n-1}z + a_n$ be a polynomial with real coefficients a_k (note a_k is the coefficient of z^{n-k}). Then all roots of p have negative real part if and only if $D_1 > 0, D_2 > 0, \dots, D_n > 0$ where D_k is the determinant of the $k \times k$ matrix,

$$D_k = \det \begin{bmatrix} a_1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ a_3 & a_2 & a_1 & 1 & 0 & 0 & \cdots & 0 \\ a_5 & a_4 & a_3 & a_2 & a_1 & 1 & \cdots & 0 \\ a_{2k-1} & a_{2k-2} & a_{2k-3} & a_{2k-4} & a_{2k-5} & a_{2k-6} & \cdots & a_k \end{bmatrix} \quad (7.83)$$

and we set $a_k = 0$ if $k > n$.

Note that $D_1 = a_1$ in Theorem 7.6.4.

■ **Example 7.23** Let $p(z) = z^3 + 5z^2 + 9z + 5$, a polynomial of degree 3, so $a_1 = 5, a_2 = 9$, and

$a_3 = 5$. Then

$$\begin{aligned} D_1 &= \det [5] = 5, \\ D_2 &= \det \begin{bmatrix} 5 & 1 \\ 5 & 9 \end{bmatrix} = 40, \\ D_3 &= \det \begin{bmatrix} 5 & 1 & 0 \\ 5 & 9 & 5 \\ 0 & 0 & 5 \end{bmatrix} = 200. \end{aligned}$$

Since $D_j > 0$ for $j = 1, 2, 3$ it follows that all roots of $p(z)$ have negative real part. Indeed, the roots of $p(z)$ are $-1, -2+i$, and $-2-i$. ■

■ **Example 7.24** Let $p(z) = z^3 + rz^2 + 9z + 5$, a polynomial of degree 3; here r is some unspecified constant. For what values of r will this polynomial have roots with negative real part? We have $a_1 = r, a_2 = 9$, and $a_3 = 5$ and find

$$\begin{aligned} D_1 &= \det [5] = r, \\ D_2 &= \det \begin{bmatrix} 5 & 1 \\ 5 & 9 \end{bmatrix} = 9r - 5, \\ D_3 &= \det \begin{bmatrix} 5 & 1 & 0 \\ 5 & 9 & 5 \\ 0 & 0 & 5 \end{bmatrix} = 45r - 25. \end{aligned}$$

The polynomial $p(z)$ will have roots with negative real part when $D_j > 0$ for $j = 1, 2, 3$. The condition $D_1 > 0$ requires $r > 0$, while each of $D_2 > 0$ and $D_3 > 0$ requires $r > 5/9$, so all roots of $p(z)$ have negative real part exactly when $r > 5/9$. ■

Back to the Three Species Food Chain

The characteristic polynomial for $\mathbf{J}(\mathbf{x}^*)$ in (7.82) is

$$p(\lambda) = \det(\mathbf{J}(\mathbf{x}^*) - \lambda \mathbf{I}) = \lambda^3 + r(1 - 1/K)\lambda^2 + (K + r - 2 - r/K)\lambda + 2r/K$$

where recall we used $a = b = c = d = e = f = 1$ in (7.79). Then $a_1 = r(1 - 1/K), a_2 = (K + r - 2 - r/K)$, and $a_3 = 2r/K$. To apply the Routh-Hurwitz theorem compute

$$\begin{aligned} D_1 &= \frac{r(K-1)}{K}, \\ D_2 &= \frac{r^2(K-1)^2}{K^2}, \\ D_3 &= \frac{r^3(K-1)^3(K-2)}{K^3}. \end{aligned}$$

Recall that $r > 0$ by assumption. It is clear that each of $D_1 > 0, D_2 > 0$, and $D_3 > 0$ holds precisely when $K > 2$. Under these conditions the fixed point at $(K-1, 1, K-2)$ will be asymptotically stable. Since this fixed point is physically relevant only when $K > 2$, we have deduced that it is always stable when it exists, for any $r > 0$.

Reading Exercise 7.6.8 Show that $\mathbf{x}^* = \langle 1, K-1, 0 \rangle$ is a fixed point for (7.79) when $a = b = c = d = e = f = 1$, corresponding to extinction of the third species and coexistence of the first two species. Compute $\mathbf{J}(\mathbf{x}^*)$ using (7.81). Then compute D_1, D_2 , and D_3 in the Routh-Hurwitz theorem

and show that

$$\begin{aligned} D_1 &= \frac{-K^2 + 2K + r}{K}, \\ D_2 &= \frac{r(K(K-2)^2 + r)}{K^2}, \\ D_3 &= -\frac{r^2(K-1)(K-2)(K(K-2)^2 + r)}{K^3}. \end{aligned}$$

Use this (and recall r and K are positive by assumption) to show that this fixed point is asymptotically stable when $r > 0$ and $1 < K < 2$. (Hint: $K(K-2)^2 + r$ is always positive. Also $-K^2 + 2K + r = r + 1 - (K-1)^2$.) What physical interpretation would you give to the requirement that $1 < K < 2$ in order for this situation (species three extinct, species one and two in stable coexistence) to hold?

7.6.4 Exercises

Exercise 7.6.1 Consider a particle of mass m moving along the horizontal (x) axis in accordance with Newton's second law of motion. If no forces act on the particle then $m\ddot{x} = 0$.

- (a) Let $x_1(t) = x(t)$ and $x_2(t) = \dot{x}(t)$. Formulate $m\ddot{x} = 0$ as a coupled pair of first-order ODEs in the form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ where $\mathbf{x} = \langle x_1, x_2 \rangle$.
- (b) Verify that the quantity $P(u_1, u_2) = mu_2$ is a first integral for this system (although m could be omitted). What physical interpretation can you attach to the fact that P remains constant in time?
- (c) Verify that the quantity $E(u_1, u_2) = \frac{m}{2}u_2^2$ is a first integral for this system (although the m or $m/2$ factor could be omitted). What physical interpretation can you attach to the fact that E remains constant in time?

Exercise 7.6.2 A frictionless spring-mass system with mass m and spring constant k obeys $m\ddot{x} + kx = 0$.

- (a) Let $x_1(t) = x(t)$ and $x_2(t) = \dot{x}(t)$. Formulate this ODE as a coupled pair of first-order ODEs in the form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ where $\mathbf{x} = \langle x_1, x_2 \rangle$.
- (b) Verify that the quantity $E(u_1, u_2) = \frac{k}{2}u_1^2 + \frac{m}{2}u_2^2$ is a first integral for this system; recall (7.72). What does this say about the trajectories of the system in the x_1x_2 phase plane?
- (c) Suppose the system now has damping and is governed by $m\ddot{x} + c\dot{x} + kx = 0$. Repeat part (a) with the same function E and show that E acts as a Lyapunov function by showing that $\nabla E \cdot \mathbf{f} = -cu_2^2$, so that $\nabla E \cdot \mathbf{f} \leq 0$ for all u_1 and u_2 . What does this say about this equilibrium point's stability? You may find Theorem 7.6.1 useful, and Theorem 7.6.3 for an even stronger conclusion (using $V = E$ in those theorems).

Exercise 7.6.3 For each system of ODEs below, show that the origin is an isolated fixed point for the system. Then use the suggested form for V to show that V is a Lyapunov function for this fixed point (V may be strict or it may not be), and so deduce what you can about the stability of the origin. Finally, linearize the system at the origin and compute the eigenvalues of the Jacobian. What can be deduced about the stability using the eigenvalues? Does the Lyapunov approach give any information that linearization does not? In each two-dimensional case a direction field may help to visualize the situation.

- (a) $\dot{x}_1 = -x_1^3, \dot{x}_2 = -x_2^3$. Try $V(x_1, x_2) = x_1^2 + x_2^2$.

- (b) $\dot{x}_1 = x_2, \dot{x}_2 = -x_1$ (this system is linear and explicitly solvable). Adjust a, b , and c so that $V(x_1, x_2) = ax_1^2 + bx_1x_2 + cx_2^2$ is a Lyapunov function. Can it be made strict?
- (c) $\dot{x}_1 = -x_1 - 2x_2, \dot{x}_2 = -2x_1 - 4x_2$ (note this system is linear and explicitly solvable). Try $V(x_1, x_2) = x_1^2 + x_2^2$. Hint: $\nabla V \cdot \mathbf{f}$ is a quadratic polynomial in x_1 and x_2 ; factor this polynomial.
- (d) $\dot{x}_1 = -x_1 - 2x_2, \dot{x}_2 = -2x_1 - 4x_2$ (note this system is linear and explicitly solvable). Try $V(x_1, x_2) = x_1^2 + x_2^2$. Hint: $\nabla V \cdot \mathbf{f}$ is a quadratic polynomial in x_1 and x_2 ; factor this polynomial.
- (e) $\dot{x}_1 = -x_2^3, \dot{x}_2 = x_1^3$. Try $V(x_1, x_2) = x_1^4 + x_2^4$.
- (f) $\dot{x}_1 = -2x_1(x_2^4 + x_3^2), \dot{x}_2 = -4x_1^2x_2^3 - 2x_2, \dot{x}_3 = -2x_1^2x_3 - 2x_3$. Try $V(x_1, x_2, x_3) = ax_1^2 + bx_2^2 + cx_3^2$. It should be easy to find choices that make $\nabla V \cdot \mathbf{f} \leq 0$. Can you obtain strict inequality for all nonzero x, y, z ?
- (g) $\dot{x}_1 = -4x_1^3 - 2x_1x_2^2 - 2x_1x_3^2, \dot{x}_2 = -2x_1^2x_2 - 2x_2, \dot{x}_3 = -2x_1^2x_3 - 2x_3$. Try $V(x_1, x_2, x_3) = ax_1^2 + bx_2^2 + cx_3^2$, adjust a, b , and c as necessary.

Exercise 7.6.4 Consider the LSD metabolism model (6.1)-(6.2) from Section 6.1. With $x_1 = u_P$ and $x_2 = u_T$ this model is of the form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ with

$$\mathbf{f}(\mathbf{x}) = \langle -(k_b + k_e)x_1 + k_a x_2, k_b x_1 - k_a x_2 \rangle.$$

Show that the function $V(\mathbf{x}) = k_b x_1^2 + k_a x_2^2$ acts as a strict Lyapunov function for this system at the fixed point $\mathbf{x}^* = \langle 0, 0 \rangle$ and so this fixed point is asymptotically stable. Hint: show that $\nabla V \cdot \mathbf{f} = -2(k_b x_1 - k_a x_2)^2 - 2k_b k_e x_1^2$.

Exercise 7.6.5 Lyapunov's second method works perfectly well for one-dimensional systems. Consider the logistic equation $\dot{x} = f(x)$ where $x = x(t)$, $f(x) = rx(1 - x/K)$ and r and K are positive constants. This system has a fixed point at $x = K$ (as well as $x = 0$). Let $V(x) = (x - K)^2$. Show that $(dV/dx)f(x) < 0$ for x near $x = K$ and so use Theorem 7.6.1 to conclude that this fixed point is asymptotically stable. What is the basin of attraction for this fixed point?

Exercise 7.6.6

- (a) Let $p(z) = z^2 + a_1z + a_2$. Use the Routh-Hurwitz theorem to show that the roots of $p(z)$ have negative real part if and only if $a_1 > 0$ and $a_2 > 0$.
- (b) Let $p(z) = z^3 + a_1z^2 + a_2z + a_3$. Use the Routh-Hurwitz theorem to show that the roots of $p(z)$ have negative real part if and only if $a_1 > 0, a_1a_2 - a_3 > 0$, and $a_3(a_1a_2 - a_3) > 0$. Show these are equivalent to the conditions $a_1 > 0, a_3 > 0$, and $a_1a_2 - a_3 > 0$.
- (c) Let $p(z) = z^4 + a_1z^3 + a_2z^2 + a_3z + a_4$. Use the Routh-Hurwitz theorem to find conditions on a_1, a_2, a_3 , and a_4 that guarantee the roots of p have negative real part.

Exercise 7.6.7 Let \mathbf{A} be a 2×2 matrix with real entries a_{jk} , $1 \leq j, k \leq 2$. Write out the characteristic polynomial of \mathbf{A} and use the Routh-Hurwitz theorem to find conditions on the a_{jk} under which the eigenvalues of \mathbf{A} both have negative real part. Show that this is equivalent to the trace-determinant conclusions of Theorem B.4.1.

Exercise 7.6.8 Consider a quadratic form $Q(x, y) = ax^2 + bxy + cy^2$. We want a simple condition on the constants a, b , and c that guarantee $Q(x, y) < 0$ for all x and y where $(x, y) \neq (0, 0)$; such a quadratic form is said to be **negative definite**. Note that $(x, y) \neq (0, 0)$ means that at least one of x or y is nonzero.

- Suppose Q in part (a) satisfies $Q(x, y) < 0$ for all $(x, y) \neq (0, 0)$. By considering the case in which $x = 1$ and $y = 0$, argue that $Q(1, 0) < 0$ implies $a < 0$.
- Suppose Q in part (a) satisfies $Q(x, y) < 0$ for all $(x, y) \neq (0, 0)$. Consider the case in which $x = -b/2a$ and $y = 1$. Show that $Q(-b/2a, 1) = c - b^2/4a$ and argue that $Q < 0$ implies $b^2 - 4ac < 0$. (Hint: keep in mind that we've shown $a < 0$.)
- Parts (a) and (b) show that if $Q(x, y) < 0$ for all $(x, y) \neq (0, 0)$ then $a < 0$ and $b^2 - 4ac < 0$. In the next two parts we show the converse. Verify that we can write

$$Q(x, y) = a \left(\left(x + \frac{b}{2a}y \right)^2 + \left(\frac{4ac - b^2}{4a^2} \right) y^2 \right).$$

- Use part (c) to argue that if $a < 0$ and $b^2 - 4ac < 0$ then $Q(x, y) < 0$ for all $(x, y) \neq (0, 0)$.

Exercise 7.6.9 Consider two species, a prey species with population $x_1(t)$ and a predator species with population $x_2(t)$. Of course the predator species eats the prey. It is assumed that in the absence of the predator species the prey would grow exponentially, but the presence of the predator negatively impacts the prey population. In the absence of the prey species the predators would slowly starve, but the presence of the prey species gives the predators sustenance.

One model for how the populations change over time is given by the ODE system

$$\begin{aligned}\dot{x}_1 &= ax_1 - bx_1x_2 \\ \dot{x}_2 &= -cx_2 + dx_1x_2\end{aligned}\tag{7.84}$$

where a, b, c , and d are positive constants. These ODEs are known as the **Lotka-Volterra predator-prey equations**. Of course we assume that x_1 and x_2 are nonnegative. (A slightly more refined model of this situation is presented and analyzed in the project “Predator-Prey Model” in Section 7.7.2.)

- Justify each term in the model (7.84), given the assumptions that were made.
- Show that the only fixed points for this system are $(0, 0)$ and $(c/d, a/b)$.
- Compute the Jacobian for this system and show that $(0, 0)$ is always an unstable saddle point. Show that eigenvalue analysis for the stability of $(c/d, a/b)$ fails to determine the stability of this fixed point (the eigenvalues should be purely imaginary).
- Verify that the function $V(x_1, x_2) = dx_1 - c \ln(x_1) + bx_2 - a \ln(x_2)$ is a first integral for this system and that $(c/d, a/b)$ is a strict local minimum for V ; the second derivative test for a function of two variables from multivariable calculus may be helpful. (See Exercise 7.6.11 if you want to know how to come up with this V .) In fact $(c/d, a/b)$ is a global minimum for V in the first quadrant $x_1, x_2 > 0$.
- Let $a = 1, b = 2, c = 1, d = 3$ (nothing special about these values—try your own). Plot $V(x_1, x_2)$ on the range $0 < x_1 < 1, 0 < x_2 < 2$, and also construct a contour plot for V . Argue that this shows the solutions to (7.84) are closed curves, indicating the system has periodic solutions. Further argue that this shows that $(c/d, a/b)$ is stable, but not asymptotically stable.
- Sketch the nullclines for this system in the first quadrant (you can leave a, b, c , and d

undefined), and then use the information from parts (b)-(e) to sketch a phase portrait.

- (g) Critique this model. For example, based on your analysis, can either species ever go extinct?

Exercise 7.6.10 The SIR epidemic model introduced in Section 7.1.1 is a versatile compartmental approach that is easily modified to incorporate a variety of assumptions. Consider an epidemic modeled as in Section 7.1.1 with susceptible (S), infected (I), and recovered (R) categories, but in which

- The overall population $N(t) = S(t) + I(t) + R(t)$ grows according to $\dot{N} = rN$ for some positive growth rate r . The new members of the population enter into the S category.
 - Some of the infected die and are removed from the population.
 - Recovery from the disease does not confer permanent immunity. Rather, members of the R category re-enter the S category at a rate proportional to the number R of recovered individuals.
- (a) Justify the following system of ODEs (7.85) as a model for this situation. In particular, identify what each term means and which terms correspond to the assumptions above. Here a, b, c, d , and r are positive constants.

$$\begin{aligned}\dot{S} &= r(S+I+R) - aSI + dR \\ \dot{I} &= aSI - bI - cI \\ \dot{R} &= bI - dR.\end{aligned}\tag{7.85}$$

- (b) Show that the fixed points for the system (7.85) are

$$(0, 0, 0) \text{ and } \left(\frac{b+c}{a}, \frac{rd(b+c)}{a((c-r)d-rb)}, \frac{rb(b+c)}{a((c-r)d-rb)} \right)$$

- (c) Show that the Jacobian matrix for this system is

$$\mathbf{J}(S, I, R) = \begin{bmatrix} r - aI & r - aS & r + d \\ aI & aS - b - c & 0 \\ 0 & b & -d \end{bmatrix}.$$

Then compute the eigenvalues of $\mathbf{J}(0, 0, 0)$ (they should be simple) and demonstrate that the origin is always unstable for this system.

- (d) Let \mathbf{q} denote the second fixed point in part (b). Let us focus on the specific choices $a = 1/2, b = 1/2, c = 1, d = 1$, and $r = 1$, with e undefined. Show that the resulting fixed point is $\mathbf{q} = \langle 3, 6r/(2-3r), 3r/(2-3r) \rangle$ and that this fixed point is physically relevant when $0 < r < 2/3$.
- (e) Compute $\mathbf{J}(\mathbf{q})$ and show that the characteristic polynomial of this matrix is $p(\lambda) = \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3$ where

$$a_1 = \frac{3r^2 - 2r + 2}{2 - 3r}, \quad a_2 = \frac{11r}{6 - 4r}, \quad a_3 = \frac{3r}{2}.$$

Use the Routh-Hurwitz theorem to show that this fixed point is always stable (when it exists).

- (f) Solve the system (7.85) numerically with $r = 1/2$. The fixed point is $\mathbf{q} = \langle 3, 6, 3 \rangle$, so choose an initial condition near (but not at) that point. Choose some initial conditions farther away. Does this fixed point seem to be globally asymptotically stable? What are the eigenvalues of $\mathbf{J}(\mathbf{q})$ for this choice of r ?
- (g) Solve the system numerically with $r = 1$ and some initial data of your choosing. Since the only physically relevant fixed point $(0, 0, 0)$ is unstable, solutions won't approach the origin. What do they do? Can you provide any rationale for why solutions for small values of r should do one thing while those with larger values do something different?
- (h) Critique this model. What assumptions are missing? (Might people die from causes other than the disease?)

Exercise 7.6.11 Consider an autonomous planar system of ODEs of the form $\dot{x} = f(x, y)$, $\dot{y} = g(x, y)$. How can we find a first integral or conserved quantity for such a system? The following exercises outline a constructive procedure. We begin with a specific pair of ODEs, $\dot{x} = y$, $\dot{y} = -x$.

Suppose that $V(x, y)$ is a function on which all solutions to this ODE system (in some region) are constant. That is, the level curves for $V(x, y)$ are parameterized as $x = x(t)$, $y = y(t)$ where $x(t)$ and $y(t)$ satisfy $\dot{x} = y$, $\dot{y} = -x$.

- (a) Use the fact that on a level curve we have $dy/dx = (dy/dt)/(dx/dt)$ (if dy/dx is defined, which means the curve is not vertical) to show that for the ODE pair $\dot{x} = y$, $\dot{y} = -x$ we have

$$\frac{dy}{dx} = -\frac{x}{y}.$$

This can be considered as a scalar ODE with x as the dependent variable and y as the independent variable.

- (b) The ODE in part (a) is separable. Use separation of variables to solve the ODE in part (a) and show that the level curves for $V(x, y)$ are of the form $x^2 + y^2 = c$ for some constant c (the solution can be left in this implicit form, no need to solve for y as a function of x). As a result we may take $V(x, y) = x^2 + y^2$ as a first integral for this system.
- (c) Repeat this procedure for the system Lotka-Volterra predator-prey system $\dot{x} = ax - bxy$, $\dot{y} = -cy + dxy$ (see Exercise 7.6.9). You may assume any trajectory for the system remains in the first quadrant, so x and y are both positive.

Exercise 7.6.12 Recall from Section 4.5 that a hard spring obeys a displacement/force relation of the form (4.81), or

$$F = k_1x + k_2x^3,$$

for positive constants k_1 and k_2 . This displacement $x(t)$ of the mass m from equilibrium in an undamped spring-mass system with such a spring is governed by the nonlinear second-order ODE

$$m\ddot{x} + k_1x + k_2x^3 = 0. \quad (7.86)$$

- (a) Let $x_1 = x$ and $x_2 = \dot{x}$. Show that (7.86) is equivalent to the system

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{k_1}{m}x_1 - \frac{k_2}{m}x_1^3.\end{aligned}\tag{7.87}$$

- (b) Show that $x_1 = x_2 = 0$ is the unique fixed point for (7.87). Compute the Jacobian of the system at this point and show that its eigenvalues do not determine the stability of this fixed point.
(c) Let

$$E(x_1, x_2) = \frac{k_1}{2}x_1^2 + \frac{k_2}{4}x_1^4 + \frac{m}{2}x_2^2.$$

Use (7.72) to verify that E provides a first integral for this system. (Here E is the total energy of the system, the sum of the kinetic energy $m\dot{x}^2/2$ and the potential energy stored in the compressed spring, $k_1x^2/2 + k_2x^4/4$.)

- (d) Take $m = 1, k_1 = 5$, and $k_2 = 1$ and construct a contour plot of E on the range $-1 \leq x_1, x_2 \leq 1$ to see the level curves of E , or the phase plane trajectories of the hard spring ODE system. How do solutions behave? What do you conclude about the stability of the fixed point $x_1 = x_2 = 0$?
(e) A damped hard spring-mass system obeys $m\ddot{x} + c\dot{x} + k_1x + k_2x^3 = 0$ for some positive constant c . Repeat the analysis of parts (a)-(b) to show that E is a Lyapunov function for this system. Conclude that $x_1 = x_2 = 0$ is an asymptotically stable fixed point (as an eigenvalue analysis should also show). Hint: use Theorem 7.6.3 and mimic its application to the damped nonlinear pendulum.

Exercise 7.6.13 Consider the three species food chain (7.79) whose fixed points are listed in (7.80). We do not assume any specific values for any of the parameters (except that all are positive). The goal here is to show that the fixed point in which all species coexist is always stable, if it exists.

- (a) Argue that the fixed point corresponding to the mutual coexistence of all three species exists if and only if each of $K - af/e > 0$ and $Kbe - fac - de > 0$ holds.
(b) Show that the Jacobian of the system at this fixed point \mathbf{x}^* is given by

$$\mathbf{J}(\mathbf{x}^*) = \begin{bmatrix} -r(1 - af/Ke) & -ra(1 - af/Ke) & 0 \\ bf/e & 0 & -cf/e \\ 0 & (Kbe - abf - de)/c & 0 \end{bmatrix}.$$

Then show that the characteristic polynomial of this matrix is $p(\lambda) = \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3$ where

$$\begin{aligned}a_1 &= r \left(1 - \frac{af}{Ke} \right) \\ a_2 &= \frac{(bK^2e^2 - Ke(de + ab(f - r)) - a^2bfr)f}{Ke^2} \\ a_3 &= \frac{((Kb - d)e - abf)(Ke - af)rf}{Ke^2}.\end{aligned}$$

- (c) Compute D_1, D_2 , and D_3 according to (7.83) and show that

$$\begin{aligned}D_1 &= r \left(1 - \frac{af}{Ke}\right) \\D_2 &= \frac{r^2 fab}{e} \left(1 - \frac{af}{Ke}\right)^2 \\D_3 &= \frac{r^3 abf^2(Kbe - abf - de)}{e^2} \left(1 - \frac{af}{Ke}\right)^3.\end{aligned}$$

Then use the results of part (a) to show that if this fixed point exists then $D_1 > 0, D_2 > 0$, and $D_3 > 0$ holds, and so this fixed point is asymptotically stable.

Exercise 7.6.14 Suppose that $\mathbf{x} = \mathbf{x}^*$ is an equilibrium solution for a system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ of ODEs. Let E be a first integral for this system and suppose that \mathbf{x}^* is a strict local minimum for E . Show that \mathbf{x}^* cannot be asymptotically stable, by showing that no solution that starts near \mathbf{x}^* can approach \mathbf{x}^* . Hint: Suppose \mathbf{x}^* is asymptotically stable. Then we can find a nearby point \mathbf{x}_0 such that the solution to $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ with $\mathbf{x}(0) = \mathbf{x}_0$ satisfies $\mathbf{x}(t) \rightarrow \mathbf{x}_0$ as t approaches infinity. Since E is continuous this means $E(\mathbf{x}(t)) \rightarrow E(\mathbf{x}_0)$ as t approaches infinity. Why is this not possible?

7.7 Modeling Projects

7.7.1 Project: Homelessness Revisited

Recap of the Linear Model

In the Section 6.5 modeling project “Homelessness” we examined a linear system of ODEs to model the fraction of homeless people in a city. You may find it helpful to review that project. We made the following definitions:

$R(t)$ is the number of renting households at time t ,

$E(t)$ is the number of evicted households at time t .

We assumed that there were a total of N households, so $R + E = N$ at all times.

A Nonlinear Model

In our first model we assumed that a constant fraction $\alpha \frac{R(t)}{N}$ of the renting group transitioned to the evicted group each year, regardless of the vacancy rate, and that a fraction $\beta \frac{E(t)}{N}$ transitioned from the evicted to renting group. We now modify these assumptions. First, let us suppose that the city has exactly M rental units (apartments and houses), and probably $M \neq N$. If $M < N$ then there aren’t enough rentals to accommodate everyone, and if $M > N$ there is enough housing for everyone.

In this part of the project we will work with the actual count of renting and evicted households, $R(t)$ and $E(t)$, respectively, rather than the fractions $r(t) = R(t)/N$ and $e(t) = E(t)/N$ as we did in Section 6.5. We ask you to construct a model for $\frac{dR}{dt}$ and $\frac{dE}{dt}$ that satisfies the following assumptions:

- The flow rate from the renting group to the evicted group increases as the number of vacancies decreases (namely, as R gets close to M).
- The flow rate from the evicted group to the renting group decreases as the number of vacancies decreases.

Modeling Exercises

Modeling Exercise 7.1.1 Suppose your roommate suggests using $-\alpha \frac{R^2}{M}$ for the flow rate from the renting group to the evicted group. Does your roommate's suggested flow rate satisfy the first assumption? If so, explain why it does. If not, suggest a different formula and show that it satisfies the first assumption.

Modeling Exercise 7.1.2 Find a formula that can be used for the flow rate from the evicted group to the renting group. Explain why your formula satisfies the second assumption.

Modeling Exercise 7.1.3 Use your flow rates from Modeling Exercises 7.1.1 and 7.1.2 to formulate a coupled pair of ODEs for $R(t)$ and $E(t)$, based on the compartmental reasoning of Figure 6.10. Hint: based on the assumptions we've made, why should $dR/dt + dE/dt = 0$ be true? What does this imply about $R(t) + E(t)$ as a function of time?

Modeling Exercise 7.1.4 If your model in Modeling Exercise 7.1.3 is not autonomous, why isn't it? If it is autonomous, sketch a phase portrait for the model. Confine your attention to the first quadrant in the RE plane, with R as the horizontal axis. You need not assume specific values for the positive parameters α, β, M , or N ; try to sketch a phase portrait that does not rely on specific choices for these parameters. How does the phase portrait depend on these parameters? For example, how does the stability of any fixed points depend on α, β, M , and N ? Does your phase portrait guarantee that R and E remain nonnegative?

Modeling Exercise 7.1.5 Suppose that $M = 118500$. Choose a value for α from the 2016 data for city evictions from the Eviction Lab [85], and state which city's information you chose. Set $\beta = \frac{1}{3}\alpha$, as a start. Use technology to sketch the phase plane and the solution curves that satisfy the initial conditions $R(0) = 112100$ and $E(0) = 5900$ (so initially, 95% of the non-homeowners are renting, and the total number of renting households is $N = 112100 + 5900 = 118000$). What does this model predict about the long term percentages of non-homeowner households who are renting and who are evicted? How does this compare to your results in the first model (if you did that project)? Experiment by varying the parameters.

7.7.2 Project: Predator-Prey Model

In this project we analyze one model of an ecosystem in which two species exist. Unlike the competing yeast species of Section 7.1, the setting is asymmetric: one species is the prey, the other is the predator that eats the prey.

Predator and Prey Species

Consider a prey species with population $x_1(t)$. Suppose that, in the absence of any predator, the prey species grows according to the logistic equation (1.10), as

$$\dot{x}_1 = r_1 x_1 (1 - x_1/K) \tag{7.88}$$

where r_1 is a positive intrinsic growth rate for the prey species and K is the carrying capacity of the environment for this species. However, if the predator species is present with population $x_2(t)$ then this negatively impacts the prey species' growth. We modify (7.88) in the same spirit as was done in (7.3) or (7.4). Specifically, we now take

$$\dot{x}_1 = r_1 x_1 \underbrace{\left(\frac{K - x_1(t) - ax_2(t)}{K} \right)}_{f_1(x_1, x_2)} \tag{7.89}$$

for some positive constant a .

Modeling Exercise 7.2.1 Provide some justification for (7.89). What does the constant a quantify? Examine how \dot{x}_1 behaves when $x_2 \approx 0$ versus when x_2 is very large. What is the physical interpretation of each situation?

We assume that the predator species relies on the prey species as its sole source of food; in the absence of any prey the predator population would dwindle according to $\dot{x}_2 = -r_2 x_2$ for some positive constant r_2 , as the predators starve to death. However, if $x_1 > 0$ then the predators are sustained by this food source and $\dot{x}_2 = -r_2 x_2$ is modified according to

$$\dot{x}_2 = \underbrace{(-r_2 + bx_1)x_2}_{f_2(x_1, x_2)} \quad (7.90)$$

for some positive constant b . The ODEs (7.89)-(7.90) are a variation on the traditional **Lotka-Volterra predator-prey equations**. See also Exercise 7.6.9 for the usual formulation of these equations (which assume exponential rather than logistic growth for the prey population.)

Modeling Exercise 7.2.2 Provide some justification for (7.90). What does the constant b quantify? Examine how \dot{x}_2 behaves when $x_1 \approx 0$ versus when x_1 is very large. What is the physical interpretation of each situation?

Equilibrium Solutions

We will confine our attention to the case in which $x_1, x_2 \geq 0$, since these variables quantify populations.

Modeling Exercise 7.2.3 Show that the equilibrium solutions for the system (7.89)-(7.90) are

$$(0,0), (K,0), \left(\frac{r_2}{b}, \frac{Kb - r_2}{ab} \right). \quad (7.91)$$

What is the physical interpretation of each fixed point in (7.91)? Why is the last fixed point relevant only when $Kb - r_2 > 0$?

How do each of K , b , and r_2 influence whether the inequality $Kb - r_2 > 0$ holds, and why does this make physical sense?

The Case $Kb - r_2 > 0$

Suppose that $Kb - r_2 > 0$, so there is an equilibrium solution in which the predators are not extinct.

Modeling Exercise 7.2.4 Show that the x_1 nullcline defined by $f_1(x_1, x_2) = 0$ consists of the vertical axis $x_1 = 0$ and the line $x_2 = K/a - x_1/a$. Sketch this nullcline for $x_1, x_2 \geq 0$, and label the intercepts of the line $x_2 = K/a - x_1/a$ in terms of K and a . Then draw appropriate arrows to indicate the horizontal motion of solutions in each region into which the nullcline divides the plane.

Modeling Exercise 7.2.5 Show that the x_2 nullcline defined by $f_2(x_1, x_2) = 0$ consists of the horizontal axis $x_2 = 0$ and the vertical line $x_1 = r_2/b$. Sketch this nullcline for $x_1, x_2 \geq 0$, and label the x_1 intercept in terms of r_2 and b . Then draw appropriate arrows to indicate the vertical motion of solutions in each region into which the nullcline divides the plane.

Modeling Exercise 7.2.6 Use the Jacobian matrix to linearize the system at each equilibrium solution $(0,0)$ and $(K,0)$ and use this to show that both of these are saddle points for any choice of r_1, r_2, a, b , and K in which all of the parameters are positive and $Kb - r_2 > 0$.

Modeling Exercise 7.2.7 Use the Jacobian matrix to linearize the system at the equilibrium solution $\left(\frac{r_2}{b}, \frac{Kb - r_2}{ab} \right)$ and use this to show that this fixed point is either a stable node or stable spiral point in the case under consideration (namely, $Kb - r_2 > 0$). Theorem B.4.1 in Section B.4 will be extremely helpful.

Modeling Exercise 7.2.8 Based on your analysis, sketch a typical phase portrait for this system under the assumption that $Kb - r_2 > 0$. What is the long-term fate of each species?

The Case $Kb - r_2 < 0$

Modeling Exercise 7.2.9 Analyze the case in which $Kb - r_2 < 0$. What happens to each species' population in the long run?

7.7.3 Project: Parameter Estimation for Competing Yeast Species

In this project we will consider the estimation of the parameters r_1 , K_1 , r_2 , K_2 , a , and b in the competing species model (7.3)-(7.4) from data collected for two competing species of yeast.

The data in Table 7.3 comes from [49]. The data comes from several different experiments performed by Gause concerning the populations of two species of yeast, *saccharomyces cerevisiae* and *schizosaccaromyces kefir*. In each experiment a nutrient-filled vessel was inoculated with a fixed amount of either the *saccharomyces* species, the *schizosaccaromyces* species, or both. The population of each species was measured at the listed times in column 1 of Table 7.3, though not all times have data points for each species. Moreover, Gause measured the volume of yeast cells present and used volume as a proxy for the actual yeast population; see [49] for precise experimental procedures.

	<i>saccharomyces</i>	Mixed Population	<i>schizosaccharomyces</i>	Mixed Population
Age in hours	Volume of yeast	Volume of yeast	Volume of yeast	Volume of yeast
6	0.37	0.375	-	0.291
16	8.87	3.99	1.00	0.98
24	10.66	4.69	-	1.47
29	12.50	6.15	1.70	1.46
40	13.27	-	-	-
48	12.87	7.27	2.73	1.71
53	12.70	8.30	-	1.84
72	-	-	4.87	-
93	-	-	5.67	-
117	-	-	5.80	-
141	-	-	5.83	-
7.5	1.63	0.923	-	0.371
15.0	6.20	3.082	1.27	0.630
24.0	10.97	5.780	-	1.220
31.5	12.60	9.910	2.33	1.112
33.0	12.90	9.470	-	1.225
44.0	12.77	10.570	-	1.102
51.5	12.90	9.883	4.56	0.961

Table 7.3: The growth of the yeast volume and the number of cells in pure cultures of *saccharomyces cerevisiae* (column 1), *schizosaccaromyces kefir* (column 3) and in the mixed population of these species (column 2 and 4 respectively). [49, p. 395]

The data in column 2 of Table 7.3 are for the *saccharomyces* species alone in the vessel, while column 4 is for the *schizosaccaromyces* species alone in the vessel. Column 3 tabulates the *saccharomyces* population when both yeast species are present, and Column 5 tabulates the *schizosaccaromyces* population when both yeast species are present. Finally, the data represents

a number of different experiments. In the first series of three experiments, either *saccharomyces* alone, or *schizosaccaromyces* alone, or both, were grown and their populations measured at the times 6, 16, ..., 141 hours listed in the table. These are the first eleven rows in Table 7.3. Then another set of three experiments was performed under identical conditions, with the results tabulated in the last seven rows, at times 7.5, 15, ..., 51.5 hours. Since the experimental conditions were the same in each series, we will amalgamate the data in each column. That is, we will assume the data for *saccharomyces* alone was taken at time 6, 7.5, ..., 141 hours, and similarly for the other experimental configurations.

Estimating Parameters for Each Species

In this section we will use the data in Table 7.3 to estimate the parameters r_1 and K_1 for the *saccharomyces* species when it is alone in the culture, then r_2 and K_2 for the *schizosaccaromyces* species when alone, and finally the competition parameters a and b in (7.3)-(7.4) when both species are present. Gause had no computer to aid his analysis. He relied on basic algebra, estimation of slopes, and graphical procedures to obtain these estimates.

We will assume that when the *saccharomyces* species alone is present, the population grows in accordance with logistic equation. If $u_1(t)$ denotes the yeast population then

$$\dot{u}_1 = r_1 u_1 (1 - u_1/K_1) \quad (7.92)$$

where r_1 is the growth rate and K_1 the carrying capacity of the environment. Recall from Section 2.2.5 that the solution to (7.92) with $u(0) = u_0$ is $u_1(t) = K_1 u_0 / (u_0 + e^{-r_1 t} (K_1 - u_0))$. However, we don't have initial data at time $t = 0$ in Table 7.3, but rather initial data $u(t_0) = u_0$. In this case the solution to (7.92) is

$$u_1(t) = \frac{K_1 u_0}{u_0 + e^{-r(t-t_0)} (K_1 - u_0)}. \quad (7.93)$$

Modeling Exercise 7.3.1 Use the data in column 2 of Table 7.3, along with (7.93), to estimate the parameters K_1 and r_1 . A quick glance at the data should make the approximate value of K_1 obvious. You can fit the parameters visually (guess and plot) or use the least-squares procedure of Section 3.4. Gause obtained estimates $r_1 = 0.21827$ and $K_1 = 13.0$, but these may not be the best.

Modeling Exercise 7.3.2 Repeat Modeling Exercise 7.3.1 for the *schizosaccaromyces* species, using the data in column 4 of Table 7.3, to estimate the growth rate r_2 and carrying capacity K_2 for this species. Gause obtained estimates $r_2 = 0.06069$ and $K_2 = 5.8$, but your estimate may well be different.

Estimating the Competition Parameters a and b

Modeling Exercise 7.3.3 Fix the values for r_1 , K_1 , r_2 , and K_2 as you obtained them in Modeling Exercises 7.3.1 and 7.3.2. Use the data in columns 3 and 5 of Table 7.3 to estimate a and b in (7.3)-(7.4). For this case, unlike the previous two Modeling Exercises, we do not have the luxury of a closed-form solution. A numerical solution to (7.3)-(7.4) may be required, in conjunction with plotting. Gause estimated $a = 3.15$ and $b = 0.439$, but these estimates may depend to some extent on his approach.

Modeling Exercise 7.3.4 Based on your estimates for r_1 , K_1 , r_2 , K_2 , a , and b and the analysis in Section 7.4, what would be the long-term fate of each species competing in this setting? How sensitive is this conclusion to the values of a and b you obtained?

8. An Introduction to Partial Differential Equations

8.1 Conservation of Stuff and the Continuity Equation

Conservation laws are central to many mathematical models of the world. By *conservation law* we mean a principle that dictates that in a certain situation of interest there is something—mass, electric charge, heat energy, momentum, or other—that is neither created nor destroyed. Such a conservation law can often be used to say something about how the situation evolves over time. We've already employed this philosophy a number of times, for example, in the various pharmacokinetic models in Sections 1.2 and 6.1.1. Even in settings where such quantities are created or destroyed, a careful accounting of how this occurs may still allow us to make conclusions. This section is devoted to developing these ideas, in the form of the **continuity equation**, a basic tool in mathematical modeling that can be used to describe incredibly diverse physical scenarios.

8.1.1 Industrial Furnaces and Metal Production

This introductory material is based on the SIMIODE modeling project “It’s a Blast (Furnace)!” [33].

Refined metal is foundational for modern civilization, and industrial furnaces are central to the production of refined metals. These furnaces come in many variations: some are used for the extraction of metal from raw ore (a blast furnace), others to further refine already existing metals (for example, make steel from iron), others to reprocess recycled material. There are a wide variety of furnace types and processes.

Whichever type of furnace and process is used, molten material must be contained in a large vessel, as illustrated in the left panel in Figure 8.1. One natural question comes to mind: why don’t the walls of the furnace melt? The answer is that in many cases the vessel, though it may have walls that are themselves made of metal, is lined with a *refractory material*, a substance that is resistant to the harsh internal environment of the furnace and protects the walls from the molten charge inside. The refractory material must obviously resist high heat, but also mechanical wear and chemical attack. Refractory material might be something as simple as brick or other durable minerals. The walls may also be actively cooled, but we will not model that here.

Nonetheless, the refractory lining can wear and break down over time. If the lining becomes

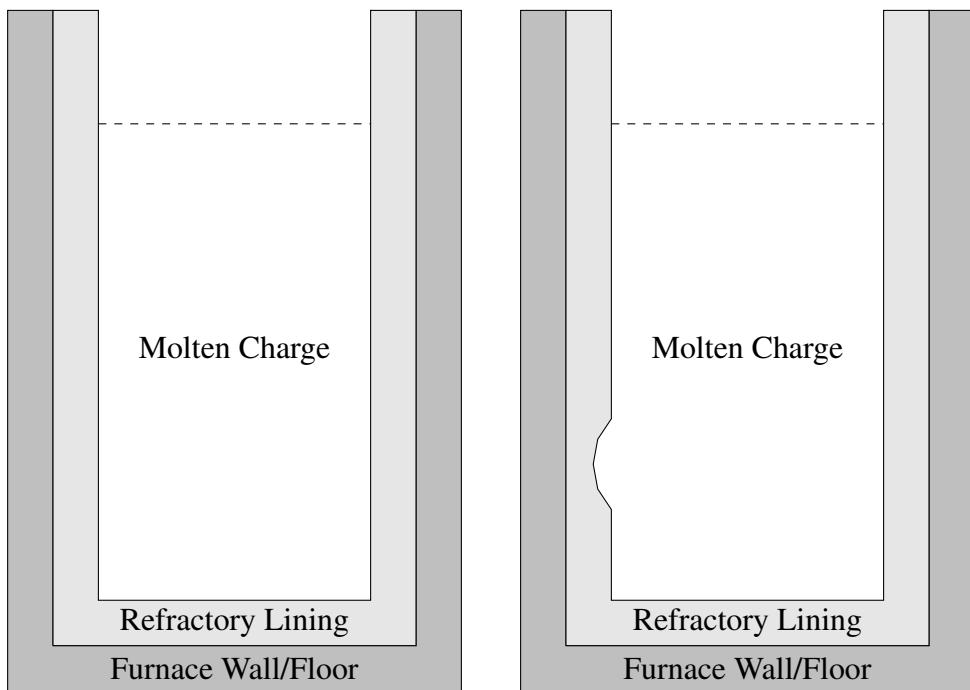


Figure 8.1: Left panel: Schematic of industrial furnace. Right panel: Furnace with thinning region in lower left refractory material.

too thin in any location, as illustrated in the right panel of Figure 8.1, this may result in serious and dangerous damage to the walls of the furnace, so it's useful if one can monitor the condition of the refractory lining. This can be done if the furnace is emptied and not in operation, but shutting down the furnace is usually unacceptable. It would be better if the lining condition could be determined while the furnace is in operation. This obviously can't be done in any direct fashion. Can it be done from outside the furnace?

One approach is this: if the lining or wall of the furnace begins to thin in a localized area, we might expect that the corresponding portion of the furnace wall on the outside would develop a hot spot, a region where the outer wall of the furnace is hotter than the surrounding area. Given that the outside of the furnace is much more easily accessible than the inside, perhaps this can be exploited to infer the condition of the inner lining. In fact, to a limited extent the temperature inside the furnace wall can be measured, just not too close to the molten charge inside. One way to do this is to embed temperature sensors (for example, a thermocouple, a passive electrical device that can measure temperature) on or even inside the wall of the furnace; see [102] or [44]. Temperature data can be collected from these sensors and used to estimate the condition of the furnace wall and lining.

To do this estimation accurately requires a quantitative understanding of how the temperature of the wall behaves. More generally, we want a model that describes how heat energy flows through material objects. Such a model is developed in this section. The main principle used will be a conservation law, similar to those we've seen before when dealing with physical systems like compartmental models (in which some substance was conserved, or at least its gain or loss accounted for), or systems like the frictionless pendulum, in which energy is conserved. However, the conservation philosophy and modeling principles we employ will have application far beyond the flow of heat. We will return to modeling the furnace in the project “It’s a Blast (Furnace)!” in Section 8.5.1 at the end of the chapter.

8.1.2 Conservation of Stuff

Conduits and Stuff

Our focus will be on using a conservation philosophy to analyze the flow of something—which we refer to simply as “stuff”—through a thin conduit of indeterminate length. Refer to Figure 8.2. We use x to indicate spatial position along the conduit with respect to some origin $x = 0$ and t to indicate time. We assume that the conduit is thin enough that the flow may be considered one-dimensional, purely along the x axis.

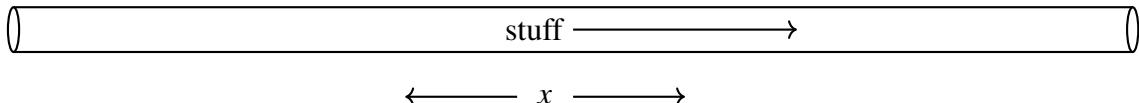


Figure 8.2: A thin conduit through which stuff flows; the variable x indicates horizontal position.

This framework is quite flexible. The conduit and stuff might be any of the combinations shown in Table 8.1. Many other possibilities exist. In each case stuff will possess some natural physical dimension, for example, volume or mass for water, electric charge for electrons, a simple numerical count (perhaps dimensionless) for cars, mass for pollution, and, of course energy for thermal energy.

Conduit	Stuff
pipe	water
wire	electrons
road	cars
river	pollutant
metal bar	thermal energy

Table 8.1: Conduits and stuff.

Reading Exercise 8.1.1 Think of another physical scenario in which stuff flows through a conduit. What is the conduit? What is the stuff? What is the physical dimension for the stuff?

Stuff is Conserved

Our guiding principle for modeling the flow of stuff through the conduit will be this:

Inside the conduit stuff is neither created nor destroyed.

That is to say, stuff is conserved. To leverage this principle for modeling we introduce two functions.

- $\rho(x, t)$ is the *stuff density*, the amount of stuff per unit length of the conduit at position x , time t . If stuff has the physical dimension of S then ρ has the dimension of SL^{-1} .
- $q(x, t)$ is the *stuff flux*, the amount of stuff per unit time flowing in the direction of increasing x . In Figure 8.2 this is from the left to the right past position x at time t . If stuff has the physical dimension of S then q has the dimension of ST^{-1} . If $q(x, t) < 0$ this means stuff is flowing in the direction of decreasing x at position x and time t .

For example, if the conduit is a pipe and the stuff is water measured on a per mass basis then ρ has dimension ML^{-1} , mass per length. The function q has dimension MT^{-1} , mass per time. For now assume that ρ and q and their first derivatives are continuous functions.

Reading Exercise 8.1.2 For each scenario listed in Table 8.1, state the dimension of ρ and q . Use Q to denote electric charge, assume that pollutant is measured in mass (which would be dissolved in water, but we are only interested in the pollutant), and that cars are measured with a simple count, which is dimensionless.

8.1.3 The Continuity Equation

The Control Volume

Let us focus our attention on a short length of the conduit from $x = x_0$ to $x = x_0 + \Delta x$ where Δx is positive, as illustrated in Figure 8.3. We will use Ω to denote this short region of the conduit; note this region is a purely hypothetical construct, to be used for analytical purposes. The region Ω is called a **control volume** (though in this case **control length** might be more accurate). We are going to compute the rate at which the amount of stuff in Ω is changing with respect to time in two different ways, one involving ρ and one involving q . Carrying out this careful accounting will yield an important relationship between the stuff density function ρ and the stuff flux function q , that holds in all situations in which stuff is conserved.

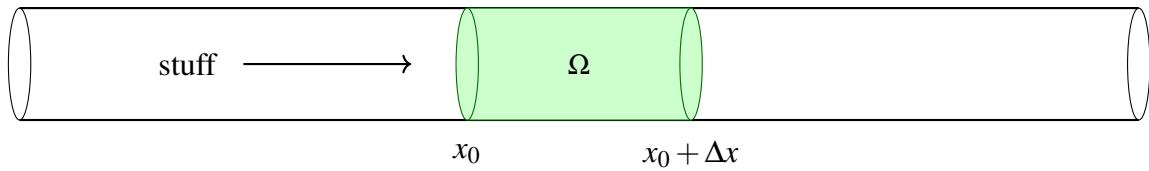


Figure 8.3: Control volume (or length) Ω inside conduit.

Stuff Rate of Change Computation I

Given that ρ is the local density with the dimension of stuff per length, the amount of stuff in Ω at time t is given by

$$\text{stuff in } \Omega = \int_{x_0}^{x_0 + \Delta x} \rho(x, t) dx. \quad (8.1)$$

Convince yourself this is dimensionally and physically correct. As a result, the rate at which the amount of stuff in Ω is changing with respect to time can be obtained by differentiating both sides of (8.1) with respect to t to find

$$\begin{aligned} \frac{d(\text{stuff in } \Omega)}{dt} &= \frac{d}{dt} \int_{x_0}^{x_0 + \Delta x} \rho(x, t) dx \\ &= \int_{x_0}^{x_0 + \Delta x} \frac{\partial \rho}{\partial t}(x, t) dx \end{aligned} \quad (8.2)$$

where we assume the $\partial/\partial t$ derivative can be passed to the inside of the integral. Equation (8.2) quantifies the rate at which the amount of stuff inside Ω is changing in terms of the function ρ .

Stuff Rate of Change Computation II

Let us now compute the rate at which the amount of stuff inside Ω is changing by using the function q . Consider the rate at which stuff is entering or leaving Ω at the points $x = x_0$ and $x = x_0 + \Delta x$. These are the only places stuff can enter or exit, since we assume stuff cannot escape through the conduit walls. Since q quantifies the rate at which stuff flows in the direction of increasing x , the rate at which stuff is entering Ω at $x = x_0$ at time t is $q(x_0, t)$ (though if $q(x_0, t)$ is negative then stuff is flowing to the left and out of Ω .) Similarly, the rate at which stuff is *entering* Ω at $x = x_0 + \Delta x$ is given by $-q(x_0 + \Delta x, t)$. Be careful: the minus sign on $-q(x_0 + \Delta x, t)$ is needed, since we are quantifying the rate at which stuff enters (not exits). All in all the net rate at which stuff enters Ω is given by

$$\text{net rate stuff enters } \Omega = q(x_0, t) - q(x_0 + \Delta x, t). \quad (8.3)$$

The claim is that the rate at which stuff enters Ω in (8.3) must equal the rate at which the amount of stuff in Ω is changing as quantified by (8.2), but this claim hinges upon the crucial assumption that stuff is neither created nor destroyed in Ω .

Reading Exercise 8.1.3 You are in charge of security in a high-tech building in which all entrances and exits are monitored. Moreover, infrared cameras keep track of how many people are in the building at all times. Suppose that at time $t = t_0$ hours there are 19 people in the building. Over the course of the next hour 14 people enter the building and 7 people exit. Argue that at time $t = t_0 + 1$ there are 26 people in the building. What assumption have you made? Hint: what if the infrared cameras indicated there were 27 people in the building at time $t_0 + 1$? What if there were 25? What could account for this? If there is a discrepancy, what kind of building might this be?

Conservation Consequences: The Continuity Equation

Reading Exercise 8.1.3 should help to convince you that the assertion that “The rate of change of stuff in Ω equals the net rate stuff enters Ω ” relies on the assumption that stuff is neither created nor destroyed in Ω , that is, *stuff is conserved*. With this principle we can use (8.2) and (8.3) to assert that

$$\int_{x_0}^{x_0 + \Delta x} \frac{\partial \rho}{\partial t}(x, t) dx = q(x_0, t) - q(x_0 + \Delta x, t). \quad (8.4)$$

Divide both sides of (8.4) by Δx and a bit of rearrangement leads to

$$\frac{1}{\Delta x} \int_{x_0}^{x_0 + \Delta x} \frac{\partial \rho}{\partial t}(x, t) dx + \frac{q(x_0 + \Delta x, t) - q(x_0, t)}{\Delta x} = 0. \quad (8.5)$$

According the mean value theorem for integrals

$$\frac{1}{\Delta x} \int_{x_0}^{x_0 + \Delta x} \frac{\partial \rho}{\partial t}(x, t) dx = \frac{\partial \rho}{\partial t}(x^*, t) \quad (8.6)$$

for some x^* such that $x_0 < x^* < x_0 + \Delta x$. In this case (8.6) in conjunction with (8.5) shows that

$$\frac{\partial \rho}{\partial t}(x^*, t) + \frac{q(x_0 + \Delta x, t) - q(x_0, t)}{\Delta x} = 0. \quad (8.7)$$

Now let Δx approach 0. The first term on the left in (8.7) approaches $\frac{\partial \rho}{\partial t}(x_0, t)$ (assume $\partial \rho / \partial t$ is a continuous function of x). The second term on the left in (8.7) approaches $\frac{\partial q}{\partial x}(x_0, t)$. Since x_0 is an arbitrary point in the conduit and t is an arbitrary time the conclusion is that

$$\frac{\partial \rho}{\partial t} + \frac{\partial q}{\partial x} = 0 \quad (8.8)$$

at all points in the conduit and times where stuff is conserved.

Equation (8.8) is a fundamental relation between ρ and q that holds in all situations in which the stuff in the conduit is conserved, and is known as the **continuity equation**. The continuity equation is an example of a **partial differential equation** (PDE), an equation involving one or more unknown functions of two or more variables (in this case ρ and q) and their partial derivatives. The continuity equation (8.8) can be adapted to situations in which stuff is not conserved as well; see Exercise 8.1.7.

It is surprising that the continuity equation applies to each of the very diverse physical situations listed in Table 8.1. Thermal energy in a metal bar does not behave at all like cars moving down a road. How can one equation allow us to model both situations? The answer is: it can't, at least not all by itself. We need more information that is specific to the situation being modeled. This additional information takes the form of an additional relation between ρ and q that captures some essential physics about the situation, beyond simple conservation of stuff. This additional relation is called a *constitutive relation*. An initial condition is also required, and possibly some *boundary conditions*. We will illustrate by using the continuity equation to model heat flow.

8.1.4 The Heat Equation

Thermal Energy Density

Let the conduit now be a material bar of some indeterminate length. Assume the bar is capable of conducting thermal energy, which is the stuff in this setting. The function ρ is the density of thermal energy in the bar, on a per length basis. To make this more precise, consider a short piece of the bar spanning the interval $x - \Delta x/2$ to $x + \Delta x/2$. At time t this piece of the bar contains an amount $E(x, \Delta x, t)$ of thermal energy. Then $\rho(x, t)$ is the amount of thermal energy in this short span at time t , divided by Δx , in the limit that Δx approaches zero, that is,

$$\rho(x, t) = \lim_{\Delta x \rightarrow 0} E(x, \Delta x, t) / \Delta x. \quad (8.9)$$

We assume this limit exists.

The quantity $q(x, t)$ is the rate at which thermal energy is flowing from left to right past at position x at time t . If thermal energy is conserved then the continuity equation (8.8) holds. We now seek a second relation between ρ and q that is independent of (8.8), a relation specific to the flow of heat energy. This second relation will be used in conjunction with (8.8) to show that $\rho(x, t)$ must satisfy a certain partial differential equation, the *heat equation*. Our derivation of the heat equation will be specific to one space dimension, though the derivation for two and three space dimensions is very similar.

Thermal Energy Density and Temperature

It will ultimately be more convenient to work with the bar's temperature $u(x, t)$ as a function of position and time, rather than the energy density $\rho(x, t)$, so let's examine a bit of physics that relates the temperature $u(x, t)$ to the thermal energy density $\rho(x, t)$. Let us suppose we have fixed a temperature scale (Celsius, Fahrenheit, Kelvin, or other). Suppose that some mass m is at a uniform temperature of u_0 degrees in this scale. This mass contains a certain amount of thermal energy, say E_0 (the units would be joules in the SI system). Suppose the temperature of the mass is raised by an amount Δu degrees to $u_0 + \Delta u$ degrees. The conventional model for the relation between energy and temperature states that the thermal energy of the mass changes by an amount ΔE that is proportional to Δu and also to m , so that

$$\Delta E = cm\Delta u \quad (8.10)$$

where c is a positive physical constant that depends on the material of which the mass is made. This constant c is called the **specific heat** of the material. By considering $\Delta u = 1$ in (8.10) we can write $c = \Delta E/m$ and so interpret c as the amount of thermal energy needed to raise the temperature of the material by one degree, on a per mass basis. The specific heat c can vary with temperature too, but remains fairly constant for most substances if Δu is not too large. The thermal energy E contained in the mass at temperature $u = u_0 + \Delta u$ is then given by $E = E_0 + \Delta E$. Making use of (8.10) then yields

$$E = E_0 + cm(u - u_0). \quad (8.11)$$

From (8.11) it is clear that the thermal energy E contained in a mass m is a linear function of the temperature u of the mass.

Let's rewrite (8.11) in terms of energy density instead of energy. Suppose that the one-dimensional bar has a constant linear mass density of λ units of mass per unit length. Consider a short section of this bar of length Δx and at temperature u , with Δx small enough that both the temperature u and energy density ρ of this piece can be considered as constant with respect to x . The mass of this portion of the bar is given by $m = \lambda \Delta x$ and from (8.11) it follows that the energy E contained in this piece of the bar is given by

$$E = E_0 + c(u - u_0)\lambda \Delta x,$$

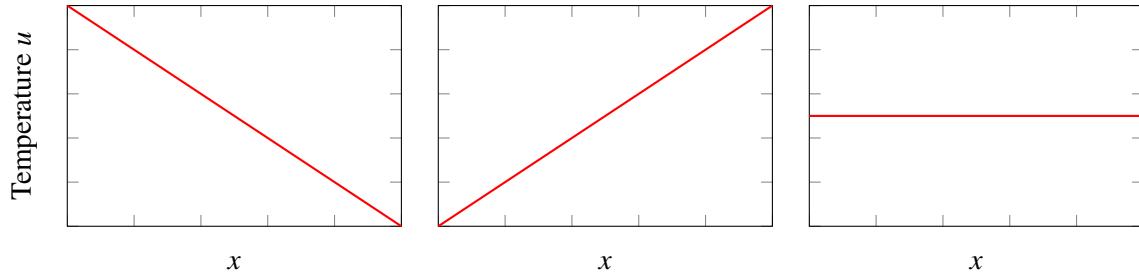


Figure 8.4: Left panel: temperature (or thermal energy density) $u(x, t)$ decreases with respect to x . Middle panel: temperature increases with respect to x . Right panel: constant temperature with respect to x . In each case time is fixed at some value $t = t^*$.

where E_0 is the amount of energy this piece contains at temperature u_0 . Divide this equation through by Δx and let Δx approach 0 to obtain

$$\rho = \rho_0 + c\lambda(u - u_0) \quad (8.12)$$

where ρ is defined by (8.9) and ρ_0 is defined similarly, as the density of thermal energy but when the bar is at constant temperature u_0 throughout. We may consider u_0 and ρ_0 as constant reference quantities that depend on the temperature scale we adopt. Equation (8.12) gives a simple linear relationship between the thermal energy density ρ and the temperature u of the bar, that involves the bar's specific heat c and mass density λ , as well as u_0 and ρ_0 .

Temperature and the Flow of Thermal Energy

Let's examine how the heat flux q depends on ρ or u . We begin with a thought experiment (often referred to as a **Gedankenexperiment**, from German, an activity supposedly used by Einstein when first formulating special relativity) to posit a relation between ρ (or u) and q . Consider the panels in Figure 8.4. In each panel the horizontal axis is the spatial or x position along the bar through which heat flows. The vertical axis quantifies the temperature u of the bar. In each case the graph is a snapshot at a specific time $t = t^*$ and shows the temperature $u(x, t^*)$ of the bar.

Consider what the heat flux q would look like for the temperature profile in the left panel of Figure 8.4. Given our intuition that heat flows from hotter regions to colder regions, the thermal energy in this case should be flowing to the right, in the positive x direction, and so $q(x, t^*) > 0$ at each x coordinate. In the middle panel the same reasoning shows that thermal energy should be flowing from right to left, so $q(x, t^*) < 0$. In the right panel the bar is at a constant temperature; there is no hotter or colder region and thermal energy has no impetus to flow, so $q(x, t^*) = 0$ at all points. A simple model that captures these observations is that

$$q(x, t) = -k \frac{\partial u}{\partial x}(x, t) \quad (8.13)$$

for some nonnegative constant k .

The relation (8.13) is known as **Fourier's law** and models the flow of heat energy as occurring from hotter to colder regions, in proportion to the steepness $\partial u / \partial x$ of the temperature gradient. The constant k depends on the material from which the bar is made and is called the material's **thermal conductivity**. The larger the value of k , the more heat energy flows per unit time for a given temperature gradient. For values of k close to zero, the heat flux is smaller. The extreme case $k = 0$ corresponds to zero heat energy flux for any temperature gradient; this would model a perfect thermal insulator.

The Heat Equation

Now let us combine (8.13) and (8.12). Specifically, differentiate both sides of (8.12) with respect to x and note that ρ_0 and u_0 were constant reference quantities that do not depend on x or t , so that

$$\frac{\partial \rho}{\partial x} = c\lambda \frac{\partial u}{\partial x}. \quad (8.14)$$

Combine (8.14) with (8.13) to obtain

$$q(x,t) = -\frac{k}{c\lambda} \frac{\partial \rho}{\partial x}(x,t) \quad (8.15)$$

Equation (8.15) is an example of a **constitutive relation**, a second relation between ρ and q that stems from the specific physics involved in this situation, beyond the conservation assumption that dictates the continuity equation. Combining (8.15) with the continuity equation will yield an equation that governs the behavior of ρ .

Using (8.15) to replace q in the continuity equation (8.8) shows that the function $\rho(x,t)$ must satisfy

$$\frac{\partial \rho}{\partial t} - \frac{k}{c\lambda} \frac{\partial^2 \rho}{\partial x^2} = 0 \quad (8.16)$$

for all x . Equation (8.16) is called the **heat equation**. It is a partial differential equation that the thermal energy density $\rho(x,t)$ obeys. The goal is to solve (8.16) for the function $\rho(x,t)$. Additional information will be needed to arrive at a unique solution, for example, initial data of the form $\rho(x,0)$, and other data if the bar has ends.

However, it will be more convenient and intuitive to work with the temperature $u(x,t)$. Use (8.12) to replace ρ in (8.16) with u (noting that ρ_0 and u_0 are reference constants that do not depend on x or t) to see that u also satisfies (8.16). That is, u satisfies

$$\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = 0 \quad (8.17)$$

where the constant

$$\alpha = \frac{k}{c\lambda} \quad (8.18)$$

is called the **thermal diffusivity of the bar**. The heat equation (8.17) is a direct consequence of the conservation of thermal energy in the bar in combination with Fourier's law, that heat energy flows from hot to cold in proportion to the temperature gradient or slope. The same modeling process works for heat flow in two and three dimensions, although in those cases mass density has the dimension of mass per area or mass per volume, rather than mass per length.

For any specific choice of α , finding a unique solution $u(x,t)$ to the heat equation (8.17) requires additional information, specifically, initial data. If the bar is of infinite length, so that $-\infty < x < \infty$, then specifying $u(x,0) = f(x)$ for some initial temperature $f(x)$ is sufficient to determine a unique solution (if f is anything reasonable). However, our focus from this point on will be a bar of finite length.

Reading Exercise 8.1.4 Verify that the function

$$u(x,t) = \frac{e^{-\frac{x^2}{4(t+1)}}}{\sqrt{t+1}}$$

satisfies the heat equation (8.17) for $-\infty < x < \infty$ and $t > 0$, with thermal diffusivity $\alpha = 1$. Plot $u(x,0)$ for $-10 < x < 10$ (this is the initial temperature of the bar), as well as $u(x,1)$ and $u(x,5)$; overlay the plots if possible. Does this seem like a reasonable temperature profile if heat diffuses from hot to cold in a bar of indefinite length?

Bars of Finite Length and Boundary Conditions

Let us now assume that the bar is of finite length and defined by $0 \leq x \leq L$. As in the case of a bar of infinite length, in order to find a unique solution $u(x,t)$ to the heat equation (8.17) an initial temperature condition is needed, typically in the form

$$u(x,0) = f(x) \quad (8.19)$$

for some function $f(x)$ defined for $0 \leq x \leq L$. But we must also specify what is happening at the ends of the bar, $x = 0$ and $x = L$. One of the most common choices is **Dirichlet boundary conditions**, in which the temperature of the endpoints are specified functions of time. This means that

$$\begin{aligned} u(0,t) &= u_0(t) \\ u(L,t) &= u_L(t) \end{aligned} \quad (8.20)$$

where the functions $u_0(t)$ and $u_L(t)$ are given. A very common special case is when $u_0(t) = u_L(t) = 0$, so-called **homogeneous Dirichlet boundary conditions** in which the bar ends are kept at 0 degrees at all times.

It is a fact that the heat equation (8.17) (with a specified diffusivity α), along with an initial condition (8.19) and boundary conditions (8.20) has a unique solution $u(x,t)$ for $0 \leq x \leq L$ and $t \geq 0$. You can prove this in Exercise 8.3.9. The central question of interest for now is “How do we find this solution?” This will be addressed in the remainder of this section and in the next section.

Reading Exercise 8.1.5 Consider a bar of length $L = 1$, so $0 < x < 1$. The initial temperature data is $u(x,0) = \sin(\pi x)$, and the boundary conditions are the homogeneous Dirichlet conditions $u(0,t) = 0$ and $u(1,t) = 0$ for $t > 0$. Verify that the function $u(x,t) = e^{-\pi^2 t} \sin(\pi x)$ satisfies the heat equation (8.17) for $0 < x < 1$ and $t > 0$, with thermal diffusivity $\alpha = 1$ and with the given boundary and initial data. Plot the solution as a function of x on the range $0 < x < 1$ at times $t = 0, 0.5$, and $t = 2$. Does this seem like reasonable behavior for the temperature of a bar with the endpoints always kept at 0 degrees?

An alternative to Dirichlet boundary conditions are **Neumann boundary conditions** in which the actual thermal energy flux at the ends of the bar is specified. Recall that $q(x,t)$ quantifies the flow of heat energy through the bar in the direction of increasing x . As a result, $q(0,t)$ is the rate at which heat energy is entering the bar at the left end, $x = 0$. The quantity $q(L,t)$ is the rate at which heat energy exits the right end of the bar, at $x = L$. If the rate at which heat energy at the ends of the bar are specified as functions of time, on the left as $g_0(t)$ and on the right as $g_L(t)$, then the boundary conditions are $q(0,t) = g_0(t)$ and $-q(L,t) = g_L(t)$ (the minus sign on q since g_L is the rate heat energy enters, not exits). By using (8.13) these conditions can be put in terms of the temperature u as

$$\begin{aligned} -k \frac{\partial u}{\partial x}(0,t) &= g_0(t) \\ k \frac{\partial u}{\partial x}(L,t) &= g_L(t) \end{aligned} \quad (8.21)$$

where k is the thermal conductivity and $g_0(t)$ and $g_L(t)$ specify the input heat energy flux at the ends of the bar. A very common special case is that in which the ends of the bar do not allow heat energy to enter or exit, so $g_0(t) = g_L(t) = 0$ for all $t > 0$. These are known as **insulating boundary**

conditions and can be expressed as

$$\begin{aligned}\frac{\partial u}{\partial x}(0, t) &= 0 \\ \frac{\partial u}{\partial x}(L, t) &= 0,\end{aligned}\tag{8.22}$$

after dividing out the conductivity k .

As in the case of Dirichlet boundary conditions, the heat equation (8.17), along with an initial condition (8.19) and boundary conditions (8.20) has a unique solution $u(x, t)$ for $0 \leq x \leq L$ and $t \geq 0$. The proof of this is also addressed in Exercise 8.3.9. In the next section we'll consider methods for finding an analytical solution to this problem.

Reading Exercise 8.1.6 Consider a bar of length $L = 1$, so $0 < x < 1$. The initial temperature data is $u(x, 0) = 1 + \cos(\pi x)$, and the boundary is insulated for $t > 0$, so (8.22) holds. Verify that the function $u(x, t) = 1 + e^{-\pi^2 t} \cos(\pi x)$ satisfies the heat equation (8.17) for $0 < x < 1$ and $t > 0$, with thermal diffusivity $\alpha = 1$ with the given boundary and initial data. Plot the solution as a function of x on the range $0 < x < 1$ at times $t = 0, 0.5$, and $t = 2$. Does this seem like reasonable behavior for the temperature of a bar in which no heat energy can enter or exit?

8.1.5 Some Solutions to the Heat Equation: Separation of Variables and Linearity

As already noted, the heat equation (8.17), when paired with an initial condition (8.19) and boundary conditions of the form (8.20) or (8.21), possesses a unique solution. Many other types of physically meaningful and important boundary conditions are common as well, and the heat equation is solvable with these conditions. But we are not yet in a position to write out analytical solutions to the heat equation with desired initial or boundary conditions. That task will have to wait until we have considered *Fourier series* in the next section. But we are in a position to motivate why Fourier series are important, and why they provide the key to solving the heat equation.

Reading Exercise 8.1.7 Verify that the function $u(x, t) = e^{-\alpha\lambda^2 t} \sin(\lambda x)$ satisfies the heat equation (8.17) for any real number λ . Do the same for the function $v(x, t) = e^{-\alpha\lambda^2 t} \cos(\lambda x)$.

Based on Reading Exercise 8.1.7 we have a huge supply of solutions to the heat equation, at least one solution for each real number λ . In a later section we will use this arsenal of solutions to construct solutions to the heat equation with any desired boundary and initial data. But how in the world would anyone have arrived at the solutions in Reading Exercise 8.1.7 in the first place?

Separation of Variables

To obtain these solutions, a philosophy similar to that employed in the method of undetermined coefficients from Section 4.3.2 is useful: when trying to find solutions to a differential equation—whether an ODE or a PDE—make a structured but flexible guess. Then substitute that guess into the DE of interest and adjust to make it work. In the case of the heat equation we will attempt to find solutions that are of the form

$$u(x, t) = T(t)X(x)\tag{8.23}$$

for some functions $T(t)$ and $X(x)$, which we must now determine. Solutions that split into a product of a function of t and a function of x as on the right in (8.23) are said to be **separable**.

Inserting $u(x, t)$ from (8.23) into the heat equation (8.17) and making use of $\frac{\partial u}{\partial t} = T'(t)X(x)$ and $\frac{\partial^2 u}{\partial x^2} = T(t)X''(x)$ yields the requirement that

$$T'(t)X(x) - \alpha T(t)X''(x) = 0$$

for all choices of x and t . A bit of rearrangement shows that this is equivalent to

$$\frac{T'(t)}{\alpha T(t)} = \frac{X''(x)}{X(x)}. \quad (8.24)$$

Here is a curious observation: the left side of (8.24) is a function of t only, while the right side is a function of x , yet t and x are completely independent variables. This can only happen if both sides of (8.24) are constant, and they must be the same constant. For example, if $x = 0$ then (8.24) forces $T'(t)/T(t) = X''(0)/X(0)$ for all t , which makes it clear that $T'(t)/T(t)$ must be constant. Similarly taking $t = 0$ requires that $X''(x)/X(x) = T'(0)/T(0)$ is also constant, clearly the same constant as $T'(t)/T(t)$. Call this constant γ to conclude that

$$\begin{aligned} \frac{T'(t)}{\alpha T(t)} &= \gamma \\ \frac{X''(x)}{X(x)} &= \gamma. \end{aligned} \quad (8.25)$$

The equations for T and X in (8.25) can be written as

$$\begin{aligned} T'(t) - \gamma\alpha T(t) &= 0 \\ X''(x) - \gamma X(x) &= 0. \end{aligned} \quad (8.26)$$

We have deduced in (8.26) that T must satisfy a simple first order ODE, and a general solution to this ODE is $T(t) = Ce^{\gamma\alpha t}$ for some constant C . At this point let us make an observation that will simplify matters. First, solutions to the heat equation should not grow exponentially in time, and so it makes sense to assume that γ in (8.26) cannot be positive (recall that α is positive). To emphasize this fact, let us replace γ by $-\gamma$ in (8.26) with the provision that $\gamma \geq 0$, so that $T'(t) = -\gamma\alpha T(t)$ (recall Modeling Tip 1.1.1). This ODE has solutions of the form

$$T(t) = Ce^{-\gamma\alpha t} \quad (8.27)$$

that do not grow exponentially, but rather decay (or are constant).

With the switch from γ to $-\gamma$ the ODE for $X(x)$ then becomes

$$X''(x) + \gamma X(x) = 0 \quad (8.28)$$

which is the equation of the undamped harmonic oscillator from Section 4.1, at least when $\gamma > 0$ (for $\gamma = 0$ see Reading Exercise 8.1.8). In this case the techniques of Chapter 4 provide a general solution for (8.28) of the form

$$X(x) = c_1 \sin(x\sqrt{\gamma}) + c_2 \cos(x\sqrt{\gamma})$$

for arbitrary constants c_1 and c_2 . In order to avoid writing $\sqrt{\gamma}$ many, many times in the next sections, let us define a constant $\lambda = \sqrt{\gamma}$ (or $\gamma = \lambda^2$), so that the general solutions for $T(t)$ and $X(x)$ can be written as

$$\begin{aligned} T(t) &= Ce^{-\alpha\lambda^2 t} \\ X(x) &= c_1 \sin(\lambda x) + c_2 \cos(\lambda x) \end{aligned} \quad (8.29)$$

for arbitrary constants C, c_1 , and c_2 .

Returning to the starting assumption (8.23) and using $T(t)$ and $X(x)$ as in (8.29) yields a family of solutions to the heat equation that includes those introduced in Reading Exercise 8.1.7,

$$u(x, t) = C_1 e^{-\alpha\lambda^2 t} \sin(\lambda x) + C_2 e^{-\alpha\lambda^2 t} \cos(\lambda x) \quad (8.30)$$

where we have set $C_1 = Cc_1$ and $C_2 = Cc_2$ as arbitrary constants. For any choice of the constants C_1, C_2 , and λ , the function $u(x, t)$ defined in (8.30) satisfies the heat equation (8.17). This yields a large reservoir of solutions to the heat equation that we will make use of in the following sections. The process that led from (8.23) to (8.29) and (8.30) is called **separation of variables**, the same term that was introduced for the solution procedure used for separable ODEs in Section 2.2.

Reading Exercise 8.1.8 Show that in the special case $\gamma = 0$ (8.27) and (8.28) yield solutions $u(x, t) = C_1x + C_2$ to the heat equation, and verify that these do in fact satisfy (8.17). These are **steady-state** solutions to the heat equation, that is, solutions that do not depend on time. We'll encounter them later.

Linearity

Not only does (8.30) yield an infinite supply of solutions to the heat equation, we can also take linear combinations of these solutions to produce more solutions. Specifically, let $u_1(x, t)$ and $u_2(x, t)$ be solutions to (8.17) and let a and b be any scalars. Let

$$u(x, t) = au_1(x, t) + bu_2(x, t).$$

Then u also satisfies the heat equation, for

$$\begin{aligned} \frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} &= \frac{\partial(au_1 + bu_2)}{\partial t} - \alpha \frac{\partial^2(au_1 + bu_2)}{\partial x^2} \\ &= a \frac{\partial u_1}{\partial t} + b \frac{\partial u_2}{\partial t} - \alpha a \frac{\partial^2 u_1}{\partial x^2} - \alpha b \frac{\partial^2 u_2}{\partial x^2} \\ &= a \left(\frac{\partial u_1}{\partial t} - \alpha \frac{\partial^2 u_1}{\partial x^2} \right) + b \left(\frac{\partial u_2}{\partial t} - \alpha \frac{\partial^2 u_2}{\partial x^2} \right) \\ &= a \cdot + b \cdot 0 \\ &= 0 \end{aligned} \tag{8.31}$$

since u_1 and u_2 are themselves solutions to the heat equation. Thus the superposition $u(x, t)$ also provides a solution to the heat equation. This computation extends to any finite number of summands, that is, the function $u = c_1u_1 + c_2u_2 + \dots + c_nu_n$ satisfies the heat equation for any scalars c_k if each u_k satisfies the heat equation.

Boundary Conditions

Let us now consider how we can implement boundary conditions, focusing for the moment on the homogeneous Dirichlet case (8.20) in which $u_0(t) = 0$ and $u_L(t) = 0$. Consider a solution $u(x, t)$ of the form (8.30). Requiring $u(0, t) = 0$ for all t leads to the requirement that

$$u(0, t) = C_1e^{-\alpha\lambda^2 t} \sin(\lambda 0) + C_2e^{-\alpha\lambda^2 t} \cos(\lambda 0) = C_2e^{-\alpha\lambda^2 t} = 0$$

for all $t > 0$. Since $e^{-\alpha\lambda^2 t}$ is never zero it follows that $C_2 = 0$ in (8.30) if a homogeneous Dirichlet boundary condition at $x = 0$ holds.

We are then left to consider solutions of the form

$$u(x, t) = C_1e^{-\alpha\lambda^2 t} \sin(\lambda x).$$

To obtain the homogeneous Dirichlet boundary condition $u(L, t) = 0$ for all $t > 0$ we need

$$u(L, t) = C_1e^{-\alpha\lambda^2 t} \sin(\lambda L) = 0.$$

Again, the exponential $e^{-\alpha\lambda^2 t}$ is never zero, so $C_1 \sin(\lambda L) = 0$. One way to obtain this is to take $C_1 = 0$, but that leads to solutions $u(x, t) = 0$ that are identically zero and of no value in what is to come. Instead, let us require that

$$\sin(\lambda L) = 0. \quad (8.32)$$

The equation $\sin(z) = 0$ has solutions $z = \pi j$ where j is an integer, so in order to satisfy (8.32) we must take $\lambda L = \pi j$ for an integer j , or

$$\lambda = \pi j/L. \quad (8.33)$$

The case $j = 0$ yields a useless solution that is identically zero, but the nonzero values of j equation (8.33) in conjunction with (8.30) yield solutions to the heat equation that also satisfy the boundary conditions $u(0, t) = u(L, t) = 0$. These solutions are of the form

$$u(x, t) = C_1 e^{-\alpha\pi^2 j^2 t/L^2} \sin(\pi j x/L) \quad (8.34)$$

for any integer j . Moreover, the linearity of the heat equation shows that any linear combination

$$u(x, t) = \sum_{j=1}^n c_j e^{-\alpha\pi^2 j^2 t/L^2} \sin(\pi j x/L) \quad (8.35)$$

will also satisfy the heat equation, as well as the boundary conditions $u(0, t) = u(L, t) = 0$.

Reading Exercise 8.1.9 Show that if $u(x, t)$ as in (8.30) satisfies an insulating boundary condition $\frac{\partial u}{\partial x}(0, t) = 0$ for all $t > 0$ then u is either constant or of the form

$$u(x, t) = C_2 e^{-\alpha\lambda^2 t} \cos(\lambda x)$$

(that is, C_1 in (8.30) must be zero). Then show that if u is not constant and also satisfies an insulating boundary condition $\frac{\partial u}{\partial x}(L, t) = 0$ for $t > 0$ then $\lambda = \pi j/L$ for some integer j .

We will use strategically chosen linear combinations of solutions of the form (8.34) or those in Reading Exercise 8.1.9 to solve the heat equation with given initial and boundary conditions in a later section.

8.1.6 Exercises

Exercise 8.1.1 Suppose that some stuff flows through a conduit of indefinite length and is conserved at each point and each time. Also suppose that the density $\rho(x, t)$ of stuff is independent of time t . Use the continuity equation (8.8) to argue that the flux $q(x, t)$ must be the same at all positions x (but may depend on time t). Also argue the converse, that if the flux is independent of position then the density cannot depend on time t .

Exercise 8.1.2 In each case below find a solution $u(x, t)$ to the heat equation (8.17) with diffusivity $\alpha = 1$ on the interval $0 < x < 1$ that has homogeneous Dirichlet boundary conditions $u(0, t) = 0$ and $u(1, t) = 0$ for $t > 0$ with the given initial condition for $u(x, 0)$. Hint: use (8.34) and if necessary, linearity in the form of (8.35). Plot each solution as a function of x for $0 \leq x \leq 1$, at the times $t = 0, 0.01, 0.05$, and 0.5 . How does the solution behave as t approaches infinity?

- (a) $u(x, 0) = 3 \sin(\pi x)$.
- (b) $u(x, 0) = 3 \sin(\pi x) + 5 \sin(6\pi x)$.
- (c) $u(x, 0) = \sin(2\pi x) - 3 \sin(8\pi x)$.
- (d) $u(x, 0) = 4 \sin(4\pi x) + 2 \sin(14\pi x)$.

Exercise 8.1.3 In each case below, find a solution $u(x,t)$ to the heat equation (8.17) with diffusivity $\alpha = 1$ on the interval $0 < x < 1$ that has insulating boundary conditions $\frac{\partial u}{\partial x}(0,t) = 0$ and $\frac{\partial u}{\partial x}(1,t) = 0$ for $t > 0$ with the given initial condition for $u(x,0)$. Hint: use the result of Reading Exercise 8.1.9 and, if necessary, linearity. Note that the solutions of Reading Exercise 8.1.9 are valid when $\lambda = 0$. Plot each solution as a function of x for $0 \leq x \leq 1$, at the times $t = 0, 0.01, 0.05$, and 0.5 . How does the solution behave as t approaches infinity?

- (a) $u(x,0) = 3 \cos(\pi x)$.
- (b) $u(x,0) = 4 + 3 \cos(\pi x)$.
- (c) $u(x,0) = 2 + 5 \cos(6\pi x)$.
- (d) $u(x,0) = 4 \cos(4\pi x) + 2 \cos(14\pi x)$.

Exercise 8.1.4 Suppose the temperature $u(x,t)$ of some bar satisfies the heat equation (8.17) with diffusivity $\alpha = 1$ on the interval $0 < x < L$ with insulating boundary conditions $\frac{\partial u}{\partial x}(0,t) = 0$ and $\frac{\partial u}{\partial x}(L,t) = 0$ for $t > 0$. In this problem we show that the total amount of thermal energy in the bar is conserved over time.

- (a) Show that the value of the integral

$$\int_0^L u(x,t) dx$$

does not change over time. Hint: Differentiate the integral above with respect to t , assume you can pass the t derivative under the integral, and use the fact that u satisfies the heat equation to obtain

$$\int_0^L \frac{\partial u}{\partial t} dx = \alpha \int_0^L \frac{\partial^2 u}{\partial x^2} dx.$$

Evaluate the integral on the right with the fundamental theorem of calculus and use the insulating boundary conditions to show this integral is zero. See Figure 8.15 in Example 8.10 in Section 8.3 for an illustration.

- (b) Use (8.12) to argue that the integral

$$\int_0^L \rho(x,t) dx$$

does not depend on time, where ρ is the thermal energy density in the bar. What physical quantity does this integral represent, and why does it make sense that it should be constant in the case of insulating boundary conditions?

Exercise 8.1.5 Consider a bar spanning the x axis from $x = 0$ to $x = L$. The bar sits in an ambient environment with temperature A degrees and gains or loses thermal energy to the environment only at the ends at $x = 0$ and $x = L$. Let $u(x,t)$ denote the temperature of the bar and $q(x,t) = -k \frac{\partial u}{\partial x}$ the heat flux, with k as the thermal conductivity of the bar.

- (a) Suppose the bar loses heat at $x = 0$ in proportion to the difference between the temperature $u(0,t)$ and the ambient environment temperature. Argue that the appropriate model for

this is $q(0,t) = -\mu(u(0,t) - A)$ where μ is nonnegative constant. Hint: do a thought experiment in which $u(0,t)$ is much larger than A ; what direction should heat energy be flowing?

- (b) Express the boundary condition at $x = 0$ from part (a) using u instead of q .
- (c) Formulate an equivalent boundary condition at $x = L$.

Boundary conditions of the type in part (b) or (c) are called **Robin boundary conditions**.

Exercise 8.1.6 In many situations the solution $u(x,t)$ to the heat equation on an interval $0 < x < L$ approaches a steady-state value, that is, as t approaches infinity the solution asymptotically approaches a function $U(x)$ that is independent of time t . The function $U(x)$ is itself a solution to the heat equation, however.

- (a) Show that $U(x)$ satisfies $d^2U/dx^2 = 0$ on $0 < x < L$, a steady-state version of the heat equation. Conclude that $U(x) = c_1x + c_2$ for some constants c_1 and c_2 .
- (b) Suppose one end of a bar with left end at $x = 0$ is held at a temperature of 20 degrees Celsius. The other end at $x = 5$ is held at 80 degrees Celsius. The temperature $U(x)$ of the bar is at steady-state. Find a formula for $U(x)$ in this case.
- (c) Suppose we have Dirichlet data $U(0) = u_0$ and $U(L) = u_L$ where u_0 and u_L are constants, the temperature of $U(x)$ at the endpoints of the bar, independent of time. Show that there is a unique function $U(x)$ of the form in part (a) that satisfies these Dirichlet boundary conditions. Find a formula for $U(x)$ in terms of x, u_0, u_L , and L .
- (d) Consider a solution $U(x)$ to the steady-state heat equation as in part (a) but with steady-state Neumann boundary conditions (8.21), which now take the form $-kU'(0) = g_0$ and $kU'(L) = g_1$. Here g_0 and g_1 are constants that specify the input heat energy flux at each endpoint of the bar. Argue that no such steady-state solution exists unless $g_1 = -g_0$. Also argue that if $g_1 = -g_0$ does hold then solutions do exist, and in fact there are infinitely many solutions.

Why does the condition $g_1 = -g_0$ make sense physically? Hint: if we pump 2 joules of energy per second into one end of the part and withdraw 1 joule per second from the other end, why would there be no steady-state solution?

Exercise 8.1.7 Suppose stuff flows through a one-dimensional conduit with flux $q(x,t)$ and stuff density $\rho(x,t)$. Let Ω denote a control volume in the conduit spanning the interval $x = x_0$ to $x = x_0 + \Delta x$ as illustrated in Figure 8.3. Suppose further that stuff is not conserved in the conduit, but rather is being created or destroyed. Let $r(x,t)$ denote the rate at which stuff is being created or destroyed at position x and time t in the conduit on a stuff per unit length of the conduit per unit time, where of course $r > 0$ indicates creation, $r < 0$ indicates destruction. Thus the net rate at which stuff is being created or destroyed in Ω at time t is given by the integral

$$\text{rate of stuff creation in } \Omega = \int_{x_0}^{x_0 + \Delta x} r(x,t) dx, \quad (8.36)$$

with the dimension of stuff per time.

- (a) Argue that the net rate at which the amount of stuff in Ω is changing at time t is the net rate at which stuff enters Ω plus the net rate of stuff creation in Ω (the former quantified by (8.3), the latter by (8.36).)

- (b) Based on part (a), argue that (8.4) should be modified to read

$$\int_{x_0}^{x_0+\Delta x} \frac{\partial \rho}{\partial t}(x, t) dx = q(x_0, t) - q(x_0 + \Delta x, t) + \int_{x_0}^{x_0+\Delta x} r(x, t) dx. \quad (8.37)$$

- (c) Let Δx approach 0 in (8.37) and mimic the reasoning that led to (8.8) to show that

$$\frac{\partial \rho}{\partial t} + \frac{\partial q}{\partial x} = r(x, t). \quad (8.38)$$

This version of the continuity equation is appropriate to the situation in which stuff is not conserved. Of course $r = 0$ yields the usual continuity equation.

8.2 Fourier Series

Based on the results of Section 8.1, we now know how to solve the heat equation (8.17) on an interval $0 \leq x \leq L$ in certain circumstances. For example (8.35) provides a solution in the case of homogeneous Dirichlet boundary conditions $u(0, t) = u(L, t) = 0$ and an initial temperature that is a finite linear combination of sine functions of the form $\sin(j\pi x/L)$ where j is an integer. This follows from the linearity of the heat equation as shown in (8.31) and the fact that we found solutions of the form (8.34) by using separation of variables. A similar observation holds for insulating boundary conditions $\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(L, t) = 0$ and initial data that consists of linear combinations of functions of the form $\cos(j\pi x/L)$. In this case a superposition of solutions of the form in Reading Exercise 8.1.9 yields a solution. Such observations were used in Exercises 8.1.2 and 8.1.3.

But what if the initial data is not of these convenient forms, a finite sum of suitable sines or cosines? One of the great discoveries of 19th century mathematics and science was this:

Every reasonable function defined on an interval $0 \leq x \leq L$ can be expressed as a sum of sines or cosines, if we are willing to consider infinite sums and interpret these infinite sums carefully.

The situation is not unlike that of Taylor polynomials and series from elementary calculus, which allow us to approximate functions like e^x , $\sin(x)$, and so on, as polynomials. When we let the degree of the polynomial increase without bound we obtain a Taylor series that, in many circumstances, converges to the function of interest.

8.2.1 An Example

Let's begin with a concrete illustration. Let $f(x) = 30x^3 - 41x^2 + 17x - 2$; there is nothing magic about this function, other than that it is mildly interesting, as it has two local extrema and several roots. A graph of $f(t)$ is shown in left panel of Figure 8.5, on the interval $0 \leq x \leq 1$. The claim is that $f(x)$ can be approximated well on this interval by using a function $s_n(x)$ that is a sum of the form

$$s_n(x) = a_0 \cos(0\pi x) + a_1 \cos(1\pi x) + a_2 \cos(2\pi x) + \cdots + a_n \cos(n\pi x), \quad (8.39)$$

if the coefficients are a_k are carefully chosen and n is large enough. Note that the first term on the right in (8.39) is really just a_0 since $\cos(0\pi x) = \cos(0) = 1$, but we write $\cos(0\pi x)$ just to emphasize the pattern.

Consider the case in which $n = 0$ in (8.39), so the approximation to $f(x)$ is $s_0(x) = a_0$, a constant. The right panel in Figure 8.5 shows the approximation when $a_0 = 1/3$. As will be shown, this constant is the best choice, in a certain quantifiable sense, though clearly the approximation

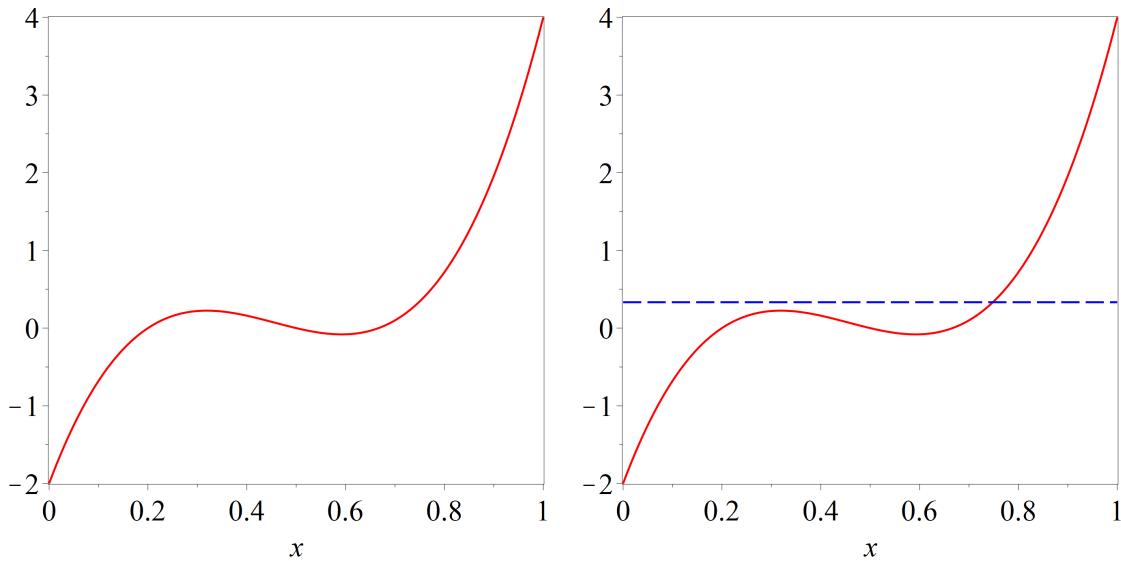


Figure 8.5: Left panel: function $f(x) = 30x^3 - 41x^2 + 17x - 2$. Right panel: function $f(x)$ (solid red) and constant approximation $s_0(x) = 1/3$ (dashed blue).

is not very good. However, consider $n = 1$, as in the left panel of Figure 8.6. The function that approximates f here is

$$s_1(x) = 1/3 - 1.1195 \cos(\pi x)$$

where we use $a_1 = -1.1195$. This is a slight improvement on $s_0(x)$, but not much. With $n = 2$ the approximation becomes

$$s_2(x) = 1/3 - 1.1195 \cos(\pi x) + 0.4053 \cos(2\pi x),$$

obtained with $a_2 = 0.4053$. This is shown in the right panel of Figure 8.6 and still doesn't seem to do a very good job.

However, the approximation

$$\begin{aligned} s_5(x) = & 1/3 - 1.1195 \cos(\pi x) + 0.4053 \cos(2\pi x) - 0.8544 \cos(3\pi x) \\ & + 0.1013 \cos(4\pi x) - 0.3286 \cos(5\pi x) \end{aligned}$$

(using the previous a_0, a_1, a_2 values along with $a_3 = -0.8544, a_4 = 0.1013$, and $a_5 = -0.3286$) shown in the left panel of Figure 8.7 is clearly a big improvement, and $s_{10}(x)$ (whose expansion is not shown, but the final term is $0.0162 \cos(10\pi x)$) is better still.

Two questions concerning this approximation process naturally arise:

1. Where do the coefficients in front of the $\cos(k\pi x)$ terms of $s_n(x)$ come from?
2. If we continue this process with appropriately chosen a_k , will $s_n(x)$ converge to $f(x)$? In what precise sense?

8.2.2 Approximating Functions

Based on Figures 8.5 to 8.7, it appears that if n is large and each coefficient a_k in front of the corresponding term $\cos(k\pi x/L)$ is chosen properly then $s_n(x)$ is a good approximation to $f(x)$. But what precisely does this mean? We need a method to quantify how well one function approximates another, a way to measure the distance between two functions.

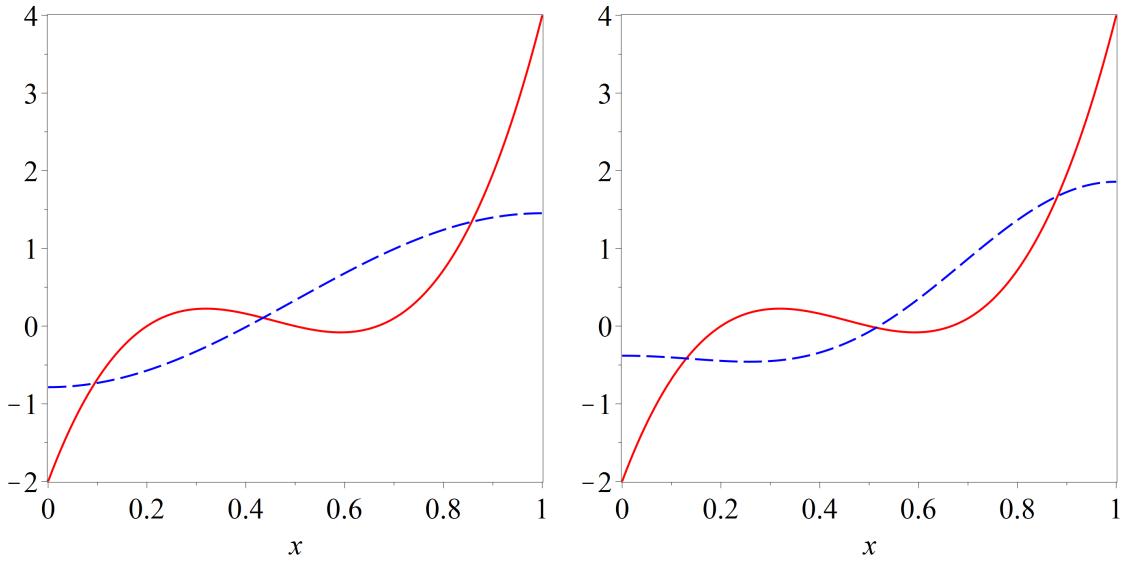


Figure 8.6: Left panel: function $f(x) = 30x^3 - 41x^2 + 17x - 2$ (solid red) and cosine approximation $s_1(x)$ (dashed blue). Right panel: function $f(x) = 30x^3 - 41x^2 + 17x - 2$ (solid red) and cosine approximation $s_2(x)$ (dashed blue).

Let's look at one mathematically elegant way to quantify how close two functions are to each other. This approach meshes particularly well with the idea of approximating functions with cosines and sines. First, recall Definition 5.2.2 for a piecewise continuous function. Also, we say a function f is **bounded** on its domain (in this case, $a \leq x \leq b$) if there is some constant M such that $|f(x)| \leq M$ for all x in the domain of f . With this terminology let us make the following definition.

Definition 8.2.1 Let $f(x)$ be a bounded piecewise continuous real-valued function on an interval $a \leq x \leq b$. We define the **L^2 norm** of f as

$$\|f\|_2 = \left(\int_a^b f^2(x) dx \right)^{1/2}.$$

The assumption that f is bounded and piecewise continuous ensures that the integral on the right in the definition for $\|f\|_2$ exists and is a real number (note f^2 is also bounded, by M^2 if $|f| \leq M$).

The L^2 norm of f provides a way to quantify the size of f . It's not hard to show that $\|cf\|_2 = |c|\|f\|_2$ for any real scalar c . Also, $\|f\|_2 = 0$ if and only if f is the zero function on the interval $[a, b]$; see Exercise 8.2.5.

■ **Example 8.1** Let $f(x) = x^2$ be defined on the interval $0 \leq x \leq 2$. Then

$$\|f\|_2 = \left(\int_0^2 (x^2)^2 dx \right)^{1/2} = \left(\int_0^2 x^4 dx \right)^{1/2} = \sqrt{32/5}.$$

■

The distance between two functions f and g is quantified by using $\|f - g\|_2$ (much as the distance between two real numbers p and q is quantified by using the absolute value $|p - q|$).

■ **Example 8.2** Let $f(x) = 30x^3 - 41x^2 + 17x - 2$ be defined on the interval $0 \leq x \leq 1$. This is the

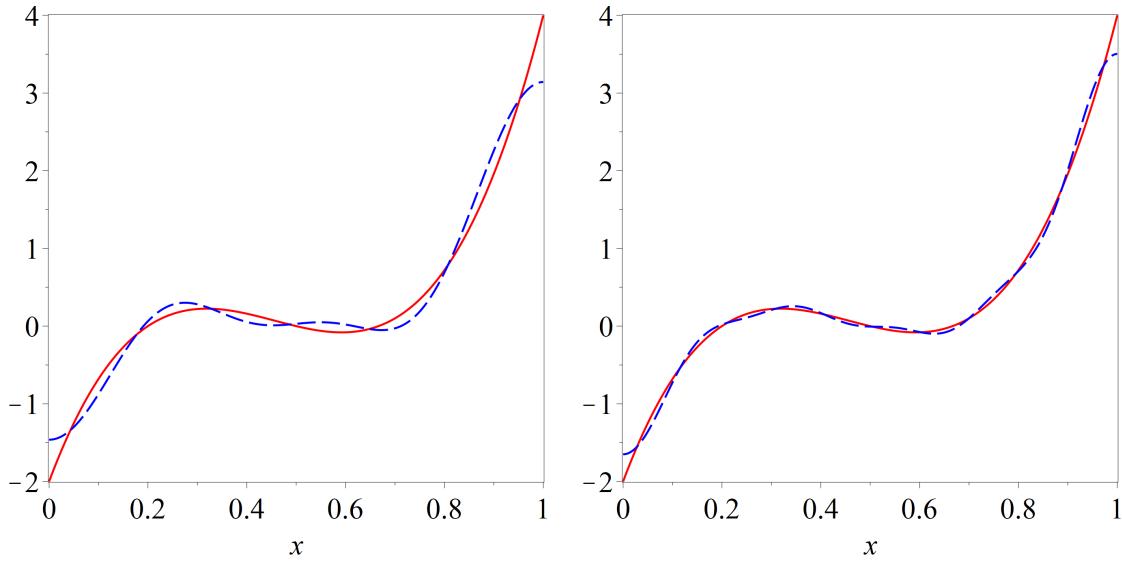


Figure 8.7: Left panel: function $f(x) = 30x^3 - 41x^2 + 17x - 2$ (solid red) and cosine approximation $s_5(x)$ (dashed blue). Right panel: function $f(x) = 30x^3 - 41x^2 + 17x - 2$ (solid red) and cosine approximation $s_{10}(x)$ (dashed blue).

function from Section 8.2.1. Let $s_0(x) = 1/3$ as in that section. Then

$$\|f - s_0\|_2 = \left(\int_0^1 (30x^3 - 41x^2 + 17x - 2 - 1/3)^2 dx \right)^{1/2} = \sqrt{731/630} \approx 1.0772.$$

We can also compute (use $s_1(x) = 1/3 - 1.1195 \cos(\pi x)$ from that section) that

$$\|f - s_1\|_2 = \left(\int_0^1 (30x^3 - 41x^2 + 17x - 2 - (1/3 - 1.1195 \cos(\pi x)))^2 dx \right)^{1/2} \approx 0.7306.$$

Similar computations show that $\|f - s_2\|_2 \approx 0.6720$, $\|f - s_5\|_2 \approx 0.1657$, and $\|f - s_{10}\|_2 \approx 0.0776$.

In Example 8.2 it appears that as n increases s_n provides a better and better approximation to f in the L^2 norm, at least for f and the s_n from Section 8.2.1. But this still doesn't answer how to obtain the mysterious coefficients in front of the cosines.

8.2.3 The Fourier Cosine Expansion

Given a function $f(x)$ defined on an interval $0 \leq x \leq L$, let us consider how to compute the coefficients a_k in an expansion of the form

$$s_n(x) = a_0 + a_1 \cos(\pi x/L) + \cdots + a_n \cos(n\pi x/L) \quad (8.40)$$

so that s_n is a good approximation to f , and improves as the number of terms n increases, in the sense of the L^2 distance. The idea is simple: Choose the numbers a_0, a_1, \dots, a_n so that the quantity $\|f - s_n\|_2$ is minimized. It's worth noting at this point that minimizing a nonnegative quantity like $\|f - s_n\|_2$ is entirely equivalent to minimizing its square, so we will proceed by minimizing $\|f - s_n\|_2^2$. This makes the computation a bit easier, as there is then no square root involved as in Definition 8.2.1.

The Best Choice for a_0

■ **Example 8.3** Let's begin with a simple example. Consider a function $f(x)$ defined on $0 \leq x \leq L$. We will construct an approximation to f in the case that $n = 0$ in (8.40). With $s_0(x) = a_0$, we seek that value of a_0 so that $\|f - s_0\|_2$ is minimized, which as remarked, is equivalent to minimizing $\|f - s_0\|_2^2$. With a_0 unspecified compute

$$\begin{aligned}\|f - s_0\|_2^2 &= \int_0^L (f(x) - a_0)^2 dx \\ &= \int_0^L (f^2(x) - 2f(x)a_0 + a_0^2) dx \\ &= \int_0^L f^2(x) dx - 2a_0 \int_0^L f(x) dx + a_0^2 \int_0^L 1 dx \\ &= \int_0^L f^2(x) dx - 2a_0 \int_0^L f(x) dx + La_0^2.\end{aligned}\tag{8.41}$$

Think of the right side of (8.41) as a function ϕ of a_0 , of the form $\phi(a_0) = P - Qa_0 + La_0^2$ where $P = \int_0^L f^2(x) dx$ and $Q = 2 \int_0^L f(x) dx$. Since $L > 0$ the function ϕ is concave up and has a unique minimum with respect to a_0 , when $\phi'(a_0) = -Q + 2La_0 = 0$. This minimum occurs when $a_0 = Q/(2L)$. In other words, the optimal value for a_0 is

$$a_0 = \frac{1}{L} \int_0^L f(x) dx.$$

For example, with $f(x) = 30x^3 - 41x^2 + 17x - 2$ and $L = 1$ as in Section 8.2.1 it turns out that $a_0 = 1/3$. This is the best constant approximation $s_0(x) = a_0$ to $f(x)$ possible, where *best* means the constant a_0 that minimizes $\|f - a_0\|_2$. ■

This approach works more generally for finding a_0, a_1, \dots, a_n to minimize $\|f - s_n\|_2$. It is aided enormously by the fact detailed in Reading Exercise 8.2.1.

Reading Exercise 8.2.1 Use the trigonometric identity

$$\cos(x)\cos(y) = \frac{\cos(x+y) + \cos(x-y)}{2}$$

to show that

$$\int_0^L \cos(j\pi x/L) \cos(k\pi x/L) dx = 0\tag{8.42}$$

when j and k are positive integers with $j \neq k$. Then use the same identity to show that with $j = k$ on the left in (8.42) the integral yields

$$\int_0^L \cos^2(j\pi x/L) dx = L/2\tag{8.43}$$

if $j \neq 0$.

Let us make the following definition.

■ **Definition 8.2.2 — Orthogonality.** Two real-valued functions $g(x)$ and $h(x)$ defined on an

interval $a \leq x \leq b$ are **orthogonal** if

$$\int_a^b g(x)h(x) dx = 0.$$

This is not unlike the definition of orthogonality for two vectors \mathbf{v} and \mathbf{w} , where orthogonality means that $\mathbf{v} \cdot \mathbf{w} = 0$, or $\sum_{k=1}^n v_k w_k = 0$. With functions instead of vectors, an integral replaces the sum in the dot product and the vector components are replaced by the values of the functions at specific x coordinates. Thus according to Definition 8.2.2 and Reading Exercise 8.2.1, when j and k are distinct positive integers, the functions $\cos(j\pi x/L)$ and $\cos(k\pi x/L)$ are orthogonal on the interval $0 \leq x \leq L$.

The Best Choice for the Cosine Coefficients a_k

Let's consider the problem of minimizing $\|f - s_n\|_2^2$ where s_n is defined as in (8.40). Compute

$$\|f - s_n\|_2^2 = \int_0^L (f(x) - a_0 - a_1 \cos(\pi x/L) - a_2 \cos(2\pi x/L) - \cdots - a_n \cos(n\pi x/L))^2 dx.$$

This looks formidable, but consider what happens when the squared integrand is expanded out. There will be a term $f^2(x)$, a term a_0^2 , terms of the form $a_j^2 \cos^2(j\pi x/L)$ for $j = 1$ to $j = n$, and cross terms of the form $2f(x)a_j \cos(j\pi x/L)$, $2a_0a_j \cos(j\pi x/L)$, and $2a_ja_k \cos(j\pi x/L) \cos(k\pi x/L)$. All of these will be integrated from $x = 0$ to $x = L$. However, from the result of Reading Exercise 8.2.1, in particular (8.42), all terms $2a_ja_k \cos(j\pi x/L) \cos(k\pi x/L)$ and $2a_0a_j \cos(j\pi x/L)$ integrate to zero. That is, these functions are orthogonal on $0 \leq x \leq L$. Also, from (8.43) all terms $a_j^2 \cos^2(j\pi x/L)$ integrate to $La_j^2/2$. The a_0^2 term integrates to La_0^2 .

When all the dust has settled we find

$$\|f - s_n\|_2^2 = \int_0^L f^2(x) dx - 2 \sum_{j=0}^n a_j \int_0^L f(x) \cos(j\pi x/L) + La_0^2 + \frac{L}{2} \sum_{j=1}^n a_j^2. \quad (8.44)$$

The expression on the right in (8.44) is, as a function of the variables a_0, \dots, a_n , a simple quadratic function. It can be minimized by computing its partial derivatives with respect to each variable, setting them to zero (which yields $n+1$ linear equations) and solving for a_0, \dots, a_n .

Let $\phi(a_0, \dots, a_n)$ denote the right side of (8.44). It's easy to compute that

$$\frac{\partial \phi}{\partial a_0} = -2 \int_0^L f(x) dx + 2La_0.$$

A minor miracle occurs: this equation does not involve any of the other a_j coefficients, and the equation $\partial \phi / \partial a_0 = 0$ yields

$$a_0 = \frac{1}{L} \int_0^L f(x) dx. \quad (8.45)$$

This is the same result obtained in Example 8.3. For the other coefficients a similar computation shows that

$$\frac{\partial \phi}{\partial a_k} = -2 \int_0^L f(x) \cos(k\pi x/L) dx + La_k,$$

for any k between 1 and n . Again, all the other a_j coefficients drop out of this derivative, and setting $\partial \phi / \partial a_k = 0$ yields

$$a_k = \frac{2}{L} \int_0^L f(x) \cos(k\pi x) dx \quad (8.46)$$

for $k = 1$ to $k = n$. The values for a_0 and the a_k given by (8.45) and (8.46) yield the unique critical point for the function $\phi(a_0, \dots, a_n) = \|f - s_n\|_2^2$. This computation does not show that this critical point is a minimum (critical points can be maxima or other, for example, saddles) but in this case the critical point can be shown to be a minimum; for a proof see [105].

We have arrived at a rather beautiful and powerful conclusion.

Theorem 8.2.1 Let f be a bounded piecewise continuous function on the interval $0 \leq x \leq L$ and suppose $s_n(x)$ is defined by (8.40). Then the choices (8.45) for a_0 and (8.46) for a_k with $1 \leq k \leq n$ yield a minimum value for $\|f - s_n\|_2$.

Theorem 8.2.1 tells us what choices for the a_k coefficients in (8.40) will give the best approximation to $f(x)$ as measured in the L^2 norm. Can this approximation be made arbitrarily good?

The Convergence of the Cosine Series

It would be ideal if increasing n causes $\|f - s_n\|_2$ to approach zero so that an approximation of any desired accuracy could be obtained with a sufficiently large n . Note that $\|f - s_n\|_2$ can in fact equal zero precisely when $s_n = f$, but $\|f - s_n\|_2$ can never be negative, so zero is the best we can do. Remarkably, if the a_k are chosen according to (8.45) and (8.46) then $\|f - s_n\|_2$ is guaranteed to approach zero as n increases.

Theorem 8.2.2 Let f be a bounded piecewise continuous function on the interval $0 \leq x \leq L$ and suppose $s_n(x)$ is defined by (8.40). Let a_0 be given by (8.45) and a_k given by (8.46). Then

$$\lim_{n \rightarrow \infty} \|f - s_n\|_2 = 0. \quad (8.47)$$

For a proof of this theorem, see [105].

The punchline of Theorem 8.2.2 is that by taking appropriate linear combinations of the functions in the set

$$S = \{1, \cos(\pi x/L), \cos(2\pi x/L), \cos(3\pi x/L), \dots\}$$

any piecewise continuous function f can be approximated to arbitrary accuracy, as measured in the L^2 norm. In this case we say that the set S is **complete** in the space of piecewise continuous functions on the interval $0 \leq x \leq L$.

The conclusion of Theorem 8.2.2 is often expressed less formally by writing

$$f(x) = a_0 + \sum_{j=1}^{\infty} a_j \cos(j\pi x/L), \quad (8.48)$$

although Theorem 8.2.2 and (8.47) are always behind the scenes as the precise interpretation of the infinite sum (8.48). The expansion of (8.48) is called the **Fourier cosine series** for the function f . The type of convergence of s_n to f in (8.47) is called **L^2 convergence** or **mean square convergence**.

■ **Example 8.4** Let $f(x)$ be defined on the interval $0 \leq x \leq 1$ as

$$f(x) = \begin{cases} 1, & 0 \leq x < 1/2 \\ 2, & 1/2 \leq x \leq 1 \end{cases}$$

This function is bounded (note that $|f(x)| \leq 2$) and piecewise continuous, with a single jump discontinuity at $x = 1/2$. The Fourier coefficient a_0 is easily computed as $a_0 = \int_0^1 f(x) dx = 3/2$.

From (8.46) it follows that

$$a_k = 2 \int_0^1 f(x) \cos(k\pi x) dx = \frac{2(-1)^k}{k\pi}$$

(a computer algebra system is helpful) when k is odd and $a_k = 0$ when $k \geq 2$ and k is even. Thus $a_1 = -2/\pi, a_3 = 2/(3\pi), a_5 = -2/(5\pi)$, and so on. This makes it easy to write out $s_n(x)$ to any number of terms. In particular, (8.48) becomes

$$f(x) = \frac{3}{2} - \frac{2}{\pi} \cos(\pi x) + \frac{2}{3\pi} \cos(3\pi x) - \frac{2}{5\pi} \cos(5\pi x) + \dots$$

The left panel of Figure 8.8 shows a graph of $f(x)$ and the function $s_0(x) = 3/2$, while the right panel shows $f(x)$ and $s_1(x) = 3/2 - 2\cos(\pi x)/\pi$. One can compute that $\|f - s_0\|_2 = 1/2$ and $\|f - s_1\|_2 \approx 0.2176$. The left panel of Figure 8.9 shows a graph of $f(x)$ and the function

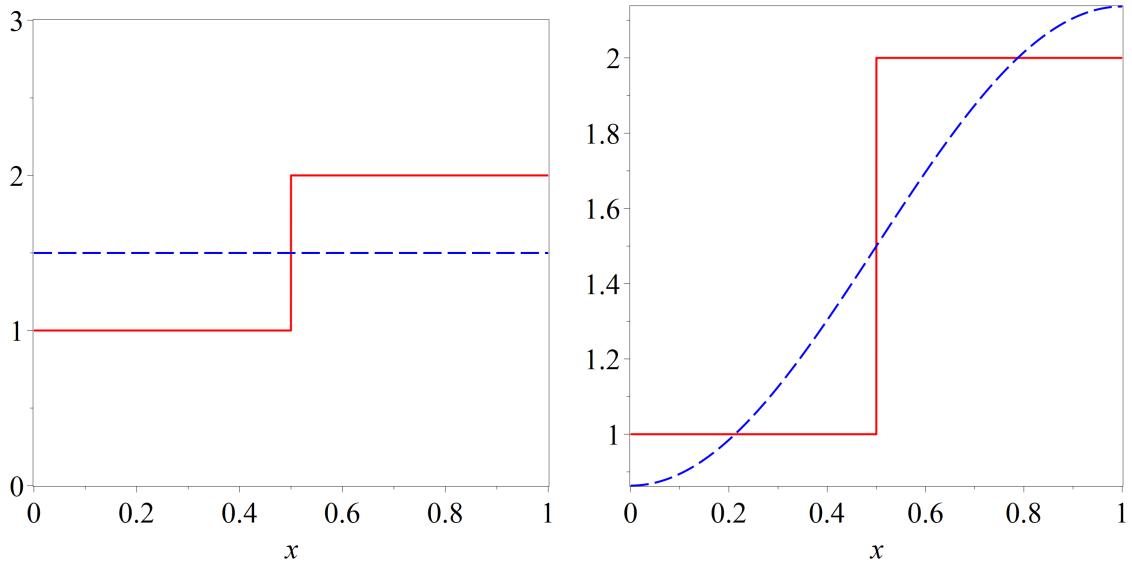


Figure 8.8: Left panel: function $f(x)$ (solid red) and cosine approximation $s_0(x) = 3/2$ (dashed blue). Right panel: function $f(x)$ (solid red) and cosine approximation $s_1(x)$ (dashed blue).

$s_{10}(x)$, while the right panel shows $f(x)$ and $s_{50}(x)$. One can compute that $\|f - s_{10}\|_2 = 0.1005$ and $\|f - s_{50}\|_2 \approx 0.0450$. ■

In Example 8.4 note that for any finite value of n the function $s_n(x)$ is a finite sum of cosines and so is continuous, but s_n has the unhappy task of trying to approximate a function $f(x)$ that is not continuous. This causes obvious difficulty at the discontinuity, an issue we'll look at more carefully later in this section. Nonetheless, it is the case that $\|f - s_n\|_2 \rightarrow 0$ as $n \rightarrow \infty$.

A Minor Variation on Cosine Expansions

Note that the expansion (8.40) for s_n requires that a_0 be computed using (8.45) while if $k \geq 1$ then a_k is computed with (8.46). The formula (8.46) makes sense when $k = 0$, but yields a value for a_0 that is two times too large. It is therefore quite common to actually use (8.46) for all $k \geq 0$ and then compensate by taking

$$s_n(x) = a_0/2 + a_1 \cos(\pi x/L) + \dots + a_n \cos(n\pi x/L) \quad (8.49)$$

as the Fourier cosine approximation to $f(x)$.

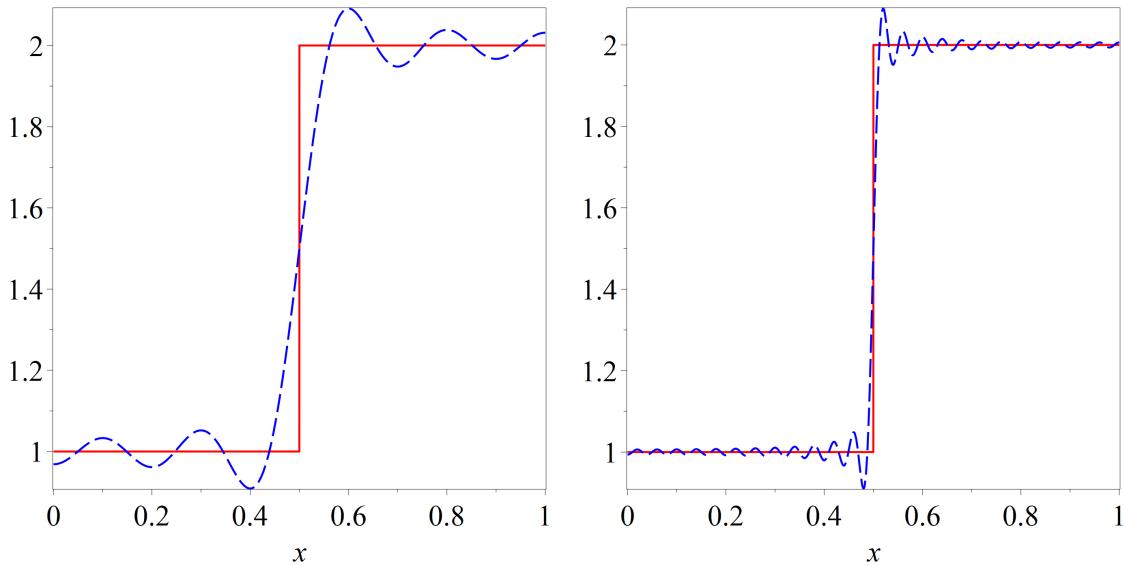


Figure 8.9: Left panel: function $f(x)$ (solid red) and cosine approximation $s_{10}(x)$ (dashed blue). Right panel: function $f(x)$ (solid red) and cosine approximation $s_{50}(x)$ (dashed blue).

It's also worth noting that even though $f(x)$ need not be defined outside the interval $0 \leq x \leq L$, the Fourier cosine expansion $s_n(x)$ is defined for all real x and has a certain symmetries and periodicity on the real line; see Exercise 8.2.3.

Reading Exercise 8.2.2 Let $f(x) = \cos(2\pi x)$ on the interval $0 \leq x \leq 1$. Use (8.49) with (8.46) to write out $s_0(x), s_1(x), s_2(x), s_3(x)$, and $s_4(x)$. Why do the results make perfect sense?

Application of the Cosine Expansion to JPEG Image Compression

The Fourier cosine series is the basis for JPEG image compression. Images are naturally two-dimensional entities, functions of x and y , but for simplicity we will describe the general process for a function of a single independent variable. In the examples that follow we'll use t for this independent variable (instead of x) and think of t as time. So consider a function $f(t)$ where t is time; you might think of $f(t)$ as quantifying an audio signal, for example, $f(t)$ could be a measurement of the pressure level at a microphone at time t . Also, once the signal is digitized and in the computer we wouldn't have an actual function of time t , but rather sampled values of the function $f(t)$ at periodic time intervals. See the Project “Frequency Analysis of Signals” in Section 8.5.4 for more on audio signals and sampling. Let us ignore these issues in order to focus on the main idea behind this type of compression.

To consider a concrete example, let

$$f(t) = t - 2t^2$$

be defined on the interval $0 \leq t \leq 1$. The goal is to store this function in a computer, or transmit it over a channel, using as little data as possible. The idea behind JPEG compression is to

1. Compute the Fourier cosine coefficients a_k for $f(t)$ up to some maximum value N using (8.46).
2. Discard Fourier coefficients whose magnitude falls below some threshold, then round all others to fixed values, for example, the nearest multiple of 0.1. This should result in many coefficients being rounded to zero.
3. The compressed signal consists of the nonzero rounded Fourier cosine coefficients a_k ; these are what we transmit or store instead of $f(t)$.

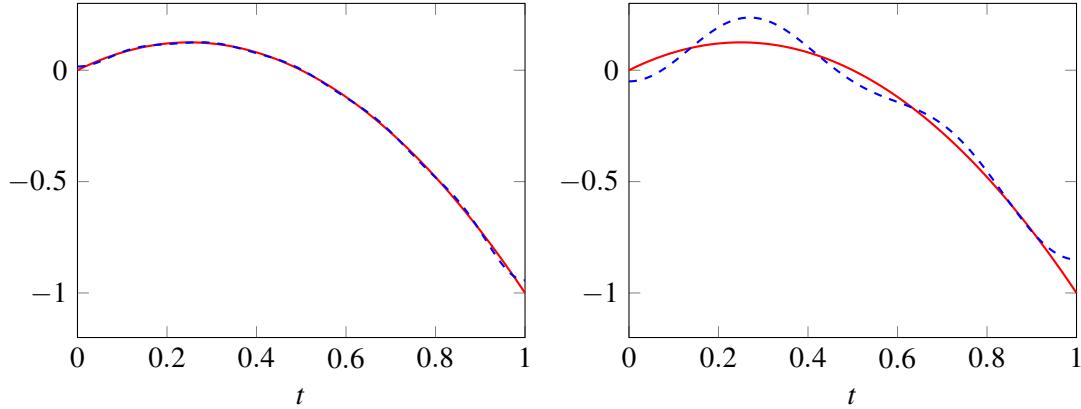


Figure 8.10: Left panel: graph of function $f(t) = t - 2t^2$ (solid red) and 10-term Fourier cosine approximation $s_{10}(t)$ (dashed blue). Right panel: graph of $f(t)$ (solid red) and truncated/rounded Fourier cosine approximation $\tilde{f}(t)$ (dashed blue).

4. An approximation to the original signal $f(t)$ is obtained by using these rounded Fourier cosine coefficients in (8.49).

■ **Example 8.5** Let's apply this to the function $f(t) = t - 2t^2$. Step 1 is to compute the cosine coefficients a_k for $f(t)$ up to some limit $k = N$. In this case we will use $N = 10$ and we find that

$$\begin{aligned} s_{10}(t) \approx & -0.1667 + 0.4053 \cos(\pi t) - 0.2026 \cos(2\pi t) + 0.0450 \cos(3\pi t) \\ & - 0.0507 \cos(4\pi t) + \cdots - 0.0081 \cos(10\pi t). \end{aligned}$$

The function $s_{10}(t)$ is a good approximation to $f(t)$, as shown in the left panel of Figure 8.10. However, s_{10} contains a lot of information (11 coefficients a_k , each a floating point number that requires at least four bytes to store). In Step 2 we choose a positive compression parameter r and then round each cosine coefficient to the nearest multiple of r , which has the effect of rounding all a_k with $|a_k| < r/2$ to zero. In the present case we use $r = 0.1$, so all a_k with $|a_k| < 0.05$ will be rounded to zero. This zeros out all a_k for this particular $f(t)$, except a_0, a_1, a_2 , and a_4 , which are rounded to $a_0 = -0.3, a_1 = 0.4, a_2 = -0.2$, and $a_4 = -0.1$. This is where a lot of compression occurs. This short list of four numbers constitutes the compressed version of the signal that we store or transmit (though further economization is possible).

In order to reconstitute an approximation $\tilde{f}(t)$ to the signal $f(t)$ we use these rounded coefficients in (8.49) to form

$$\tilde{f}(t) \approx -0.15 + 0.4 \cos(\pi t) - 0.2 \cos(2\pi t) - 0.1 \cos(4\pi t)$$

The compressed version of the signal is graphed in the right panel of Figure 8.10. By making the parameter r closer to zero we retain more of the a_k , and those that are retained are less-harshly rounded, so $\tilde{f}(t)$ is a better reproduction of $f(t)$, but at the expense of requiring more storage. ■

The full JPEG algorithm has various refinements. For example, the image is not compressed as a whole, but is broken up into smaller portions. Also, more sophisticated rounding is used and the final rounded set of Fourier coefficients is subjected to additional compression algorithms. See [26] for more details.

8.2.4 The Fourier Sine Expansion

Here's some good news: The Fourier cosine expansion process also works if we replace all cosines with sines. Specifically, given a function $f(x)$ defined on $0 \leq x \leq L$ let

$$s_n(x) = b_1 \sin(\pi x/L) + b_2 \sin(2\pi x/L) + \cdots + b_n \sin(n\pi x/L) \quad (8.50)$$

where

$$b_k = \frac{2}{L} \int_0^L f(x) \sin(k\pi x/L) dx. \quad (8.51)$$

Note that (8.51) always yields $b_0 = 0$, so that's why this term is omitted in (8.50).

Theorem 8.2.3 Let f be a bounded piecewise continuous function on the interval $0 \leq x \leq L$ and suppose $s_n(x)$ is defined by (8.50). Then the quantity $\|f - s_n\|_2$ is minimized when the b_k are given by (8.51). Moreover with the b_k so defined

$$\lim_{n \rightarrow \infty} \|f - s_n\|_2 = 0. \quad (8.52)$$

Theorem 8.2.3 may be summarized by saying that the set of functions

$$S = \{\sin(\pi x/L), \sin(2\pi x/L), \sin(3\pi x/L), \dots\}$$

is **complete** in the space of piecewise continuous functions on the interval $0 \leq x \leq L$.

As in the case of Fourier cosine series, we may write

$$f(x) = \sum_{j=1}^{\infty} b_j \sin(j\pi x/L) \quad (8.53)$$

with the understanding that this really means that (8.52) holds. The expansion (8.53) is called the **Fourier sine series** for the function f .

The computation of the b_k is facilitated by the fact that the functions $\sin(j\pi x/L) \sin(k\pi x/L)$ are orthogonal on the interval $0 \leq x \leq L$ when $j \neq k$, that is,

$$\int_0^L \sin(j\pi x/L) \sin(k\pi x/L) dx = 0 \quad (8.54)$$

when j and k are distinct integers.

■ **Example 8.6** Consider $f(x) = 30x^3 - 41x^2 + 17x - 2$ defined on $0 \leq x \leq 1$, the function examined in Section 8.2.1. The Fourier sine expansion s_{10} for f is

$$s_{10}(x) \approx 0.2412 \sin(\pi x) - 0.4586 \sin(2\pi x) + \dots - 0.3704 \sin(10\pi x)$$

(intermediate terms omitted) and a plot of $f(x)$ and s_{10} is shown in the left panel of Figure 8.11. The right panel shows $s_{50}(x)$. We can compute $\|f - s_{10}\|_2 \approx 0.6045$ and $\|f - s_{50}\|_2 \approx 0.2821$ ■

Figure 8.11 of Example 8.6 illustrates one peculiarity of the Fourier sine expansion of a function on an interval $0 \leq x \leq L$. Because each term $\sin(k\pi x/L)$ is equal to zero when $x = 0$ or $x = L$, the approximation $s_n(x)$ also satisfies $s_n(0) = s_n(L) = 0$ for any finite value of n , and of course at $x = 0$ and $x = L$ these expressions limit to zero as n increases. This is the source of the strange behavior of the Fourier sine expansion at the endpoints of the interval. Nonetheless, it is still true that $\lim_{n \rightarrow \infty} \|f - s_n\|_2 = 0$.

As with the Fourier cosine expansion, the sine expansion (8.50) is defined for all real x , even though $f(x)$ need not be defined outside the interval $0 \leq x \leq L$. The sine expansion also has a certain symmetries and periodicity on the real line; see Exercise 8.2.3.

8.2.5 More on Fourier Series Convergence

Sometimes we want to say more about the convergence of s_n to f for Fourier sine or cosine series. For example, what does the Fourier cosine expansion of Example 8.4 do at the discontinuity at

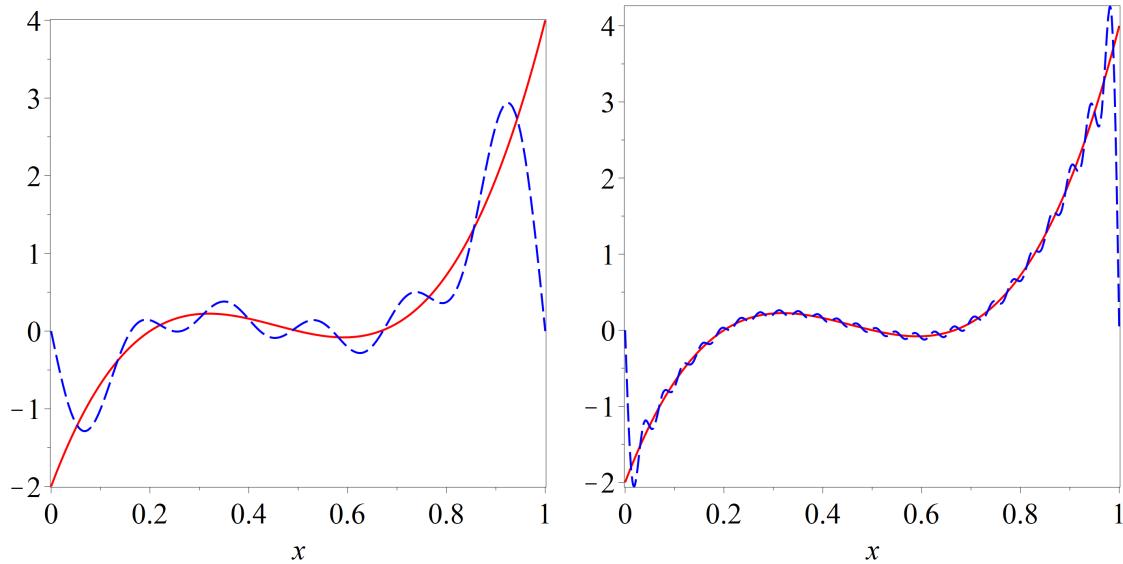


Figure 8.11: Left panel: function $f(x) = 30x^3 - 41x^2 + 17x - 2$ (solid red) and Fourier sine approximation $s_{10}(x)$ (dashed blue). Right panel: function $f(x) = 30x^3 - 41x^2 + 17x - 2$ (solid red) and Fourier sine approximation $s_{50}(x)$ (dashed blue).

$x = 1/2$ as the number of terms n increases? Or in Example 8.6, it's clear that $s_n(0)$ does not converge to $f(0)$ (since $s_n(0) = 0$ for all n and $f(0) = 2$) and similarly $s_n(1)$ does not converge to $f(1)$. But does $s_n(x)$ converge to $f(x)$ at all other points? These are more delicate issues and the details of the analysis are beyond what we will consider, but some basic truths are easily summarized.

Pointwise Convergence

Recall the definition of a piecewise continuous function given in Definition 5.2.2. In the present case we will require that f has only finitely many discontinuities in the interval $0 \leq x \leq L$, that these are jump discontinuities, and that f is bounded. We also require that f is differentiable at all but finitely many points in $0 \leq x \leq L$, and that f' is itself piecewise continuous. Informally, the graph of f has finitely many jump discontinuities and finitely many “corners.”

With these assumptions it can be shown that

1. At any point x_0 with $0 \leq x_0 \leq L$ at which f is continuous, the partial sums $s_n(x_0)$ for the Fourier cosine series converge to $f(x_0)$ as $n \rightarrow \infty$.
2. At any point x_0 with $0 < x_0 < L$ at which f is continuous, the partial sums $s_n(x_0)$ for the Fourier sine series converge to $f(x_0)$ as $n \rightarrow \infty$. But since $s_n(0)$ and $s_n(L)$ equal zero for all n , these quantities converge to 0.
3. At a point x_0 with $0 < x_0 < L$ at which f has a jump discontinuity the Fourier sine and cosine partial sums converge to

$$\frac{1}{2} \left(\lim_{t \rightarrow t_0^-} f(t) + \lim_{t \rightarrow t_0^+} f(t) \right), \quad (8.55)$$

which is the average value of the limiting value of f from the left and right at $x = x_0$.

For a proof of these facts see [105]. When $s_n(x_0)$ converges to $f(x_0)$ at a specific point $x = x_0$ we say that s_n **converges pointwise to f at $x = x_0$** . So for example, point (1) is the statement that the partial sums s_n of the Fourier cosine series converge pointwise to f at all points at which f is continuous.

■ **Example 8.7** For the piecewise constant function $f(x)$ defined in Example 8.4 we found the Fourier cosine expansion of $f(x)$, and based on that expansion the n th partial sum is (for n odd)

$$s_n(x) = \frac{3}{2} - \frac{2}{\pi} \cos(\pi x) + \frac{2}{3\pi} \cos(3\pi x) - \frac{2}{5\pi} \cos(5\pi x) + \frac{2(-1)^{(n+1)/2}}{n\pi} \cos(n\pi x),$$

while $s_{n+1} = s_n$ when n is odd. When $x = 1/2$ all of the cosine terms are zero and so $s_n(1/2) = 3/2$ for all n , which clearly converges to $3/2$ as $n \rightarrow \infty$. This is in accordance with (8.55) since the one-sided limits at $x = 1$ are $\lim_{x \rightarrow 1^-} f(x) = 1$ and $\lim_{x \rightarrow 1^+} f(x) = 2$.

An amusing result can be obtained by considering the behavior of s_n at $x = 0$. Based on the discussion of pointwise convergence for the cosine series it follows that $s_n(0)$ converges to $f(0) = 1$, which in conjunction with $\cos(0) = 1$ leads to the conclusion that

$$\lim_{n \rightarrow \infty} \left(\frac{3}{2} - \frac{2}{\pi} + \frac{2}{3\pi} - \frac{2}{5\pi} + \frac{2(-1)^{(n+1)/2}}{n\pi} \right) = 1$$

where the limit is taken over odd n . A little rearrangement (subtract $3/2$ from both sides above, multiply by $-\pi/2$, take the limit as n approaches infinity) yields the strange fact that

$$1 - 1/3 + 1/5 - 1/7 + \cdots = \pi/4.$$

■

Reading Exercise 8.2.3 Let $f(x) = x$ on the interval $0 \leq x \leq 1$. Show that the Fourier cosine coefficient $a_0 = 1$, $a_k = 0$ when k is even and $k \geq 2$, and $a_k = -4/(k\pi)^2$ when k is odd, so that

$$f(x) = \frac{1}{2} - \frac{4}{\pi^2} \sum_{k=1, \text{ odd}}^{\infty} \frac{\cos(k\pi x)}{k^2}$$

(with the understanding that the sum of the right converges to f in the L^2 norm or pointwise if $0 \leq x \leq 1$). Verify directly that the Fourier series yields $f(1/2)$ when $x = 1/2$. Then substitute $x = 0$ into the Fourier cosine series and use it to show that

$$1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots = \frac{\pi^2}{8}.$$

Uniform Convergence

As already remarked, when f is piecewise smooth, $s_n(x_0)$ converges to $f(x_0)$ at any point $x = x_0$ where f is continuous. That is, s_n converges pointwise to f at all points at which f is continuous, at least for $0 < x_0 < L$, and even at the endpoints for the cosine series. An even stronger form of convergence is that s_n converges uniformly to f . This means that

$$\lim_{n \rightarrow \infty} \max_{0 \leq x \leq L} |f(x) - s_n(x)| = 0. \quad (8.56)$$

If (8.56) holds then for any particular choice of x_0 we must have $\lim_{n \rightarrow \infty} s_n(x_0) = f(x_0)$. This means that if s_n converges uniformly to f then s_n converges pointwise to f at each x_0 . (The converse is not true, however.) The condition (8.56) means that $s_n(x)$ can be made to stay uniformly close to $f(x)$ over the entire interval $0 \leq x \leq L$ by taking n sufficiently large.

■ **Example 8.8** Uniform convergence is illustrated in Figure 8.12, with the function $f(x) = x(1 - |x - 1/2|)$. In the left panel the Fourier cosine expansion the function $s_1(x)$ is shown as the dashed blue graph, along with $f(x)$ as the solid red curve. Also shown as dotted black curves are the graphs of $f(x) + 0.05$ and $f(x) - 0.05$, which form a band of vertical span 0.1 centered on the

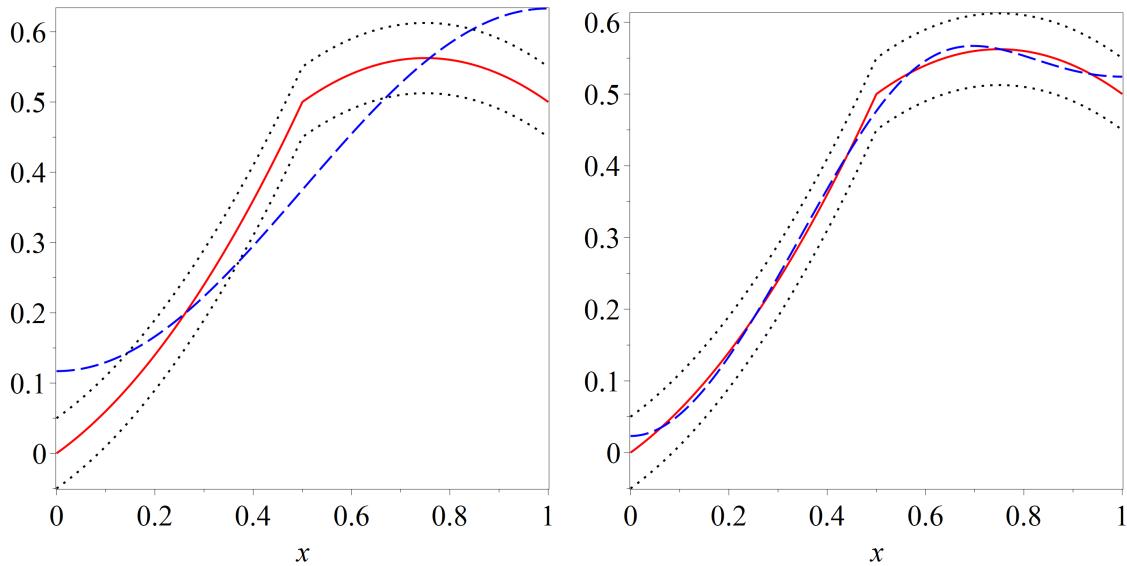


Figure 8.12: Left panel: function $f(x) = x(1 - |x - 1/2|)$ (solid red), Fourier cosine approximation $s_1(x)$ (dashed blue) and graphs of $f(x) \pm 0.05$ (dotted black). Right panel: same, but with Fourier cosine approximation $s_5(x)$.

graph of f . Note that the graph of $s_1(x)$ does not lie within this ± 0.05 band centered on f . Put another way, the inequality $|s_1(x) - f(x)| < 0.05$ is not true for all x in the interval $[0, 1]$. However, by increasing n to 5 we see that $s_5(x)$, shown as the dashed blue curve in the right panel, now lies within this ± 0.05 tolerance band around the graph of f , so $|f(x) - s_5(x)| < 0.05$ is true for all x in the interval $[0, 1]$.

If we tighten up the tolerance limits around the graph of f to $f(x) \pm 0.01$ then the graph of $s_5(x)$ may not fall within these limits, but by increasing n sufficiently we will find that $s_n(x)$ does satisfy $|s_n(x) - f(x)| < 0.01$ for all x in $[0, 1]$. Moreover, the tolerance 0.01 can be made arbitrarily small (but positive) and we can still take n large enough so that $s_n(x)$ stays within such a prescribed tolerance over the whole interval. In this case the partial sums $s_n(x)$ converge uniformly to $f(x)$ on the interval $0 \leq x \leq 1$.

Contrast the cosine expansion with the Fourier sine expansion for this function. In Figure 8.13 we show the analogous situation with the Fourier sine expansions s_{10} and s_{50} . Because $s_n(1) = 0$ for all n but $f(1) = 1/2$, it's easy to see that we can never obtain $|s_n(x) - f(x)| < 0.05$ for all x in $[0, 1]$ by using the Fourier sine expansion, in particular, when $x = 1$. More generally, no matter how large n is, we can never obtain $|s_n(x) - f(x)| < \varepsilon$ for any choice of ε if $\varepsilon < 1/2$ since $|s_n(1) - f(1)| = 1/2$ for all n . ■

There are a variety of conditions under which it can be assured that the Fourier sine or cosine approximations converge uniformly. The following theorem illustrates one such set of conditions.

Theorem 8.2.4 Suppose f is continuous and f' is piecewise continuous and bounded on the interval $0 \leq x \leq L$. Then the Fourier cosine approximations converge uniformly to f on $0 \leq x \leq L$. If $f(0) = f(L) = 0$ the Fourier sine approximations converge uniformly to f on $0 \leq x \leq L$.

For a proof of this see [105].

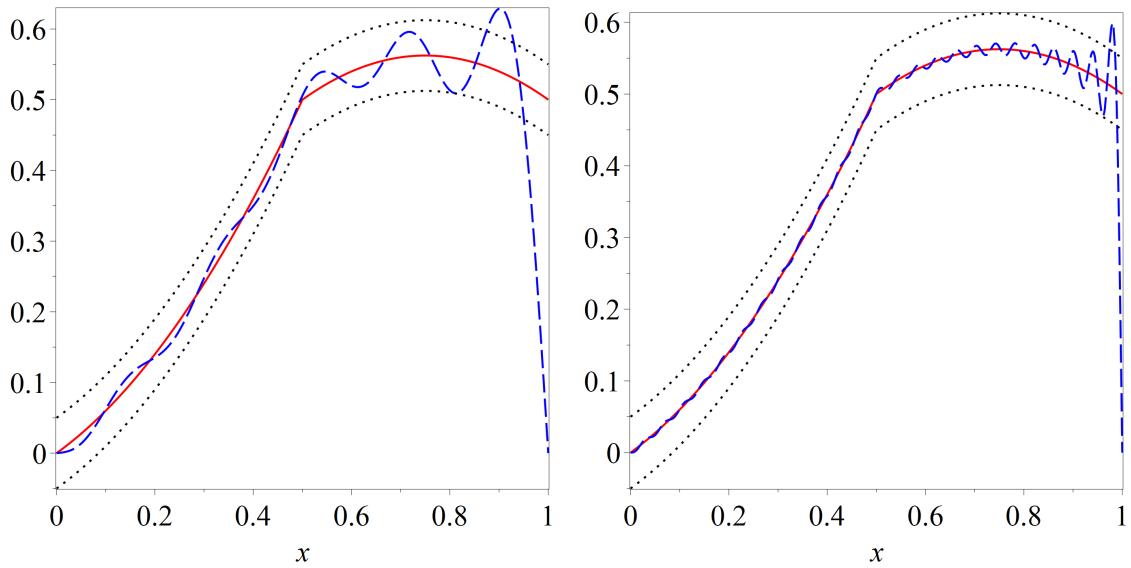


Figure 8.13: Left panel: function $f(x) = x(1 - |x - 1/2|)$ (solid red), Fourier sine approximation $s_{10}(x)$ (dashed blue) and graph of $f(x) \pm 0.05$ (dotted black). Right panel: same, but with Fourier sine approximation $s_{50}(x)$.

Conclusion

Fourier expansions can also be performed on an interval $0 \leq x \leq L$ by using complex exponential functions of the form $e^{2\pi i k x / L}$ and using these functions has certain advantages; see Exercise 8.2.10. Fourier expansions are not limited to intervals of the form $0 \leq x \leq L$ either. The interval $-L \leq x \leq L$ also commonly occurs, and in fact one can perform Fourier expansion on any interval $a \leq x \leq b$. One can even perform expansions similar to Fourier series but using families of functions that are not exponential or trigonometric in nature. For an example, see Exercise 8.2.9.

8.2.6 Exercises

Exercise 8.2.1 For each function $f(x)$ below on the indicated interval, use (8.46) and (8.49) to compute $s_n(x)$ for the Fourier cosine approximation for the indicated values of n . In each case plot $s_n(x)$ and $f(x)$ on the interval $0 \leq x \leq L$, and compute $\|f - s_n\|_2$.

- (a) $f(x) = 3\cos(2\pi x)$ on $0 \leq x \leq 1$, $n = 0, 1, 2, 3, 5$. Hint: the results of Reading Exercise 8.2.1 make short work of this.
- (b) $f(x) = 3\cos(2\pi x) - 4\cos(4\pi x)$ on $0 \leq x \leq 1$, $n = 0, 1, 2, 3, 5$. Same hint as (a).
- (c) $f(x) = x$ on $0 \leq x \leq 2$, $n = 0, 1, 5$.
- (d)

$$f(x) = \begin{cases} 0, & x < 2 \\ 5, & x \geq 2 \end{cases}$$

on $0 \leq x \leq 3$, $n = 0, 1, 5, 10$.

(e)

$$f(x) = \begin{cases} \cos(\pi x), & x < 2 \\ \cos(2\pi x), & x \geq 2 \end{cases}$$

on $0 \leq x \leq 3$, $n = 0, 3, 5, 10$.

- (f) $f(x) = \delta(x - 1)$ on $0 \leq x \leq 2$, where δ is the Dirac delta function of Section 5.4. Note that $\delta(x - 1)$ is not bounded (it's not even a function), so the conclusions of Theorem 8.2.2 don't hold or even make sense here ($\|f - s_n\|_2$ is not defined), but we can still compute the Fourier cosine coefficients a_k and form s_n . Use $n = 5, 10, 20$ and take note of how s_n behaves as n increases. Does the cosine expansion seem to approximate a delta function? Hint: The necessary integrals for a_k can be computed (even without a computer) by using (5.69).

Exercise 8.2.2 For each function $f(x)$ below on the indicated interval, use (8.50) and (8.51) to compute $s_n(x)$ for the Fourier sine approximation for the indicated values of n . In each case plot $s_n(x)$ and $f(x)$ on the interval $0 \leq x \leq L$, and compute $\|f - s_n\|_2$.

- (a) $f(x) = 1$ on $0 \leq x \leq 1$, $n = 1, 3, 10$.
- (b) $f(x) = 3 \sin(2\pi x) - 4 \sin(4\pi x)$ on $0 \leq x \leq 1$, $n = 1, 2, 3, 5$. Hint: (8.54) makes this easy.
- (c) $f(x) = x$ on $0 \leq x \leq 2$, $n = 1, 5, 10$.
- (d) $f(x) = x(2 - x)$ on $0 \leq x \leq 2$, $n = 1, 3, 5$.

Exercise 8.2.3

- (a) Recall that the cosine function is even ($\cos(-x) = \cos(x)$) and periodic with period 2π . Use this to show that the Fourier cosine expansion (8.49) of a function is even for any n , so $s_n(-x) = s_n(x)$, and that s_n is periodic with period $2L$.
- (b) Compute the Fourier cosine expansion s_{10} for the function $f(x) = x$ on the interval $0 \leq x \leq 1$ and then plot $s_{10}(x)$ on the interval $-3 \leq x \leq 3$. Explain what you see in light of part (a).
- (c) Recall that the sine function is odd ($\sin(-x) = -\sin(x)$) and periodic with period 2π . Use this to show that the Fourier sine expansion (8.50) of a function is odd for any n , so $s_n(-x) = -s_n(x)$, and that s_n is periodic with period $2L$.
- (d) Compute the Fourier sine expansion s_{10} for the function $f(x) = x$ on the interval $0 \leq x \leq 1$ and then plot $s_{10}(x)$ on the interval $-3 \leq x \leq 3$. Explain what you see in light of part (c).

Exercise 8.2.4 Use the trigonometric identity $\sin(x)\sin(y) = (\cos(x - y) - \cos(x + y))/2$ to show that (8.54) holds when j and k are positive integers with $j \neq k$. What is the value of the integral on the left in (8.54) when $j = k$?

Exercise 8.2.5 Argue that if ϕ is a nonnegative and continuous real-valued function on an interval $a \leq x \leq b$ (assume $a < b$) and

$$\int_a^b \phi(x) dx = 0$$

then $\phi(x) = 0$ for all x in this interval. Why does this show that if $\|f\|_2 = 0$ for some real-valued continuous function f then f is the zero function? Why does this show that if $\|g - h\|_2 = 0$ for continuous real-valued functions g and h then $g(x) = h(x)$ at all points? Hint: draw a picture of a nonnegative (but nonzero) function ϕ on an interval $a \leq x \leq b$ and interpret the integral

$\int_a^b \phi(x) dx$ as the area under the curve.

Exercise 8.2.6 Fourier series can also be performed on an interval $[-L, L]$. The traditional approximation for a function $f(x)$ makes use of terms of the form $\cos(j\pi x/L)$ and $\sin(j\pi x/L)$ and is of the form

$$s_n(x) = \frac{a_0}{2} + \sum_{j=1}^n (a_j \cos(j\pi x/L) + b_j \sin(j\pi x)) \quad (8.57)$$

where

$$\begin{aligned} a_j &= \frac{1}{L} \int_{-L}^L f(x) \cos(j\pi x/L) dx \\ b_j &= \frac{1}{L} \int_{-L}^L f(x) \sin(j\pi x/L) dx \end{aligned} \quad (8.58)$$

for $j \geq 0$, although $b_0 = 0$ can be omitted.

Use (8.57) and (8.58) to compute Fourier sine/cosine expansions for the following functions on the indicated interval for the given values of n . Graph the function and the approximations.

- (a) $f(x) = 3 \cos(2\pi x) - \sin(\pi x)$ on $-1 \leq x \leq 1$, $n = 0, 1, 5$.
- (b) $f(x) = 1 - x^2$ on $-1 \leq x \leq 1$, $n = 0, 2, 5$.
- (c) $f(x) = x$ on $-3 \leq x \leq 3$, $n = 0, 2, 5$.
- (d) $f(x) = e^x$ on $-2 \leq x \leq 2$, $n = 0, 2, 5$.

Exercise 8.2.7 Consider the function

$$f(t) = 30t^3 - 41t^2 + 17t - 2$$

defined on the interval $0 \leq t \leq 1$, which is the function used as an example in Section 8.2.1 and Example 8.2, although there we considered f as a function of x . Compute a cosine approximation for f out to 20 terms. Then apply the JPEG compression process used in Example 8.5 to compress $f(t)$ down to a small set a_0, \dots, a_m of rounded Fourier coefficients; round each to the nearest multiple of $r = 0.1$, which should round many coefficients to zero. How many Fourier coefficient remains nonzero after rounding? Compare the graph of $\tilde{f}(t)$ to that of $f(t)$.

Repeat for rounding parameters $r = 0.5$ and $r = 0.01$.

Exercise 8.2.8 The Fourier sine and cosine series also work for functions of two (or more) variables. Let $f(x, y)$ be a continuous function defined on a rectangle $0 \leq x \leq A$, $0 \leq y \leq B$. We can approximate $f(x, y)$ with a Fourier cosine series of the form

$$\begin{aligned} s_{m,n}(x, y) &= \frac{a_{0,0}}{4} + \frac{1}{2} \sum_{j=0}^m a_{j,0} \cos(j\pi x/A) + \frac{1}{2} \sum_{k=0}^n a_{0,k} \cos(k\pi y/B) \\ &\quad + \sum_{j=0}^m \sum_{k=0}^n a_{j,k} \cos(j\pi x/A) \cos(k\pi y/B). \end{aligned} \quad (8.59)$$

The appropriate choice for the coefficients $a_{j,k}$ is

$$a_{j,k} = \frac{4}{AB} \int_0^A \int_0^B f(x,y) \cos(j\pi x/A) \cos(k\pi y/B) dy dx. \quad (8.60)$$

It can be shown that $\lim_{m,n \rightarrow \infty} \|f - s_{m,n}\|_2 = 0$ where the L^2 norm of a function ϕ defined on the rectangle is

$$\|\phi\|_2 \left(\int_0^A \int_0^B \phi^2(x,y) dy dx \right)^{1/2}.$$

- (a) Let $f(x,y) = x(1-x) + xy$ on the rectangle $0 \leq x \leq 2, 0 \leq y \leq 1$. Compute $s_{m,n}(x,y)$ for the choices $(m,n) = (0,0)$, $(m,n) = (2,2)$, and $(m,n) = (10,10)$ (you'll need a computer algebra system for the last one). In each case plot $f(x,y)$ and $s_{m,n}(x,y)$.
- (b) Make the obvious modification to (8.59) and (8.60) (in particular, omit all terms in which $j = 0$ or $k = 0$) to find formulas for a Fourier sine expansion involving terms $\sin(j\pi x/A) \sin(k\pi y/B)$. Repeat part (a) with a Fourier sine expansion but for the choices $(m,n) = (1,1)$, $(m,n) = (2,2)$, and $(m,n) = (10,10)$

Exercise 8.2.9 Functions other than sines and cosines can be used to approximate functions on an interval. For example, consider the polynomials

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = x^2 - 1/3$$

on the interval $0 \leq x \leq 1$. These polynomials (along with higher degree polynomials) can be used in a manner analogous to sines or cosines for approximation.

- (a) Show that these polynomials are orthogonal to each other on the interval $-1 \leq x \leq 1$, that is,

$$\int_{-1}^1 p_j(x) p_k(x) dx = 0$$

for $j \neq k$.

- (b) Suppose we want to approximate a function $f(x)$ as $f(x) \approx s_2(x)$ where

$$s_2(x) = c_0 p_0(x) + c_1 p_1(x) + c_2 p_2(x)$$

by choosing the coefficients c_k to minimize

$$\|f - s_2\|_2^2 = \int_0^1 (f(x) - c_0 p_0(x) - c_1 p_1(x) - c_2 p_2(x))^2 dx.$$

Show that this requires choosing

$$c_k = \frac{\int_{-1}^1 f(x) p_k(x) dx}{\int_{-1}^1 p_k^2(x) dx}.$$

- (c) Let $f(x) = e^x$. Use part (b) to construct $s_2(x)$, which should be a quadratic polynomial in x . Plot $s_2(x)$ and e^x on $-1 \leq x \leq 1$. Also plot $e^x - s_2(x)$ and $e^x - T_2(x)$ where $T_2(x) = 1 + x + x^2/2$ is the Taylor polynomial that approximates e^x near $x = 0$. Which approximation has the smallest magnitude error on this interval?
- (d) Construct a cubic addition $p_3(x) = x^3 + Ax^2 + Bx + C$ to this orthogonal family by requiring that

$$\int_{-1}^1 p_k(x)p_3(x) dx = 0$$

for $k = 0, 1, 2$. This should yields three equations in three unknowns A, B , and C , which can be solved for this constants. Then use $p_3(x)$ in the approximation of part (b)/(c) to approximate $f(x) = e^x$ on $-1 \leq x \leq 1$.

Exercise 8.2.10 Fourier series can be performed using complex exponential functions instead of sines or cosines, and this might be viewed as the most mathematically elegant form of Fourier approximation. In particular, let $f(x)$ be a function defined on the interval $0 \leq x \leq L$ (other intervals can be used), and suppose f is piecewise smooth. Then it can be shown that f can be approximated to arbitrary accuracy in the L^2 norm using a sum of the form

$$s_n(x) = \sum_{k=-n}^n c_k e^{2\pi i k x / L} \quad (8.61)$$

if the c_k are chosen correctly. Moreover, $\lim_{n \rightarrow \infty} \|f - s_n\|_2 = 0$. In general the c_k will be complex numbers.

In what follows Definition 8.2.2 for orthogonal functions must be amended when the functions involved are complex-valued, like complex exponential function such as $e^{iy} = \cos(y) + i\sin(y)$. Two complex-valued functions $g(x)$ and $h(x)$ are **orthogonal** on the interval $a \leq x \leq b$ if

$$\int_a^b g(x) \overline{h(x)} dx = 0 \quad (8.62)$$

where $\overline{h(x)}$ denotes the complex-conjugate of $h(x)$. That is, the function h must be conjugated (the role of g and h can be reversed). If g and h are real-valued then $\overline{h} = h$ and (8.62) becomes Definition 8.2.2. Also, the definition of $\|g\|_2$ for a complex-valued function on an interval $a \leq x \leq b$ is

$$\|g\|_2 = \left(\int_a^b g(x) \overline{g(x)} dx \right)^{1/2} = \left(\int_a^b |g(x)|^2 dx \right)^{1/2} \quad (8.63)$$

since $z\bar{z} = |z|^2$ for any complex number z . If g is real-valued then $|g|^2 = g^2$ and this is the same as Definition 8.2.1.

- (a) Use (8.62) to verify that the functions $e^{2\pi i j x / L}$ and $e^{2\pi i k x / L}$ are orthogonal on the interval $0 \leq x \leq L$ if j and k are integers and $j \neq k$. Also show that the value of

$$\int_0^L e^{2\pi i k x / L} \overline{e^{2\pi i k x / L}} dx = L$$

for all integers k . Hint: $\overline{e^{2\pi i kx/L}} = e^{-2\pi i kx/L}$.

- (b) An argument parallel to that for cosine functions in Section 8.2.3 and in particular (8.44) with $e^{2\pi i kx/L}$ replacing $\cos(k\pi x/L)$, shows that the choice for the c_k in (8.61) that minimizes $\|f - s_n\|_2^2$ with the L^2 norm defined by (8.63) is

$$c_k = \frac{1}{L} \int_0^L f(x) e^{-2\pi i kx/L} dx. \quad (8.64)$$

Take $f(x) = (x - 1/2)^2$ on the interval $0 \leq x \leq 2$. Compute c_k using (8.64) for $k = -5$ to $k = 5$ and then form the sum $s_5(x)$ defined by (8.61). Plot $s_5(x)$ and $f(x)$ on $0 \leq x \leq 2$; despite the complex exponentials, s_5 will be real-valued, but you may need to tell your software to plot the real part of s_5 . Repeat for s_{20} . Does s_n approximate f better as n increases?

8.3 Solving the Heat Equation

8.3.1 Homogeneous Dirichlet Conditions

We are now in a position to exhibit the solution to the heat equation (8.17) on an interval $0 \leq x \leq L$ with homogeneous Dirichlet boundary conditions $u(0, t) = u(L, t) = 0$ and initial condition $u(x, 0) = f(x)$, at least if f is piecewise smooth. Note that we refer to “the” solution to the heat equation with this boundary and initial data: as you can show in Exercise 8.3.9, the solution is unique.

The solution to $\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2}$ on the interval $0 < x < L$ with homogeneous Dirichlet boundary conditions $u(0, t) = u(L, t) = 0$ and initial data $u(x, 0) = f(x)$ is given by the infinite sum

$$u(x, t) = \sum_{k=1}^{\infty} b_k e^{-\alpha k \pi^2 t / L^2} \sin(k\pi x / L) \quad (8.65)$$

where the b_k are the Fourier sine coefficients of $f(x)$, given by

$$b_k = \frac{2}{L} \int_0^L f(x) \sin(k\pi x / L) dx. \quad (8.66)$$

The justification that (8.65)-(8.66) provides a solution is not completely cut and dried, at least from a rigorous mathematical perspective. The manipulation of infinite sums as on the right in (8.65) can be a bit delicate, especially manipulations like term-by-term differentiation. We will not discuss these issues here, but merely note that the manipulations necessary for the following argument can be justified. Specifically, as remarked in Section 8.1.5 the heat equation is linear, so linear combinations of solutions are also solutions. Since each summand $e^{-\alpha k \pi^2 t / L^2} \sin(k\pi x / L)$ in (8.65) satisfies the heat equation, so does $u(x, t)$ for $t > 0$. Also, since each summand equals zero when $x = 0$ and $x = L$ it follows that $u(0, t) = u(L, t) = 0$ for all t . Finally, note that $u(x, 0) = \sum_{k=1}^{\infty} b_k \sin(k\pi x / L) = f(x)$, since the b_k are chosen as the Fourier sine coefficients for $f(x)$ on $0 \leq x \leq L$.

Remark 8.3.1 One word concerning boundary and initial conditions. With homogeneous Dirichlet boundary conditions $u(0, t) = u(L, t) = 0$ for $t > 0$, it makes the most physical sense to specify initial data $u(x, 0) = f(x)$ with the additional requirement that $f(0) = f(L) = 0$. It would be somewhat nonphysical to have the bar ends at a nonzero temperature at time $t = 0$ and then instantly transition to zero temperature for $t > 0$.

■ **Example 8.9** Consider the heat equation (8.17) on a bar spanned by $0 \leq x \leq 1$ with diffusivity $\alpha = 1$, homogeneous Dirichlet boundary conditions $u(0, t) = u(1, t) = 0$, and initial temperature

$u(x, 0) = f(x)$ where $f(x) = x(1 - x)$; note that $f(0) = f(1) = 0$. With this initial data the Fourier sine coefficients for $f(x)$ can be computed symbolically for all k as

$$b_k = 2 \int_0^1 x(1 - x) \sin(k\pi x) dx = \frac{4(1 - (-1)^k)}{k^3 \pi^3}$$

(a computer algebra system helps). Thus $b_k = 0$ for all even k , while $b_1 = 8/\pi^3, b_3 = 8/(27\pi^3)$, and generally $b_k = 8/(k^3\pi^3)$ when k is odd. The solution to the heat equation here can thus be written as the infinite sum

$$u(x, t) = \sum_{k=1}^{\infty} \frac{4(1 - (-1)^k)}{k^3 \pi^3} e^{-k^2 \pi^2 t} \sin(k\pi x) = \sum_{k=1, \text{odd}}^{\infty} \frac{8}{k^3 \pi^3} e^{-k^2 \pi^2 t} \sin(k\pi x).$$

A plot of $u(x, t)$ on the interval $0 \leq x \leq 1$ at times $t = 0, 0.1$, and $t = 0.3$ is shown in the left panel of Figure 8.14 and in the right panel as a function of both x and t . The temperature of the bar decreases rapidly to zero, the temperature at which the endpoints are held. ■

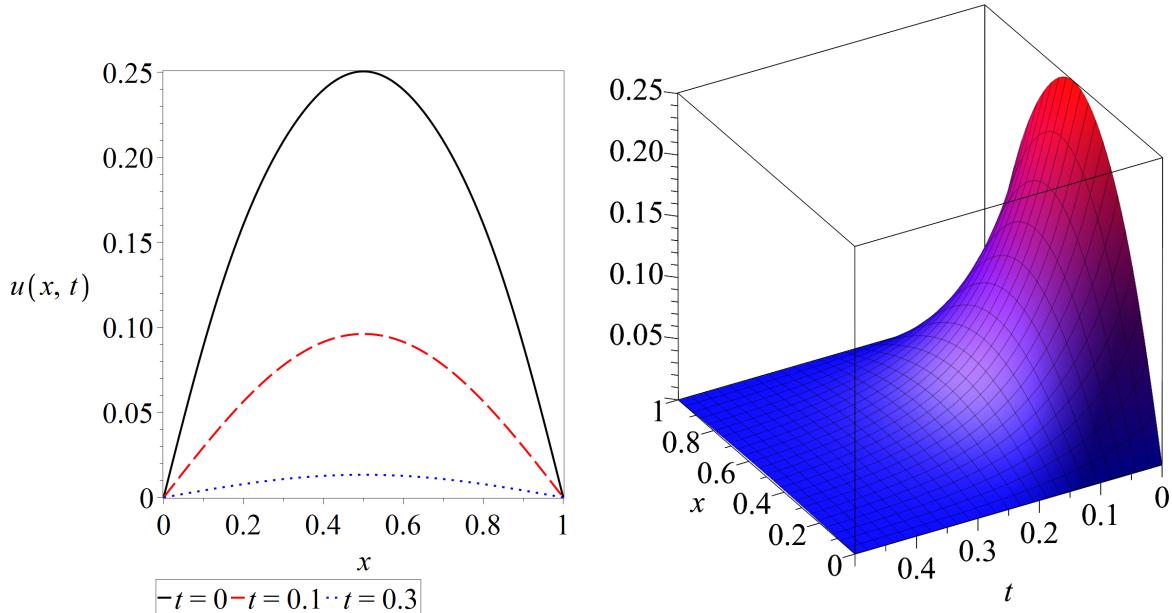


Figure 8.14: Left panel: solution $u(x, t)$ to heat equation with diffusivity $\alpha = 1$ on interval $0 \leq x \leq 1$, boundary conditions $u(0, t) = u(1, t) = 0$ and initial data $u(x, 0) = x(1 - x)$. Solution u at time $t = 0$ shown as solid black graph, $t = 0.1$ as dashed red graph, $t = 0.3$ as dotted blue graph. Right panel: same solution as a surface plot.

In practice it may not be possible to compute the b_k coefficients in any simple symbolic form, but one could compute the b_k numerically for $k = 1$ to $k = N$ for some choice of N . The solution $u(x, t)$ can then be approximated by using (8.65) with a finite upper limit of summation $k = N$.

Reading Exercise 8.3.1 Write out the solution to the heat equation in the setting of Example 8.9 but change the diffusivity to $\alpha = 2$. Plot the solution $u(x, t)$ on the interval $0 \leq x \leq 1$ at the same times, $t = 0, 0.1$, and 0.3 . What changes? Can you explain how the diffusivity affects the solution?

8.3.2 Insulating Boundary Conditions

Analogous reasoning that led to (8.65) shows that the solution to the heat equation on an interval $0 \leq x \leq L$ with insulating boundary conditions $\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(L, t) = 0$ and initial data $u(x, 0) = f(x)$

is given by

$$u(x,t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k e^{-\alpha k \pi^2 t / L^2} \cos(k \pi x / L) \quad (8.67)$$

where the a_k are the Fourier cosine coefficients of $f(x)$ on $0 \leq x \leq L$, given by

$$a_k = \frac{2}{L} \int_0^L f(x) \cos(k \pi x / L) dx. \quad (8.68)$$

■ Example 8.10 Consider the heat equation (8.17) on a bar spanned by $0 \leq x \leq 1$ with diffusivity $\alpha = 1$, insulating boundary data $\frac{\partial u}{\partial x}(0,t) = \frac{\partial u}{\partial x}(L,t) = 0$, and initial temperature $u(x,0) = f(x)$ where $f(x) = x^2(1-x)^2$. The Fourier cosine coefficients for $f(x)$ can be computed symbolically for all $k \geq 1$ as

$$a_k = 2 \int_0^1 x^2(1-x)^2 \cos(k \pi x) dx = -\frac{24(1+(-1)^k)}{k^4 \pi^4}$$

while $a_0 = 1/15$; a computer algebra system helps here too. Note that $a_k = 0$ for all odd k . The solution to the heat equation here can be written as the infinite sum

$$u(x,t) = \frac{1}{30} - \sum_{k=1}^{\infty} \frac{24(1+(-1)^k)}{k^4 \pi^4} e^{-k^2 \pi^2 t} \cos(k \pi x) = \frac{1}{30} - \sum_{k=2, \text{ even}}^{\infty} \frac{48}{k^4 \pi^4} e^{-k^2 \pi^2 t} \cos(k \pi x).$$

A plot of $u(x,t)$ on the interval $0 \leq x \leq 1$ at times $t = 0, 0.03$, and $t = 0.1$ is shown in the left panel of Figure 8.15 and a graph of $u(x,t)$ as a function of both x and t is shown in the right panel. The temperature of the bar decreases rapidly to a constant value, in this case $1/30$. If we think of the heat energy as a kind of fluid and the bar ends as walls that prevent heat energy from leaving, the energy or temperature “sloshes” to a constant average level. You can show this in Reading Exercise 8.3.2. ■

Reading Exercise 8.3.2 Use (8.67) with (8.68) to argue that the solution to the heat equation with insulating boundary conditions and initial data $u(x,0) = f(x)$ always settles to a constant value as t approaches infinity and this value is given by

$$\frac{1}{L} \int_0^L f(x) dx.$$

(Compare to the conclusions in Exercise 8.1.4.)

8.3.3 Other Boundary Conditions

Solutions to the heat equation with many other types of boundary conditions can be found using separation of variables and a Fourier-series approach. In this section we give a few examples of how this can be done. With a bit of experience, you should be able to deduce how to solve the heat equation even in settings you have not seen before, though depending on the situation the algebra can be a bit messy. Some additional examples are explored in the exercises at the end of this section.

A Mix of Dirichlet and Insulating Boundary Conditions

Consider the problem of solving the heat equation (8.17) on an interval $0 \leq x \leq L$ with initial data $u(x,0) = f(x)$ and mixed boundary conditions, $u(0,t) = 0$ at $x = 0$ and an insulating boundary condition $\frac{\partial u}{\partial x}(L,t) = 0$ on the right at $x = L$. To attack this problem let us return to separation of variables and seek solutions to the heat equation of the form $u(x,t) = T(t)X(x)$, as in Section

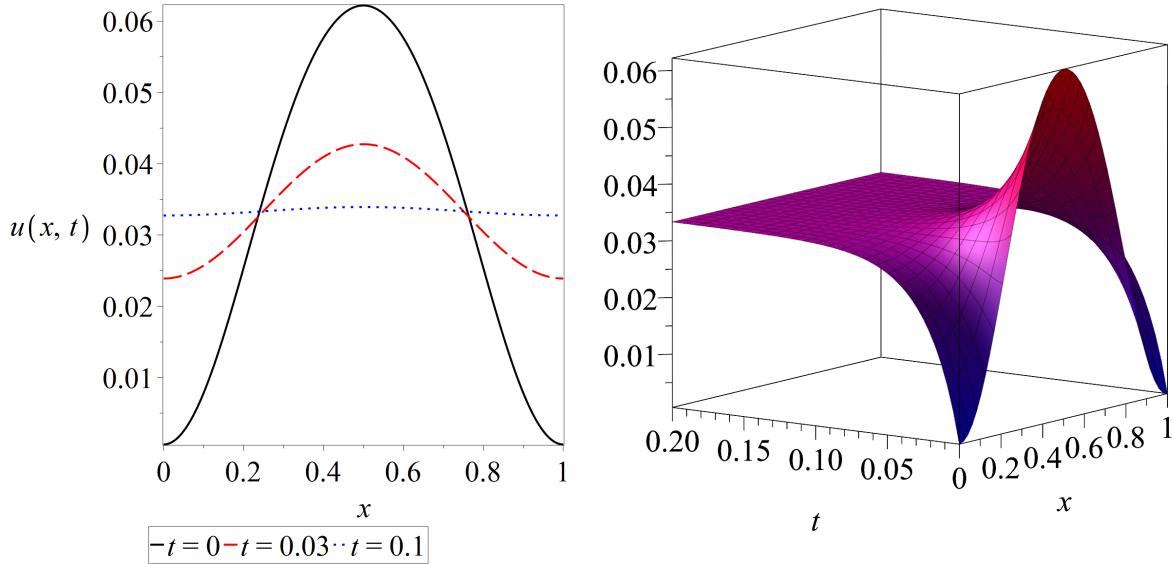


Figure 8.15: Solution $u(x, t)$ to heat equation with diffusivity $\alpha = 1$ on interval $0 \leq x \leq 1$, boundary conditions $\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(1, t) = 0$ and initial data $u(x, 0) = x^2(1-x)^2$. Solution u at time $t = 0$ shown as solid black graph, $t = 0.03$ as dashed red graph, $t = 0.1$ as dotted blue graph. Right panel: same solution as a surface plot.

8.1.5. Precisely the same analysis that led from (8.23) to (8.29) again shows that we should take $T(t) = Ce^{-\alpha\lambda^2 t}$ and $X(x) = c_1 \sin(\lambda x) + c_2 \cos(\lambda x)$ for constants C, c_1, c_2 , and λ . This again produces solutions

$$u(x, t) = C_1 e^{-\alpha\lambda^2 t} \sin(\lambda x) + C_2 e^{-\alpha\lambda^2 t} \cos(\lambda x)$$

to the heat equation.

As in the case of homogeneous Dirichlet conditions, the requirement $u(0, t) = 0$ for all t forces $C_2 = 0$, and so u must be of the form

$$u(x, t) = C_1 e^{-\alpha\lambda^2 t} \sin(\lambda x). \quad (8.69)$$

The condition that $\frac{\partial u}{\partial x}(L, t) = 0$ now becomes $C_1 \lambda e^{-\alpha\lambda^2 t} \cos(\lambda L) = 0$. Assume that $C_1 \neq 0$ or else u is identically zero and of no value. Also, $e^{-\alpha\lambda^2 t}$ is never zero, so we need $\lambda \cos(\lambda L) = 0$. But $\lambda = 0$ also yields a solution $u(x, t)$ in (8.69) that is identically zero, so let us instead require that

$$\cos(\lambda L) = 0. \quad (8.70)$$

Since $\cos(z) = 0$ precisely when $z = \pi/2 + j\pi$ for some integer j , from (8.70) we conclude that

$$\lambda = \frac{(j+1/2)\pi}{L}$$

for some integer j . Using this in (8.69) yields an infinite family of functions

$$u(x, t) = e^{-\alpha(j+1/2)^2\pi^2 t/L^2} \sin\left(\frac{\pi(j+1/2)x}{L}\right),$$

one for each integer j , and each such function satisfies the heat equation with $u(0, t) = 0$ and $\frac{\partial u}{\partial x}(L, t) = 0$.

Following our previous procedure, let us consider linear combinations of these functions; such superpositions will satisfy both the heat equation and boundary conditions. It's also not hard to see that we may as well limit our attention to $j \geq 0$ (j and $-j - 1$ yield the same function, up to a minus sign). So consider a finite linear combination of the form

$$u(x, t) = \sum_{j=0}^n c_j e^{-\alpha(j+1/2)^2 \pi^2 t / L^2} \sin\left(\frac{\pi(j+1/2)x}{L}\right). \quad (8.71)$$

This function has initial data

$$u(x, 0) = \sum_{j=1}^n c_j \sin\left(\frac{\pi(j+1/2)x}{L}\right). \quad (8.72)$$

If we can choose the c_j so the right side of (8.72) converges to $f(x)$ as $n \rightarrow \infty$ we will have produced the desired solution (again, as an infinite sum).

Let S denote the set of functions of the form $\sin(\pi(j+1/2)x/L)$ corresponding to the choices $j = 0, 1, 2, \dots$, so

$$S = \left\{ \sin\left(\frac{\pi x/2}{L}\right), \sin\left(\frac{3\pi x/2}{L}\right), \sin\left(\frac{5\pi x/2}{L}\right), \dots \right\}.$$

Here are two convenient facts:

1. The set S is orthogonal, that is

$$\int_0^L \sin\left(\frac{\pi(j+1/2)x}{L}\right) \sin\left(\frac{\pi(k+1/2)x}{L}\right) dx = 0. \quad (8.73)$$

when $j \neq k$.

2. If

$$s_n(x) = \sum_{j=0}^n c_j \sin\left(\frac{\pi(j+1/2)x}{L}\right)$$

then the choice for c_j that minimizes $\|f - s_n\|_2$ is given by

$$c_j = \frac{\int_0^L f(x) \sin\left(\frac{\pi(j+1/2)x}{L}\right) dx}{\int_0^L \sin^2\left(\frac{\pi(j+1/2)x}{L}\right) dx}. \quad (8.74)$$

3. The set S is complete, that is, if the c_j are chosen according to (8.74) then $\|f - s_n\|_2 \rightarrow 0$ as $n \rightarrow \infty$.

The proof that (8.73) holds is a simple computation; see Reading Exercise 8.3.3. The proof that (8.74) is correct follows exactly the same path that led from (8.44) to (8.46), without the a_0 term. For a proof that S is complete, see [105].

Based on this analysis, we can express the solution to the heat equation with initial data $u(x, 0) = f(x)$, Dirichlet condition $u(0, t) = 0$ at $x = 0$, and an insulating boundary condition $\frac{\partial u}{\partial x}(L, t) = 0$ on the right at $x = L$ as

$$u(x, t) = \sum_{j=0}^{\infty} c_j e^{-\alpha(j+1/2)^2 \pi^2 t / L^2} \sin\left(\frac{\pi(j+1/2)x}{L}\right) \quad (8.75)$$

with the c_j given by (8.74).

■ **Example 8.11** Suppose $u(x, t)$ satisfies the heat equation with diffusivity $\alpha = 5$ on the interval $0 \leq x \leq 3$ with boundary data $u(0, t) = 0$ at $x = 0$ and an insulating boundary condition $\frac{\partial u}{\partial x}(3, t) = 0$ on the right at $x = 3$. The initial data is $u(x, 0) = f(x)$ with $f(x) = 6x - x^2$. The solution is given by (8.75) with c_j given by (8.74). We compute

$$\begin{aligned} c_j &= \frac{\int_0^3 (6x - x^2) \sin\left(\frac{\pi(j+1/2)x}{3}\right) dx}{\int_0^3 \sin^2\left(\frac{\pi(j+1/2)x}{3}\right) dx} \\ &= \frac{288}{(2k+1)^3 \pi^3}. \end{aligned}$$

As usual, a computer algebra system is helpful. The first few terms of the solution are

$$u(x, t) = \frac{288}{\pi^3} e^{-5\pi^2 t/36} \sin(\pi x/6) + \frac{32}{3\pi^3} e^{-5\pi^2 t/4} \sin(\pi x/2) + \frac{288}{125\pi^3} e^{-125\pi^2 t/36} \sin(5\pi x/6) + \dots$$

A graph of the solution at times $t = 0, 0.2$, and 1 is shown in Figure 8.16. ■

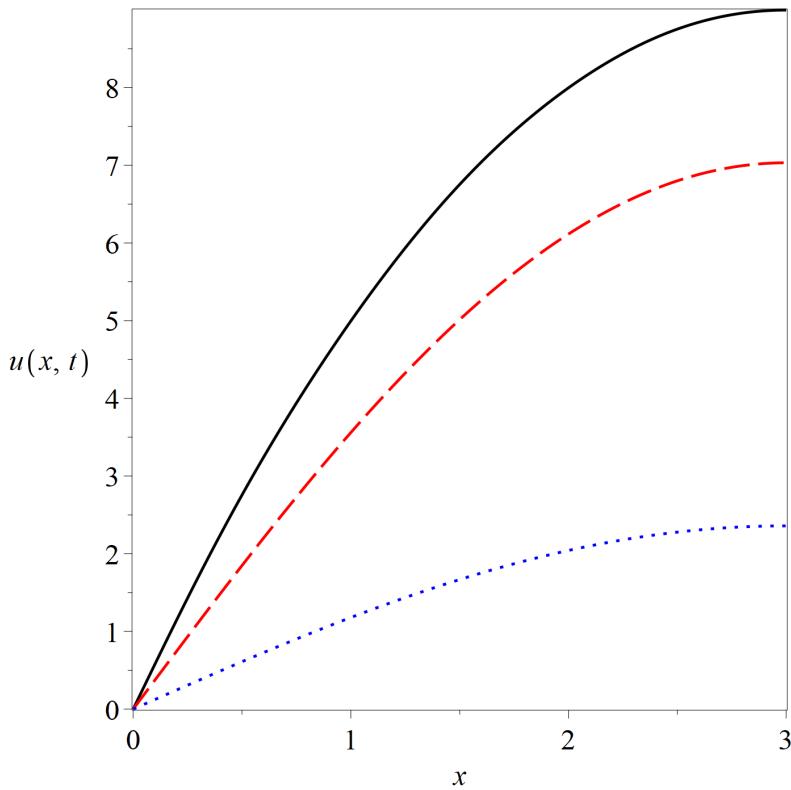


Figure 8.16: Solution $u(x, t)$ to heat equation with diffusivity $\alpha = 5$ on interval $0 \leq x \leq 3$, boundary conditions $u(0, t) = 0$ and $\frac{\partial u}{\partial x}(3, t) = 0$ and initial data $u(x, 0) = 6x - x^2$. Solution u at time $t = 0$ shown as solid black graph, $t = 0.2$ as dashed red graph, $t = 1$ as dotted blue graph.

Reading Exercise 8.3.3 Verify (8.73). Also, compute the value of

$$\int_0^L \sin^2\left(\frac{\pi(j+1/2)x}{L}\right) dx. \quad (8.76)$$

Hint: Use the identity $\sin(y)\sin(z) = (\cos(y-z) - \cos(y+z))/2$ for (8.73) and $\sin^2(y) = 1/2 - \cos(2y)/2$ may be helpful for (8.76). The value of the integral in (8.76) does not depend on j , but does depend on L .

Reading Exercise 8.3.4 Use (8.74) to compute

$$s_n(x) = \sum_{j=1}^n c_j \sin\left(\frac{\pi(j+1/2)x}{L}\right).$$

for the function $f(x) = x$ with $n = 0, 5, 10$ on the interval $0 \leq x \leq 1$. In each case plot $s_n(x)$.

Nonhomogeneous Boundary Conditions

Consider the problem of solving the heat equation (8.17) on $0 \leq x \leq L$ with an initial condition $u(x, 0) = f(x)$ and nonhomogeneous Dirichlet boundary conditions of the form $u(0, t) = u_0(t)$ and $u(L, t) = u_L(t)$, where $u_0(t)$ and $u_L(t)$ are specified functions of time. This can be done using a variation of the techniques we already seen, but let's focus for now on the easiest case in which $u_0(t)$ and $u_L(t)$ are simply constants, $u_0(t) = u_0$ and $u_L(t) = u_L$.

The technique is essentially the method of undetermined coefficients from Section 4.3.2. Specifically, we begin by guessing a particular solution $u_p(x, t)$ to the heat equation that also satisfies the Dirichlet boundary data. In this case the task is easy: take $u_p(x, t)$ to be independent of t and linear in x , as

$$u_p(x, t) = u_0 + \frac{(u_L - u_0)x}{L}.$$

The function u_p is, as a function of x , a straight line that interpolates the Dirichlet data at $x = 0$ and $x = L$. This function does not depend on t and so is a steady-state solution to the heat equation. We encountered these in Reading Exercise 8.1.8.

Next consider the function $w(x, t) = u(x, t) - u_p(x, t)$ where u is the desired solution to the heat equation with boundary data $u_0(t) = u_0$ and $u_L(t) = u_L$, and with initial data $u(x, 0) = f(x)$. Linearity shows that w is a solution to the heat equation. Moreover, w has boundary data $w(0, t) = u(0, t) - u_p(0, t) = u_0 - u_0 = 0$, and $w(L, t) = u(L, t) - u_p(L, t) = u_L - u_L = 0$. That is, $w(x, t)$ is a solution to the heat equation with homogeneous Dirichlet data. The function w has initial data $w(x, 0) = \tilde{f}(x)$ where

$$\tilde{f}(x) = u(x, 0) - u_p(x, 0) = f(x) - u_0 - \frac{u_L - u_0}{L}x.$$

We know how to find $w(x, t)$ using a Fourier sine series, just as in Section 8.3.1. We compute $w(x, t)$ accordingly and then form $u(x, t) = w(x, t) + u_p(x, t)$, the desired solution with nonhomogeneous (but constant) Dirichlet data.

■ **Example 8.12** Consider a bar of length $L = 1$ with diffusivity $\alpha = 3$. Let's find a solution $u(x, t)$ to the heat equation with Dirichlet data $u(0, t) = 20$, $u(1, t) = 80$, and initial data $f(x) = 20 + 60x^3$. Note that $f(x)$ satisfies $f(0) = 20$ and $f(1) = 80$, and is compatible with the boundary conditions.

First form the steady-state solution $u_p(x, t) = 20 + 60x$, which satisfies the heat equation as well as the Dirichlet boundary conditions $u_p(0, t) = 20$ and $u_p(1, t) = 80$. Then $w(x, t) = u(x, t) - u_p(x, t)$ satisfies the heat equation and has initial data $w(x, 0) = \tilde{f}(x)$ where

$$\tilde{f}(x) = f(x) - (20 + 60x) = 60x^3 - 60x.$$

The boundary data for w is $w(0, t) = 20 - 20 = 0$ and $w(1, t) = 60 - 60 = 0$. The function $w(x, t)$ can be found by first computing the Fourier sine coefficients for $\tilde{f}(x)$ as

$$b_k = 2 \int_0^1 \tilde{f}(x) \sin(k\pi x) dx = \frac{720(-1)^k}{k^3 \pi^3}$$

and then taking

$$w(x, t) = \sum_{k=1}^{\infty} b_k e^{-3k^2\pi^2t} \sin(k\pi x).$$

The solution $u(x, t)$ is then

$$u(x, t) = u_p(x, t) + w(x, t) = 20 + 60x + \sum_{k=1}^{\infty} \frac{720(-1)^k}{k^3\pi^3} e^{-3k^2\pi^2t} \sin(k\pi x).$$

Figure 8.17 shows $u(x, t)$ at times $t = 0, 0.01, 0.05$, and 0.1 . The solution $u(x, t)$ evolves toward the steady-state solution $u_p(x) = 20 + 60x$ as t increases. ■

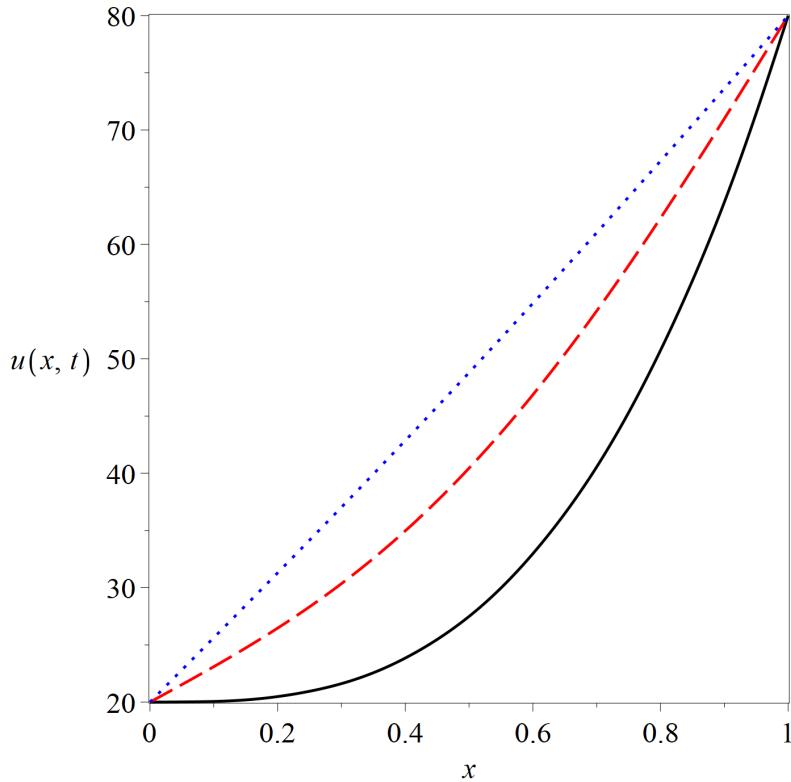


Figure 8.17: Solution $u(x, t)$ to heat equation with diffusivity $\alpha = 3$ on interval $0 \leq x \leq 1$, boundary conditions $u(0, t) = 20$ and $u(1, t) = 80$ with initial data $u(x, 0) = 20 + 60x^3$, at times $t = 0$ (solid black), $t = 0.03$ (dashed red) and $t = 0.1$ (dotted blue).

8.3.4 Diffusion

The heat equation governs more than just the flow of heat through an object. The same equation applies to many situations in which a substance spreads through another medium, for example, a liquid or gas, through the process of **diffusion**. Diffusion is the natural spread of a substance from regions of higher concentration to regions of lower concentration, often through random molecular motion. The heat equation models one instance of this phenomena, in which the substance that diffuses is thermal energy.

To illustrate diffusion more generally, consider a thin pipe of indefinite length, filled with water. Some other substance, perhaps a chemical, diffuses through the water. The water in the pipe is motionless, but provides the medium through which the chemical naturally spreads, from regions

of higher concentration to regions of lower concentration (think of a drop of food coloring put into a container of water). We will quantify the amount of the chemical on a per mass basis. As in Section 8.1 let $\rho(x,t)$ denote the density of the chemical on a mass per length basis and let $q(x,t)$ denote the flux of the chemical from left to right (mass per time). Here the density can also be considered as a concentration. If the mass of the chemical substance is conserved then the continuity equation (8.8) holds. We adopt the constitutive relation $q(x,t) = -\alpha \frac{\partial \rho}{\partial x}$ where α is a constant, as we did for the heat equation. This models the diffusion of the chemical from higher to lower concentrations with a flux proportional to the concentration or density gradient. In the context of diffusion, the constitutive relation $q(x,t) = -\alpha \frac{\partial \rho}{\partial x}$ is called **Fick's law**. Fick's law in conjunction with the continuity equation yields

$$\frac{\partial \rho}{\partial t} - \alpha \frac{\partial^2 \rho}{\partial x^2} = 0 \quad (8.77)$$

which is exactly equation (8.16) from Section 8.1. Here α is called the **diffusivity** of the chemical substance (which depends on other things too, for example, what the chemical is dissolved in). In the present setting things are somewhat simplified compared to the flow of heat, since we will work directly with the density function ρ instead of its proxy, temperature. Of course (8.77) is just the heat equation, though in this setting it is often called the **diffusion equation**.

Source Localization

In the following example we consider using the diffusion equation to model a pollutant source in a body of water. How does the pollutant spread out over time?

■ **Example 8.13** A common problem in applied mathematics is that of **source localization**. The term is quite broad, but in general refers to the problem of locating and characterizing the source of some stuff, in both time and space, using data that may be obtained far from the source itself.

To illustrate, suppose some amount of a pollutant is dumped into a body of water at time $t = 0$ at some location. We will model the spread of this pollutant as a function of time and space, under the assumption that the pollutant diffuses through the water according to Fick's law and that there is no current in the water. An accurate model of this situation might allow us to tackle the problem of using concentration data measured at various locations in the body of water to determine where and when the pollutant was dumped, and thus perhaps nab the culprit responsible for illegally discharging toxic material into a waterway. In this example we will set up a model for how the pollutant spreads. In the Project “Finding Polluters” in Section 8.5.2 you can consider how this model might be used to determine the pollutant source location from data collected away from the pollution source.

Let us focus on a simple version of the problem that illustrates the essential mathematics. The body of water will be one-dimensional, modeled as a thin pipe filled with water with both ends sealed. Assume there is no current, that is, the water itself is motionless. Suppose the pipe spans $x = 0$ to $x = 1$ in a horizontal direction. An amount A mass units of some chemical is introduced into the pipe at time $t = 0$ at a point $x = x_0$ where $0 < x_0 < 1$. Suppose this chemical substance is conserved, so that the density or concentration of the chemical for positive times t obeys the diffusion equation (8.77). The fact that the ends of the pipe are sealed yields insulating boundary conditions $q(0,t) = q(1,t) = 0$, or equivalently, $\frac{\partial \rho}{\partial x}(0,t) = \frac{\partial \rho}{\partial x}(1,t) = 0$. Let us take $\alpha = 1$ for simplicity.

A Point Initial Source

What is the appropriate initial condition? If A mass units of the substance are introduced into the pipe at a single point $x = x_0$ at time $t = 0$ then based on the modeling of Chapter 5 and Section 5.4, at time $t = 0$ the density ρ can be modeled as

$$f(x) = A\delta(x - x_0) \quad (8.78)$$

where δ denotes the Dirac delta function. This choice for $f(x)$ does not fit the mold of a piecewise continuous function for the initial condition in the diffusion equation— f here isn't even a function—but let's press on and see what happens when we try to solve the diffusion equation.

Let's use $A = 1$ and $x_0 = 1/2$, so the initial condition (8.78) is $f(x) = \delta(x - 1/2)$. The solution procedure for the diffusion (or heat) equation here is to compute a Fourier cosine expansion of $f(x)$, which yields coefficients

$$a_k = 2 \int_0^1 \delta(x - 1/2) \cos(k\pi x) dx.$$

Based on the rules for handling delta functions, in particular (5.69), it follows that

$$a_k = 2 \cos(k\pi/2). \quad (8.79)$$

Thus $a_0 = 2, a_1 = 0, a_2 = -2, a_3 = 0, a_4 = 2$, and so on. In general $a_k = 2(-1)^{k/2}$ when k is even and $a_k = 0$ when k is odd.

Before writing out the solution to the heat equation, let's look at the corresponding cosine expansion for $\delta(x - 1/2)$, given by

$$\begin{aligned} s_n(x) &= \frac{a_0}{2} + \sum_{k=1}^n a_k \cos(k\pi x) \\ &= 1 + 2 \sum_{k=2, \text{ even}}^n (-1)^{k/2} \cos(k\pi x) \end{aligned} \quad (8.80)$$

to n terms. Plots of $s_n(x)$ on the interval $0 \leq x \leq 1$ for $n = 0, 2, 10$, and 50 are shown in Figure 8.18. Although $\delta(x - 1/2)$ does not fit into the framework in which we've developed Fourier series, it does seem like the cosine approximation s_n is doing the right thing as $n \rightarrow \infty$. It can be shown in

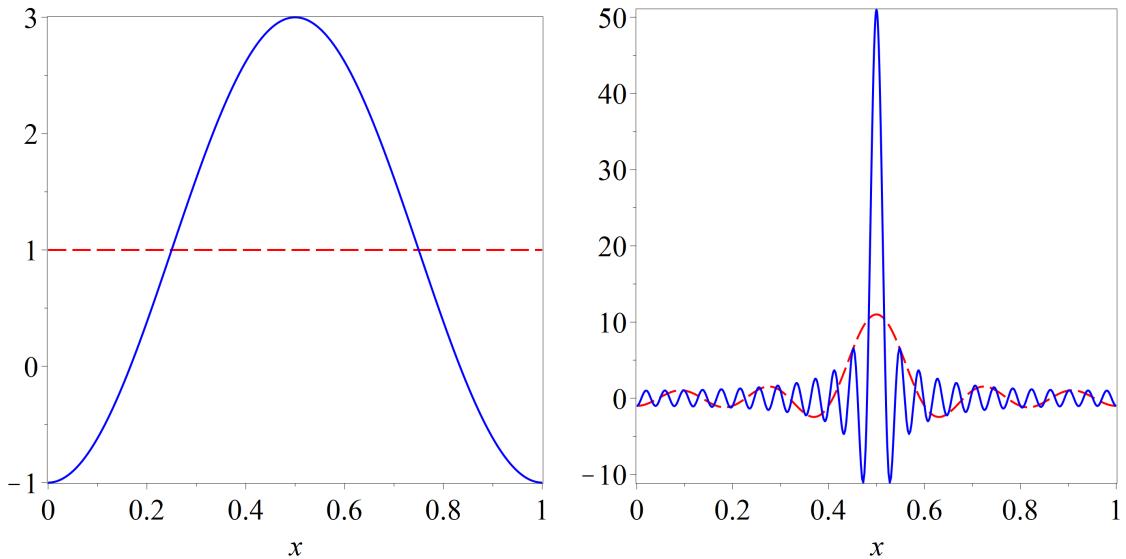


Figure 8.18: Left panel: Fourier cosine approximations $s_0(x)$ (dashed red) and $s_2(x)$ (solid blue) to Dirac delta function $\delta(x - 1/2)$. Right panel: Fourier cosine approximations $s_{10}(x)$ (dashed red) and $s_{50}(x)$ (solid blue) to Dirac delta function $\delta(x - 1/2)$.

a precise sense that when used in the solution procedure for the heat equation, the Fourier series coefficients (8.80) yield a solution that is consistent with the Dirac delta function initial condition, and the results provide an accurate depiction of how the concentration of the pollutant evolves over time.

Reading Exercise 8.3.5 Show that with $s_n(x)$ as defined by (8.80), $s_n(1/2) = n + 1$ when n is even and $s_n(1/2) = n$ when n is odd. What is $\lim_{n \rightarrow \infty} s_n(1/2)$ equal to, and why does this make intuitive sense?

Solution to the Diffusion Equation With an Initial Point Source

Let us continue and write out the solution to the diffusion equation with a point source initial condition. With the insulating boundary conditions the cosine expansion (8.67) in conjunction with (8.79) yields

$$\begin{aligned}\rho(x, t) &= 1 + 2 \sum_{k=2, \text{even}}^{\infty} (-1)^{k/2} e^{-\alpha k \pi^2 t / L^2} \cos(k \pi x / L) \\ &= 1 - 2e^{-4\pi^2 t} \cos(2\pi x) + 2e^{-16\pi^2 t} \cos(4\pi x) - 2e^{-36\pi^2 t} \cos(6\pi x) + \dots\end{aligned}\quad (8.81)$$

where recall that $\alpha = 1$. Figure 8.19 shows the solution $\rho(x, t)$ in (8.81) graphed as a function of x on the interval $0 \leq x \leq 1$ at times $t = 0.01, 0.03$, and 0.1 . The initial mass of pollutant deposited at $x = 1/2$ begins to diffuse over time, eventually reaching a uniform concentration throughout the pipe, which makes intuitive sense. ■

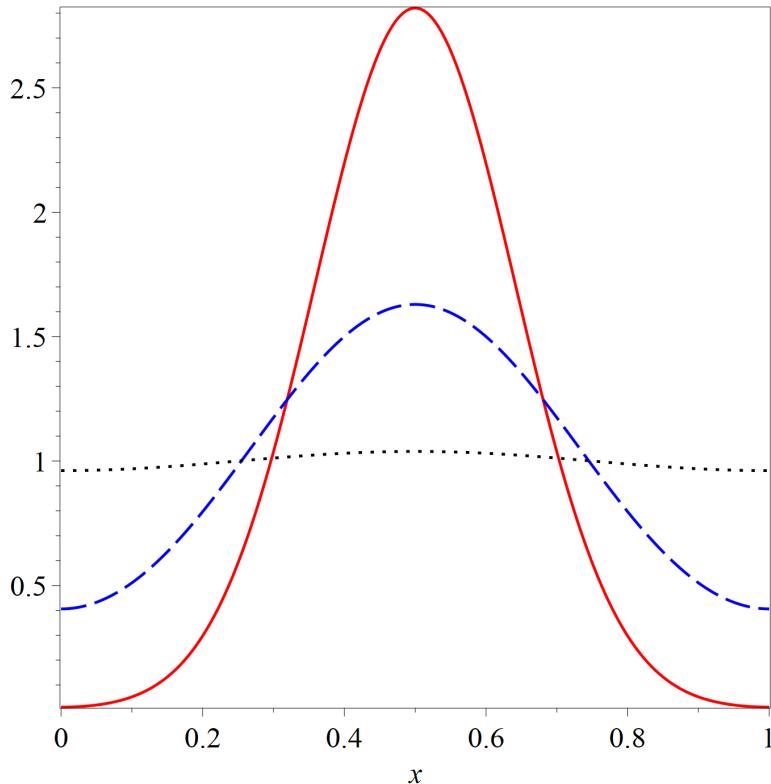


Figure 8.19: Solution to diffusion equation with insulating boundary conditions and initial data $\rho(x, 0) = \delta(x - 1/2)$ at times $t = 0.01$ (solid red), $t = 0.03$ (dashed blue), and $t = 0.1$ (dotted black).

Forward and Inverse Problems

The process of solving the diffusion equation with specified boundary and initial conditions to find $\rho(x, t)$ as we've been doing is called a **forward problem** or a **direct problem**. These terms refer to the usual thing we do with a differential equation, namely find the solution. In the project “Finding Polluters” in Section 8.5.2 you can explore the following twist: consider an initial condition

$\rho(x, 0) = A\delta(x - x_0)$ with source location x_0 and initial pollutant amount A as unknowns. The task is to determine A and x_0 from data such as $\rho(0, t)$ for $t > 0$ (measurements of the pollutant concentration at the location $x = 0$, remote from the source). This is an example of an **inverse problem**—given information about the solution to a differential equation we seek to reverse engineer something about the initial conditions, or the boundary conditions, or perhaps parameters that appear in the DE. It's really just another version of the parameter estimation problems we've encountered throughout the text.

8.3.5 Solving the Nonhomogeneous Heat or Diffusion Equation

Adding Creation and Destruction of Stuff

Consider what the pollutant model of Example 8.13 would look like if the pollutant was not a single instantaneous discharge at time $t = 0$ at a point location $x = x_0$, but instead occurred over time and possibly over a distributed portion of space. Specifically, suppose the pollutant is introduced into the region $0 < x < L$ at a rate of $r(x, t)$ mass units per length of the conduit per time unit. In this case stuff (the pollutant) is not conserved inside the bar, but is being created or introduced, so the usual continuity equation will not hold. In Exercise 8.1.7 you were asked to modify the continuity equation for a situation such as this and demonstrate that (8.38) holds, which is reproduced here

$$\frac{\partial \rho}{\partial t} + \frac{\partial q}{\partial x} = r(x, t). \quad (8.82)$$

Reading Exercise 8.3.6 Verify that each term in (8.82) has the dimension of mass per length per time, if the stuff is measured in units of mass.

Combining Fick's law $q(x, t) = -\alpha \frac{\partial \rho}{\partial x}$ with (8.82) yields

$$\frac{\partial \rho}{\partial t} - \alpha \frac{\partial^2 \rho}{\partial x^2} = r(x, t). \quad (8.83)$$

This is the **nonhomogeneous heat equation** or **nonhomogeneous diffusion equation**. The function $r(x, t)$ is often called a **source term** and represents the local rate at which stuff is being created (if $r > 0$) or destroyed (if $r < 0$) at position x and time t , on a stuff per length per time basis. Our goal is to develop a procedure for solving (8.83) for the function $\rho(x, t)$ on an interval $0 \leq x \leq L$, for a given diffusivity α , right hand side $r(x, t)$, initial condition $\rho(x, 0) = f(x)$, and boundary conditions at $x = 0$ and $x = L$.

Solving the Nonhomogeneous Diffusion Equation

To illustrate how this can be done, let's use insulating boundary conditions $\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(L, t) = 0$; homogenous Dirichlet boundary conditions can be handled in a similar manner, as can other boundary conditions. We will solve (8.83) by using an expansion similar to (8.67), but in the form

$$\rho(x, t) = \frac{\phi_0(t)}{2} + \sum_{k=1}^{\infty} \phi_k(t) \cos(k\pi x/L) \quad (8.84)$$

where the functions $\phi_k(t)$ are to be determined. The homogeneous case of the heat equation in which $r(x, t) = 0$ resulted in $\phi_k(t) = a_k e^{-\alpha k^2 \pi^2 t}$ where the a_k were the Fourier cosine coefficients of the initial data, but that will no longer be the case. Note that each summand $\phi_k(t) \cos(k\pi x/L)$ on the right in (8.84) satisfies the insulating boundary conditions $\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(L, t) = 0$, so that's built right into the process. To obtain the proper initial condition $\rho(x, 0) = f(x)$, first compute the Fourier cosine coefficients f_k of $f(x)$, as

$$f_k = \frac{2}{L} \int_0^L f(x) \cos(k\pi x/L) dx \quad (8.85)$$

and then require

$$\phi_k(0) = f_k \quad (8.86)$$

for $k = 0, 1, 2, \dots$. In (8.84) this will yield the initial data we want, since then

$$\rho(x, 0) = f_0/2 + \sum_{k=1}^{\infty} f_k \cos(k\pi x/L) = f(x).$$

With these choices $\rho(x, t)$ in (8.84) will have the correct boundary and initial data, but the functions $\phi_k(t)$ must also be chosen so that (8.83) holds for $t > 0$, and this depends on $r(x, t)$. To do this begin with a Fourier cosine expansion of the function $r(x, t)$ with respect to x . Specifically, for each fixed t think of $r(x, t)$ as a function of x on the interval $0 \leq x \leq L$ and expand $r(x, t)$ into a cosine series with respect to x as

$$\rho(x, t) = \frac{a_0(t)}{2} + \sum_{k=1}^{\infty} a_k(t) \cos(k\pi x/L) \quad (8.87)$$

by taking

$$a_k(t) = \frac{2}{L} \int_0^L r(x, t) \cos(k\pi x/L) dx. \quad (8.88)$$

The coefficients a_k are functions of t because $r(x, t)$ depends on t . Using the expansion on the right in (8.87) for $r(x, t)$ and the ansatz (8.84) for $\rho(x, t)$ in the nonhomogeneous heat equation (8.83) yields (assume we can differentiate the series for $\rho(x, t)$ term by term with respect to t and x)

$$\sum_{k=0}^{\infty} (\phi'_k(t) + \alpha k^2 \pi^2 \phi_k(t)/L^2) \cos(k\pi x/L) = \frac{a_0(t)}{2} + \sum_{k=1}^{\infty} a_k(t) \cos(k\pi x/L). \quad (8.89)$$

Our task is now simple: if $\rho(x, t)$ is to be a solution to (8.83) then the left and right sides of (8.89) must match, which forces

$$\begin{aligned} \phi'_0(t) &= a_0(t), \\ \phi'_k(t) + \frac{\alpha k^2 \pi^2}{L^2} \phi_k(t) &= a_k(t), \quad k = 1, 2, 3, \dots \end{aligned} \quad (8.90)$$

where $a_k(t)$ is given by (8.88). In summary, we must solve a first-order linear ODE for each function $\phi_k(t)$; recall from (8.86) that the initial condition for each k is $\phi_k(0) = f_k$ where f_k is the Fourier cosine coefficient of $f(x)$, given by (8.85). Solving the nonhomogeneous heat equation has been turned into the problem of solving an infinite number of first-order ODEs.

The solution for $\phi_0(t)$ in (8.90) is obtained by direct integration,

$$\phi_0(t) = f_0 + \int_0^t a_0(z) dz$$

and the solution for $\phi_k(t)$ for each $k \geq 1$ can be obtained from the integrating factor approach with integrating factor $e^{\alpha k^2 \pi^2 t}$ in (8.90). After computing $\phi_k(t)$ for each k , use (8.84) to write out the solution to the nonhomogeneous diffusion or heat equation. In practice we might only compute a finite number of terms to produce an acceptable approximation.

Example 8.14 Consider a bar of length $L = 2$ with diffusivity $\alpha = 1$, insulating boundary conditions $\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(2, t) = 0$, initial data $f(x) = x^2(2-x)^2$, and nonhomogeneous term $r(x, t) = xe^{-t}$. If we take only terms up to $k = 3$ in the sum of (8.84) then the relevant coefficients f_k are

$$f_0 = 16/15, \quad f_1 = 0, \quad f_2 = -48/\pi^4, \quad f_3 = 0$$

and

$$a_0(t) = 2e^{-t}, \quad a_1(t) = -8e^{-t}/\pi^2, \quad a_2(t) = 0, \quad a_3(t) = -8e^{-t}/(9\pi^2).$$

From (8.90) with initial data $\phi_k(0) = f_k$ for $k = 0, 1, 2, 3$ we obtain

$$\begin{aligned}\phi_0(t) &= \frac{46}{15} - 2e^{-t}, \quad \phi_1(t) \approx 0.5524(e^{-2.467t} - e^{-t}), \\ \phi_2(t) &= -0.4928e^{-9.87t}, \quad \phi_3(t) = 0.0042(e^{-22.21t} - e^{-t}).\end{aligned}$$

From (8.84) an approximate solution is given by

$$\rho(x, t) \approx \phi_0(t)/2 + \phi_1(t) \cos(\pi x/2) + \phi_2(t) \cos(\pi x) + \phi_3(t) \cos(3\pi x/2).$$

A graph of $\rho(x, t)$ for $0 \leq x \leq 2$ at times $t = 0, t = 0.5, t = 2.0$, and $t = 5$ is shown in Figure 8.20. At this time the source term $r(x, t) = xe^{-t}$ has substantially decayed to zero, so little additional pollutant mass would be added after that time. ■

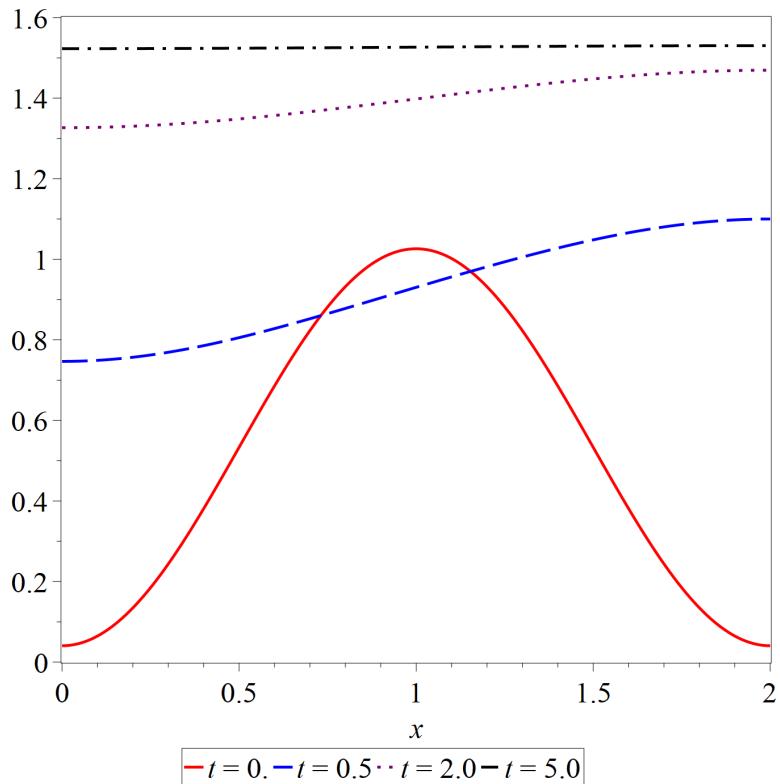


Figure 8.20: Solution to diffusion equation with insulating boundary conditions and initial data $\rho(x, 0) = x^2(2-x)^2$ at times $t = 0.0$ (solid red), $t = 0.5$ (dashed blue), $t = 2.0$ (dotted purple), and time $t = 5$ (dash-dot black).

Reading Exercise 8.3.7 Why should we expect that in the setting of Example 8.14 with insulating boundary conditions, the total amount of stuff in the bar at time $t = T$ should be the total amount present at time $t = 0$ plus the total amount introduced by the source term $r(x, t)$? Argue that these amounts are

$$\int_0^L f(x) dx \text{ and } \int_0^T \int_0^L r(x, t) dx dt$$

respectively. Compute both integrals above for Example 8.14 using $T = 5$. Use this to explain the graph of $\rho(x,t)$ at $t = 5$ in Figure 8.20.

See Exercise 8.3.5 for a guided rigorous proof of the conclusion in Reading Exercise 8.3.7.

Conclusion

In this section we've shown how to solve the heat or diffusion equation with a variety of different boundary conditions, including the nonhomogeneous version of the equation in which a source term is present. Many possible combinations of boundary conditions with and without source terms exist that have not been examined, but with a bit of practice and by utilizing the linearity of the heat equation, solutions to these problems can be pieced together. Some opportunities to try this are outlined in the exercises.

8.3.6 Exercises

Exercise 8.3.1 Solve each equation $\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = 0$ below for $u(x,t)$ on the indicated interval $0 \leq x \leq L$ with the given diffusivity α and initial condition $u(x,0) = f(x)$. Use homogeneous Dirichlet boundary conditions. Carry out the series solution to the indicated number of terms N (from b_1 to b_N in the cosine expansion of $f(x)$) and plot the solution at the given times.

- (a) $f(x) = x^2(2-x)^2$ on $0 \leq x \leq 2$, $\alpha = 5$, $N = 3$, plot at $t = 0, 0.05$, and 0.1 .
- (b) $f(x) = x\sin(\pi x)$ on $0 \leq x \leq 2$, $\alpha = 1$, $N = 3$, plot at $t = 0, 0.1$, and 0.2 .
- (c) $f(x) = x\cos(3\pi x/8)$ on $0 \leq x \leq 4$, $\alpha = 3$, $N = 4$, plot at $t = 0, 0.1$, and 0.2 .
- (d) $f(x) = x - x^5$ on $0 \leq x \leq 1$, $\alpha = 1$, $N = 5$, plot at $t = 0, 0.1$, and 0.2 . Then change α to 2 and redo the plots. How does the change in α affect things?

Exercise 8.3.2 Solve each equation $\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = 0$ below for $u(x,t)$ on the indicated interval $0 \leq x \leq L$ with the given diffusivity α and initial condition $u(x,0) = f(x)$. Use insulating boundary conditions. Carry out the series solution to the indicated number of terms N (from a_0 to a_N in the cosine expansion of $f(x)$) and plot the solution at the given times.

- (a) $f(x) = x^2(1-x)^2$ on $0 \leq x \leq 1$, $\alpha = 1$, $N = 3$, plot at $t = 0, 0.03$ and 0.1 .
- (b) $f(x) = x\sin^2(\pi x)$ on $0 \leq x \leq 2$, $\alpha = 1$, $N = 5$, plot at $t = 0, 0.05$, and 0.5 .
- (c) $f(x) = x\sin^2(\pi x/4)$ on $0 \leq x \leq 4$, $\alpha = 3$, $N = 3$, plot at $t = 0, 0.2$, and 0.5 .
- (d) $f(x) = x^2 - x^4/2$ on $0 \leq x \leq 1$, $\alpha = 1$, $N = 2$, plot at $t = 0, 0.05$, and 0.2 . Then change α to 3 and redo the plots. How does the change in α affect things?

Exercise 8.3.3 Example 8.11 may be a useful template. Solve each equation $\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = 0$ below for $u(x,t)$ on the indicated interval $0 \leq x \leq L$ with the given diffusivity α and initial condition $u(x,0) = f(x)$. Use a homogeneous Dirichlet condition on the left at $x = 0$ and an insulating boundary condition at $x = L$. Carry out the series solution to the indicated number of terms N (from c_0 to c_N in (8.71)) and plot the solution at the given times.

- (a) $f(x) = x^2(2-x)^2$ on $0 \leq x \leq 2$, $\alpha = 5$, $N = 3$, plot at $t = 0, 0.1$, and 0.5 .
- (b) $f(x) = x\sin^2(\pi x/4)$ on $0 \leq x \leq 4$, $\alpha = 3$, $N = 3$, plot at $t = 0, 0.2$, and 2 .

Exercise 8.3.4 Adapt the procedure of Example 8.11 and preceding paragraphs to find the general solution to $\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = 0$ on an interval $0 \leq x \leq L$ with initial data $u(x, 0) = f(x)$, insulating boundary condition $\frac{\partial u}{\partial x}(0, t) = 0$, and homogeneous Dirichlet boundary condition $u(L, t) = 0$. Use your general solution to express the solution with $\alpha = 1, L = 2$ and $f(x) = x^2(2-x)$ to at least 3 terms.

Exercise 8.3.5 Consider a solution $\rho(x, t)$ to the nonhomogeneous diffusion or heat equation (8.83) with initial data $\rho(x, 0) = f(x)$ and insulating boundary conditions.

- (a) Show that for any time $T > t$ we have

$$\int_0^L \rho(x, T) dx = \int_0^L f(x) dx + \int_0^T \int_0^L r(x, t) dx dt. \quad (8.91)$$

Hint: Replace $r(x, t)$ with $\frac{\partial \rho}{\partial t} - \alpha \frac{\partial^2 \rho}{\partial x^2}$ in the second integral on the right in (8.91). Then show that

$$\int_0^T \int_0^L \frac{\partial \rho}{\partial t} dx dt = \int_0^L (\rho(x, T) - f(x)) dx$$

and

$$\int_0^T \int_0^L \frac{\partial^2 \rho}{\partial x^2} dx dt = \int_0^T \left(\frac{\partial \rho}{\partial x}(L, t) - \frac{\partial \rho}{\partial x}(0, t) \right) dt = 0.$$

Conclude that

$$\int_0^T \int_0^L r(x, t) dx dt = \int_0^L (\rho(x, T) - f(x)) dx.$$

- (b) Argue that the left side of (8.91) is the total amount of stuff in the conduit for $0 \leq x \leq L$ at time $t = T$. Argue that the right side of (8.91) is the amount of stuff at time $t = 0$ plus all the stuff that was created (or destroyed) in the conduit from time $t = 0$ to time $t = T$. Why does the resulting equality make sense?

Exercise 8.3.6 Solve each equation $\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = r(x, t)$ below for $u(x, t)$ on the indicated interval $0 \leq x \leq L$ with the given diffusivity α and initial condition $u(x, 0) = f(x)$ using the method of Section 8.3.5. Use insulating boundary conditions in each case. Carry out the series solution to the indicated number of terms N (from $\phi_0(t)$ to $\phi_N(t)$) and plot the solution at the given times.

- (a) $f(x) = 0$ on $0 \leq x \leq 2$, $\alpha = 1$, $N = 3$, and $r(x, t) = xe^{-t}$. Plot at $t = 0, 0.5, 2$, and 5 .
- (b) $f(x) = 0$ on $0 \leq x \leq 2$, $\alpha = 3$, $N = 2$, and $r(x, t) = 1$. Plot at $t = 0, 0.1, 0.3$, and 1 . Why does this make sense?
- (c) $f(x) = x \sin^2(\pi x/4)$ on $0 \leq x \leq 4$, $\alpha = 3$, $N = 5$, and $r(x, t) = x - 2$. Plot at $t = 0, 0.2, 1$, and 3 .
- (d) $f(x) = 1$ on $0 \leq x \leq 4$, $\alpha = 3$, $N = 3$, and $r(x, t) = x \sin(\pi t)$. Plot at $t = 0, 0.5, 1$, and 4 .

Exercise 8.3.7 Suppose one mass unit of a pollutant is discharged into a water-filled conduit spanning the x axis from $x = 0$ to $x = 1$ as in Example 8.13 in Section 8.3.4. The pollutant is discharged at the point $x = x_0$ and the diffusivity is $\alpha = 0.1$. Assume insulating boundary conditions. Write out a 50 term Fourier cosine series solution for $\rho(x, t)$ for the cases in which $x_0 = 1/4$ and $x_0 = 1/3$. Then plot $\rho(0, t)$ in each case on the interval $0 \leq t \leq 5$. How does the pollutant concentration $\rho(0, t)$ at $x = 0$ behave in each case as t increases? In what time interval do the solutions $\rho(0, t)$ differ the most? Repeat at the other end by plotting $\rho(1, t)$ for each case.

Exercise 8.3.8 The techniques of this section are easily adapted to solve the diffusion or heat equation with a variety of boundary conditions. Let's look at how these techniques can be used to solve the heat equation with nonhomogeneous Dirichlet boundary conditions. Let $u(x, t)$ satisfy the homogeneous heat equation (8.17) with diffusivity α on the interval $0 \leq x \leq L$. Suppose the initial condition is $u(x, 0) = f(x)$ for some function $f(x)$ and the Dirichlet boundary conditions are $u(0, t) = u_0(t)$ and $u(L, t) = u_L(t)$ for some functions u_0 and u_L .

In the following you should keep in mind the specific case $\alpha = 2$ on the interval $0 \leq x \leq 4$ with initial data $f(x) = x(4 - x)$ and Dirichlet data $u_0(t) = \sin(t)$ and $u_L(t) = -\sin(2t)$.

- (a) Suppose $w(x, t)$ is any function that satisfies the Dirichlet boundary conditions, that is, $w(0, t) = u_0(t)$ and $w(L, t) = u_L(t)$; note w need not satisfy any other conditions. Let $v(x, t) = u(x, t) - w(x, t)$ (so $u = v + w$). Argue that $v(x, t)$ satisfies the nonhomogeneous heat equation

$$\frac{\partial v}{\partial t} - \alpha \frac{\partial^2 v}{\partial x^2} = r(x, t)$$

where $r(x, t) = -\left(\frac{\partial w}{\partial t} - \alpha \frac{\partial^2 w}{\partial x^2}\right)$, and that v has initial data $v(x, 0) = h(x)$ where $h(x) = f(x) - w(x, 0)$, and that v satisfies homogeneous Dirichlet boundary conditions $v(0, t) = v(L, t) = 0$.

- (b) Verify that the function $w(x, t)$ in part (a) may be taken as

$$w(x, t) = u_0(t) + (u_L(t) - u_0(t))x/L$$

(although this choice for $w(x, t)$ is one of infinitely many choices that will work). Write out $w(x, t)$ explicitly for the specific case above. Write out the nonhomogeneous heat equation (in particular, $r(x, t)$) satisfied by $v(x, t)$, and the initial data $h(x)$.

- (c) The technique of Example 8.14 in Section 8.3.5 for solving the nonhomogeneous heat equation with insulating boundary conditions is easily adapted to handle homogeneous Dirichlet boundary conditions. If the function that satisfies the homogeneous heat equation is $v(x, t)$ then the solution is of the form

$$v(x, t) = \sum_{k=1}^{\infty} \phi_k(t) \sin(k\pi x/L) \tag{8.92}$$

where the functions $\phi_k(t)$ satisfy (8.90) (with the $a_k(t)$ as the Fourier sine coefficients of $r(x, t)$ with respect to x) for $k \geq 1$ with initial data $\phi_k(0) = h_k$ where the h_k are the Fourier sine coefficients for $h(x)$ from part (a). If we compute $v(x, t)$ we may compute $u(x, t)$ as

$$u(x, t) = w(x, t) + v(x, t).$$

Compute $v(x, t)$ for the specific case above, using (8.92) with at least $N = 5$ terms in the sum on the right. Then write out $u(x, t) = w(x, t) + v(x, t)$. Plot the solution at a variety of times. Does the solution $u(x, t)$ behave as you might expect? If possible, animate the solution on $0 \leq x \leq 4$ for $t = 0$ to $t = 5$.

- (d) Adapt this technique to show how to solve the heat equation with nonhomogeneous Neumann boundary conditions $\frac{\partial u}{\partial x}(0, t) = g_0(t)$ and $\frac{\partial u}{\partial x}(L, t) = g_L(t)$. Hint: choose $w(x, t) = xg_0(t) + (g_1(t) - g_0(t))x^2/(2L)$.

Exercise 8.3.9 It's not hard to prove that any solution to the heat equation (8.17) with Dirichlet boundary data (8.20) and initial data (8.19) is unique. To see why, suppose that functions $u_1(x, t)$ and $u_2(x, t)$ both satisfy the heat equation with the same Dirichlet boundary data and initial data. We will prove that in fact $u_1(x, t) = u_2(x, t)$, so that the solution is unique.

- (a) Let $u(x, t) = u_1(x, t) - u_2(x, t)$. Argue that u also satisfies the heat equation (8.17) but with boundary data $u(0, t) = u(L, t) = 0$ and initial data $u(x, 0) = 0$.
(b) Given that $\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = 0$ for $0 < x < L$ and $t > 0$, multiply the heat equation through by u to conclude that

$$u \frac{\partial u}{\partial t} - \alpha u \frac{\partial^2 u}{\partial x^2} = 0$$

for $0 \leq x \leq L$ and $t \geq 0$. Then integrate both sides above with respect to x from $x = 0$ to $x = L$ and then integrate again in t from $t = 0$ to $t = T$ (where T is some fixed positive time). Conclude that

$$\int_0^T \int_0^L u(x, t) \frac{\partial u}{\partial t}(x, t) dx dt - \alpha \int_0^T \int_0^L u(x, t) \frac{\partial^2 u}{\partial x^2}(x, t) dx dt = 0. \quad (8.93)$$

- (c) Show that the first double integral on the left in (8.93) can be (partially) evaluated as

$$\int_0^T \int_0^L u(x, t) \frac{\partial u}{\partial t}(x, t) dx dt = \frac{1}{2} \int_0^L u^2(x, T) dx. \quad (8.94)$$

Hint: do the double integral in the order t first, then x , and note that $u \frac{\partial u}{\partial t} = \frac{1}{2} \frac{\partial(u^2)}{\partial t}$. Also note that $u(x, 0) = 0$.

- (d) Perform the x integration in the second double integral on the left in (8.93) using integration by parts to show that

$$\begin{aligned} \int_0^L u(x, t) \frac{\partial^2 u}{\partial x^2}(x, t) dx &= u(L, t) \frac{\partial u}{\partial x}(L, t) - u(0, t) \frac{\partial u}{\partial x}(0, t) - \int_0^L \left(\frac{\partial u}{\partial x}(x, t) \right)^2 dx \\ &= - \int_0^L \left(\frac{\partial u}{\partial x}(x, t) \right)^2 dx. \end{aligned} \quad (8.95)$$

Hint: remember that $u(0, t) = u(L, t) = 0$.

- (e) Use (8.93), (8.94), and (8.95) to conclude that for any $T > 0$ we have

$$\frac{1}{2} \int_0^L u^2(x, T) dx + \alpha \int_0^T \int_0^L \left(\frac{\partial u}{\partial x}(x, t) \right)^2 dx dt = 0. \quad (8.96)$$

- (f) Argue that the integrand in each integral in (8.96) is nonnegative and hence each integral is nonnegative, and further, that (8.96) shows that both integrals must be zero. Then use the conclusion of Exercise 8.2.5 to conclude that the integrand of each integral in (8.96) is identically zero. Conclude from the first integral on the left in (8.96) that $u(x, T) = 0$ for all x and all T , so that u is the zero function. Finally, use part (a) to argue that this means $u_1(x, t) = u_2(x, t)$ for all x and t .
- (g) Show that the same argument works if we use Neumann boundary data (8.21) instead of Dirichlet data.

8.4 The Advection and Wave Equations

8.4.1 The Advection Equation

Let us return to the setting of Section 8.1.2 in a context in the conduit represents a river or canal of indefinite length along the x axis. Through this conduit water flows at a constant velocity of c distance units per time unit, with $c > 0$ indicating flow in the direction of increasing x , to the right in Figure 8.2. The stuff in this setting is some substance, say a pollutant dissolved in the water; the water itself is of no interest, but is merely the medium that transports the pollutant. This pollutant is measured on a per mass basis, so here the density function $\rho(x, t)$ quantifies the amount of pollutant in the water on a mass per length basis. As with the diffusion equation, ρ may also be viewed as a concentration. The flux function $q(x, t)$ quantifies the rate at which pollutant is flowing past the point x at time t on a mass per time basis. Assume that the pollutant is being neither created nor destroyed, and so is a conserved quantity. As a result the continuity equation (8.8) holds for all x and t .

A Constitutive Relation

As has been previously noted, the continuity equation itself does not provide enough information to determine ρ or q . Another equation is needed, a constitutive relation between ρ and q based on the physics of this situation. To find this relation, let us assume that the pollutant particles are all carried by the water at speed c , without any diffusion. With this assumption, consider a short time interval $t = t_0$ to $t = t_0 + \Delta t$, and a specific location $x = x_0$ in the conduit.

Reading Exercise 8.4.1

- (a) Use Figure 8.21 to argue that the amount of pollutant that will flow past the point $x = x_0$ during the time interval $t = t_0$ to $t = t_0 + \Delta t$ is the amount of pollutant in the conduit from $x = x_0 - c\Delta t$ to $x = x_0$ at time t_0 . This is the shaded portion of the conduit in Figure 8.21. Explain why this amount (a mass) is given by the integral

$$\int_{x_0 - c\Delta t}^{x_0} \rho(x, t_0) dx.$$

- (b) Explain why, based on part (a), the instantaneous rate $q(x_0, t_0)$ at which pollutant is flowing past the point x_0 at time t_0 is given by

$$q(x_0, t_0) = \lim_{\Delta t \rightarrow 0} \left(\int_{x_0 - c\Delta t}^{x_0} \rho(x, t_0) dx \right).$$

- (c) Argue that the limit above yields $q(x_0, t_0) = c\rho(x_0, t_0)$. Hint: one approach is to let $P(x, t_0)$ be an antiderivative for $\rho(x, t_0)$ with respect to x , then use the fundamental theorem of calculus to write the limit in (b) as

$$q(x_0, t_0) = c \lim_{\Delta t \rightarrow 0} \left(\frac{P(x_0, t_0) - P(x_0 - c\Delta t, t_0)}{c\Delta t} \right).$$

- (d) Verify that $q = c\rho$ is dimensionally correct.

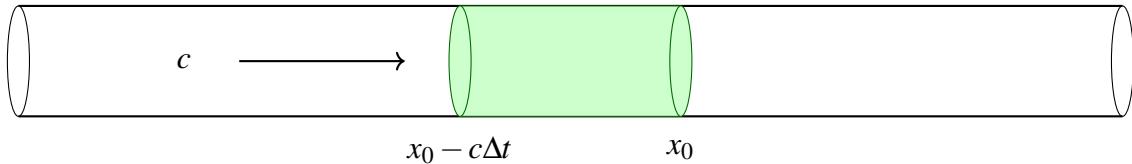


Figure 8.21: The shaded/green region is the pollutant that will be carried past $x = x_0$ from time $t = t_0$ to $t = t_0 + \Delta t$, if the conduit fluid moves at velocity c .

The Advection Equation

Reading Exercise 8.4.1 yields the constitutive relation

$$q(x, t) = c\rho(x, t) \quad (8.97)$$

at all points in the conduit and at all relevant times. We can use this to replace q in the continuity equation (8.8) and obtain the **advection equation** for $\rho(x, t)$,

$$\frac{\partial \rho}{\partial t} + c \frac{\partial \rho}{\partial x} = 0. \quad (8.98)$$

This is also known as the **one-way wave equation** or sometimes as the **transport equation**. The advection equation is a partial differential equation that governs how the density $\rho(x, t)$ evolves in time if the quantity of interest (in this case, pollutant) is rigidly transported with velocity c along the conduit.

8.4.2 Solution to the Advection Equation

Let's perform a simple thought experiment in order to solve the advection equation (8.98) when the conduit is of infinite length, so $-\infty < x < \infty$. It makes sense that we need some kind of initial concentration data, say

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

for some specified initial pollutant concentration $f(x)$. Given the assumptions that led to the constitutive relation (8.97), we should expect that the pollutant is transported along the x axis at velocity c . That is, at time t the pollutant profile $\rho(x, t)$ should simply be $f(x)$ translated ct units in the positive x direction. From elementary precalculus, the graph of $f(x)$ translated a distance ct to the right is just $f(x - ct)$. It follows that for $t > 0$ the pollutant concentration $\rho(x, t)$ should be given in terms of the initial concentration as

$$\rho(x, t) = f(x - ct). \quad (8.99)$$

Reading Exercise 8.4.2 Verify that $\rho(x, t)$ as given by (8.99) satisfies the advection equation (8.98) for any initial concentration $f(x)$, and that $\rho(x, 0) = f(x)$.

Visualizing the Solution; Characteristics

Consider the advection equation with a fixed wave speed $c > 0$ and initial data $\rho(x, 0) = e^{-x^2}$ (nothing special about this f , it just gives an aesthetically pleasing wave shape). The solution to the advection equation in this case is, from (8.99), $\rho(x, t) = e^{-(x-ct)^2}$. This solution is illustrated at times $t = 0, 1$, and 2 in the left panel of Figure 8.22. The initial density $f(x)$ is shown as the solid red curve. At time $t = 1$ the graph of the solution $\rho(x, 1) = f(x - c)$ is the graph of f shifted to the right c units (the dashed blue curve); at $t = 2$ the graph of $\rho(x, 2)$ is the graph of f shifted $2c$ units to the right (the dotted black curve).

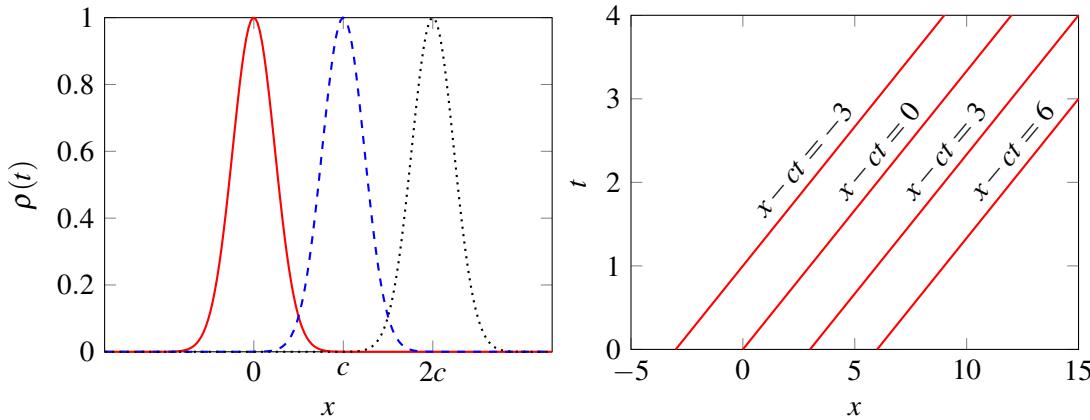


Figure 8.22: Left panel: solution to advection equation with initial data $\rho(x,0) = e^{-x^2}$ at times $t = 0$ (solid red), $t = 1$ (dashed blue) and $t = 2$ (dotted black); the solution propagates to the right at speed c . Right panel: characteristic curves $x - ct = x_0$ for $x_0 = -3, 0, 3$, and 6 .

The right panel of Figure 8.22 shows another way to think about how solutions evolve in time. This panel shows a space-time diagram, with the spatial variable x as the horizontal axis and time t as the vertical axis, limited to $t \geq 0$ (though we could include $t < 0$). This is the natural domain of the solution to the advection equation. The diagonal lines are lines of the form

$$x - ct = x_0 \quad (8.100)$$

for various values of x_0 , in this case $x_0 = -3, 0, 3$, and 6 . These lines are called the **characteristics** or **characteristic curves** for the advection equation (8.98). Note that x_0 is the x intercept of the line $x - ct = x_0$ (where $t = 0$). On these lines in the xt plane the solution to the advection equation is constant. One way to see this is to note that if $x = x_0 + ct$ (this is equivalent to (8.100)) then from (8.99) we have

$$\rho(x_0 + ct, t) = f((x_0 + ct) - ct) = f(x_0).$$

That is, the solution $\rho(x,t)$ on the characteristic $x + ct = x_0$ equals $f(x_0)$. Another way to see that ρ is constant on any characteristic curve is to compute the time derivative of ρ restricted to such a curve (where $x = x_0 + ct$). From the chain rule

$$\frac{d}{dt}(\rho(x_0 + ct, t)) = c \frac{\partial \rho}{\partial x}(x_0 + ct, t) + \frac{\partial \rho}{\partial t}(x_0 + ct, t) = 0$$

since ρ satisfies the advection equation (8.98). Thus $\rho(x_0 + ct, t)$ is constant, as asserted. If we know the value of ρ at any point on a characteristic then we know ρ at all points on the characteristic.

The characteristics in (8.100) can also be written as $t = x/c - x_0/c$, so they have a slope of $1/c$ with the t axis as vertical.

Reading Exercise 8.4.3 Suppose $f(x) = e^{-x^2}$ as in Figure 8.22 and $c = 3$ (this is the value used in both panels of that figure). Compute the value of $\rho(6,3)$ using (8.99). Then verify that the characteristic $x - 3t = -3$ passes through $(6,3)$ and trace this characteristic back to the relevant point on the x axis in Figure 8.22. Check that your result for $\rho(6,3)$ is equal to f at this point.

Reading Exercise 8.4.4 What would the characteristic curves look like if $c = 0$? Why does this make sense? What would the characteristic curves look like if $c < 0$?

8.4.3 The Wave Equation

Solutions to the advection equation propagate at velocity c along the x axis in the direction of increasing x (to the right in Figure 8.21) if $c > 0$ and in the direction of decreasing x (to the left in Figure 8.21) if $c < 0$. But in many physical systems solutions exhibit bidirectional behavior—stuff can propagate to the left or to the right. What kind of equation might allow both left- and right-moving waves at the same time? To answer this we introduce the useful notion of a *differential operator*.

Differential Operators

The expression $\frac{d}{dt}$ is an example of a **differential operator**. This differential operator acts on a function of t , say $f(t)$, and produces a new function, df/dt . It is assumed that only differentiable functions of t are presented to d/dt , so that df/dt is defined. Other examples of differential operators are

$$\frac{d^2}{dt^2}, \quad \frac{\partial^2}{\partial x \partial y}, \quad \text{and} \quad \frac{\partial}{\partial t} + c \frac{\partial}{\partial x}.$$

In each case the relevant differential operator is presented a suitable function as input ($f(t)$ for d^2/dt^2 , $f(x,y)$ for $\partial^2/\partial x \partial y$, and $f(x,t)$ for $\frac{\partial}{\partial t} - c \frac{\partial}{\partial x}$) and the differential operator acts on the function to produce the appropriate derivative or combination of derivatives.

Reading Exercise 8.4.5 Apply the differential operator $\frac{\partial}{\partial t} + 3 \frac{\partial}{\partial x}$ to each of the functions $u(x,t) = x$, $u(x,t) = xe^t$, and $u(x,t) = (x - 3t)^2$. Assume x and t are independent variables (so $\partial x/\partial t = 0$ and $\partial t/\partial x = 0$).

Differential operators can be added, subtracted, multiplied by scalars (for example, $\frac{\partial}{\partial t} - c \frac{\partial}{\partial x}$, is the sum of $\frac{\partial}{\partial t}$ and $-c \frac{\partial}{\partial x}$). Differential operators can also be composed. For example, $\frac{d^2}{dx^2} = \frac{d}{dx} \left(\frac{d}{dx} \right)$, or $\frac{\partial^2}{\partial x \partial y} = \frac{\partial}{\partial x} \left(\frac{\partial}{\partial y} \right)$. We generally assume that the differential operators commute. For example, if $f(x,y)$ is a suitably differentiable function of x and y then $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$ and so we write more abstractly

$$\frac{\partial^2}{\partial x \partial y} = \frac{\partial}{\partial x} \left(\frac{\partial}{\partial y} \right) = \frac{\partial}{\partial y} \left(\frac{\partial}{\partial x} \right) = \frac{\partial^2}{\partial y \partial x}.$$

Differential Operators and Solutions to DEs

The notation of differential operators sets up a useful parallel to that of solving algebraic equations. Consider an algebraic equation of the form $h(x) = 0$ for some function h . A solution to this equation is a root $x = x^*$ of $h(x)$, that is, a number x^* such that $h(x^*) = 0$. In a parallel manner we may think of the process of solving a differential equation like $df/dx = 0$ as that of finding a root (which here will be a function) of the differential operator d/dx . In this case there are infinitely many such roots or solutions, since any constant function of the form $f(x) = a$ satisfies $\frac{d}{dx}(f) = 0$.

A solution $\rho(x,t)$ to the advection equation (8.98) with wave velocity c can be considered as root of the differential operator $\frac{\partial}{\partial t} + c \frac{\partial}{\partial x}$, in that

$$\left(\frac{\partial}{\partial t} + c \frac{\partial}{\partial x} \right) (\rho) = 0. \tag{8.101}$$

Reading Exercise 8.4.6 Show that

$$\frac{d^2}{dt^2} + 3 \frac{d}{dt} + 2I = \left(\frac{d}{dt} + 2I \right) \left(\frac{d}{dt} + I \right) = \left(\frac{d}{dt} + I \right) \left(\frac{d}{dt} + 2I \right)$$

where “ I ” is interpreted as the identity operator, so $I(f) = f$ for any function f , and, for example, $\frac{d}{dt}I = I\frac{d}{dt} = \frac{d}{dt}$. Why does the factorization on the right above show that $u(t) = e^{-t}$ satisfies $d^2u/dt^2 + 3du/dt + 2u = 0$? (Hint: apply this differential operator to $u(t) = e^{-t}$ as $(\frac{d}{dt} + 2I)(\frac{d}{dt} + I)(u)$.) Use similar reasoning to show that $u(t) = e^{-2t}$ also satisfies $d^2u/dt^2 + 3du/dt + 2u = 0$.

The Wave Equation

Consider the task of constructing an algebraic equation that has two prescribed solutions, say $x = 2$ and $x = 5$. One way to do this is to devise an equation $h_1(x) = 0$ that has $x = 2$ as a solution, say by taking $h_1(x) = x - 2$. Do the same to obtain an equation $h_2(x) = 0$ with $x = 5$ as the solution, by taking $h_2(x) = x - 5$. Then the equation $h_1(x)h_2(x) = 0$, which is the equation $(x - 2)(x - 5) = 0$, has both $x = 2$ and $x = 5$ as solutions.

Based on this philosophy we can construct a differential equation that has both left- and right-moving waves as solutions, by using the differential operator in (8.101) with $c > 0$ and also with $-c < 0$. However, rather than using multiplication as with algebraic equations we will use a composition of these differential operators in the form

$$\begin{aligned} \left(\frac{\partial}{\partial t} - c \frac{\partial}{\partial x} \right) \left(\frac{\partial}{\partial t} + c \frac{\partial}{\partial x} \right) (\rho) &= \left(\frac{\partial}{\partial t} - c \frac{\partial}{\partial x} \right) \left(\frac{\partial \rho}{\partial t} + c \frac{\partial \rho}{\partial x} \right) \\ &= \frac{\partial^2 \rho}{\partial t^2} + c \frac{\partial^2 \rho}{\partial x \partial t} - c \frac{\partial^2 \rho}{\partial t \partial x} - c^2 \frac{\partial^2 \rho}{\partial x^2} \\ &= \frac{\partial^2 \rho}{\partial t^2} - c^2 \frac{\partial^2 \rho}{\partial x^2} \end{aligned} \quad (8.102)$$

where we assume that the function $\rho(x, t)$ is suitably differentiable so that $\frac{\partial^2 \rho}{\partial x \partial t} = \frac{\partial^2 \rho}{\partial t \partial x}$ and then the middle terms on the right in the second line of (8.102) cancel. Under the assumption that ρ is sufficiently differentiable, we can also reverse the differential operators on the left in (8.102) and find that

$$\left(\frac{\partial}{\partial t} + c \frac{\partial}{\partial x} \right) \left(\frac{\partial}{\partial t} - c \frac{\partial}{\partial x} \right) (\rho) = \frac{\partial^2 \rho}{\partial t^2} - c^2 \frac{\partial^2 \rho}{\partial x^2}.$$

So the order of the operators on the left in (8.102) doesn’t matter.

If $\rho(x, t)$ satisfies the advection equation $\frac{\partial \rho}{\partial t} + c \frac{\partial \rho}{\partial x} = 0$ (so ρ is a wave moving to the right at speed c , the direction of increasing x) then according to the computation of (8.102) the function ρ must also satisfy

$$\frac{\partial^2 \rho}{\partial t^2} - c^2 \frac{\partial^2 \rho}{\partial x^2} = 0 \quad (8.103)$$

since

$$\left(\frac{\partial}{\partial t} - c \frac{\partial}{\partial x} \right) \left(\frac{\partial}{\partial t} + c \frac{\partial}{\partial x} \right) (\rho) = \left(\frac{\partial}{\partial t} - c \frac{\partial}{\partial x} \right) (0) = 0$$

The same holds true if $\rho(x, t)$ satisfies the advection equation $\frac{\partial \rho}{\partial t} - c \frac{\partial \rho}{\partial x} = 0$ (so ρ is a wave moving to the left at speed c , the direction of decreasing x). Equation (8.103) is known as the **wave equation**. The parameter c is the **wave speed**. The wave equation is a PDE that is second-order in both the time and space variables.

There are two important observations to make about the wave equation. First, by design, both left- and right-moving waves are solutions, and these solutions move at speed c . Second, the wave equation is linear, so the principle of superposition holds: linear combinations of solutions are again solutions.

Reading Exercise 8.4.7 Verify by direct computation that $\rho(x, t) = f(x - ct) + g(x + ct)$ satisfies the wave equation (8.103), where f and g are any twice-differentiable functions.

The wave equation governs many different physical phenomena to good approximation, for example, the vibration of a string under tension. In the project “Strung Out” in Section 8.5.3 you can show directly from the basic physics that a vibrating string should be governed approximately by the wave equation; the wave speed c is related to the string’s linear density and tension. A two-space dimensional version of the wave equation governs the vibration of a drumhead. A three-space dimensional version governs the propagation of electromagnetic radiation and many phenomena involving the vibration of material objects. In the next two sections we’ll consider how to solve (8.103) when the region for x is bounded, for example, $0 \leq x \leq L$. We’ll also consider the case in which x can range over the whole real axis, $-\infty < x < \infty$.

8.4.4 Solution to the Wave Equation

A Vibrating String

Let’s consider the wave equation in the context of a vibrating string under tension. Assume the string is finite in length, like a guitar string, and spans the interval $0 \leq x \leq L$ for some length L . We will assume that the motion of the string as it vibrates is purely up and down in a fixed plane, with $u(x, t)$ denoting the vertical displacement of the string in terms of position x and time t . See Figure 8.23 for an illustration, a snapshot of the string’s vertical displacement at a fixed time $t = t_0$ for a string spanning the interval $0 \leq x \leq 1$; the displacement of the string at the point $x = 0.7$ is highlighted. The displacement $u(x, t)$ of the string will satisfy (8.103), which we rewrite here as

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0. \quad (8.104)$$

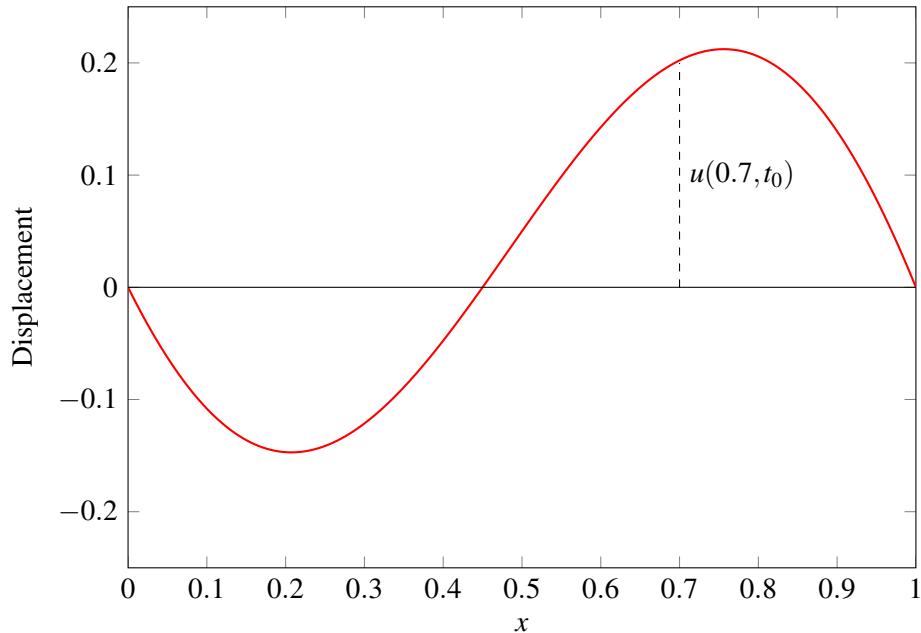


Figure 8.23: Vertical displacement $u(x, t)$ of vibrating string on interval $0 \leq x \leq 1$.

As indicated in Figure 8.23, the ends of the string at $x = 0$ and $x = L$ are fixed. For example, the ends may be clamped or tied, like a guitar string. With this assumption the ends of the string

satisfy the boundary conditions

$$u(0, t) = 0 \quad \text{and} \quad u(L, t) = 0 \quad (8.105)$$

at all times t . Our goal is to find the solution to the wave equation (8.104) on an interval $0 \leq x \leq L$ with the boundary conditions (8.105). As you might suspect, more information is needed to find a unique solution.

One piece of additional information that's needed is not surprising: the initial position of the string must be specified, that is,

$$u(x, 0) = f(x) \quad (8.106)$$

at the initial time $t = 0$. Given the boundary conditions (8.105) it also makes sense to require that $f(0) = f(L) = 0$. However, the initial position (8.106) by itself is not enough data to determine the future behavior of the string. One way to see this, at least intuitively, is to envision the string with initial position $u(x, 0) = 0$; one solution would then be to take $u(x, t) = 0$ for all x and t , but this isn't the only possibility with this initial position. The string may be in motion at time $t = 0$, even if it is at the equilibrium or zero position, and in this case the string will not remain at equilibrium in the future.

The additional information that is needed is the initial velocity of each part of the string at time $t = 0$, as

$$\frac{\partial u}{\partial t}(x, 0) = g(x) \quad (8.107)$$

for some function g . With the boundary data (8.105), initial position (8.106), and initial velocity (8.107), we will see that the string's motion for $t > 0$ is uniquely determined. Again, given the boundary conditions (8.105) it makes sense to require that $g(0) = g(L) = 0$.

Separating Variables

Let us approach solving the wave equation in the same manner as the heat equation, by first seeking separable solutions in the form $u(x, t) = T(t)X(x)$ for some functions $T(t)$ and $X(x)$. Insert $u(x, t) = T(t)X(x)$ into (8.104), compute the derivatives, then simplify to find that $T''(t)X(x) - c^2T(t)X''(x) = 0$ is required. A bit of rearrangement yields

$$\frac{T''(t)}{c^2T(t)} = \frac{X''(x)}{X(x)}. \quad (8.108)$$

The left side of (8.108) is a function solely of t and the right side is a function solely of x . The same argument given in Section 8.1.5 following (8.24) shows that both sides of (8.108) must be constant, the same constant. In anticipation of the nature of solutions to the wave equation (oscillatory in space and time) let us call this constant $-\lambda^2$; this will give rise to sines and cosines for the solution, and using λ^2 instead of λ means we won't have to write $\sqrt{\lambda}$ over and over again.

With this insight the left and right sides of (8.108) yield

$$\begin{aligned} T''(t) + c^2\lambda^2T(t) &= 0 \\ X''(x) + \lambda^2X(x) &= 0. \end{aligned} \quad (8.109)$$

Each ODE in (8.109) is the equation of an undamped oscillator. The general solutions are

$$\begin{aligned} T(t) &= A \cos(c\lambda t) + B \sin(c\lambda t) \\ X(x) &= C \cos(\lambda x) + D \sin(\lambda x). \end{aligned} \quad (8.110)$$

For any choice of the constants A, B, C, D , and λ the product $u(x, t) = T(t)X(x)$ yields a solution to the wave equation (8.104).

Invoking the Boundary Conditions

To make use of the boundary conditions (8.105) we proceed in precisely the same manner as the heat equation. Specifically, with $u(x, t) = T(t)X(x)$, the condition that $u(0, t) = 0$ forces $X(0) = 0$. In (8.110) this yields $C = 0$, so $X(x) = D\sin(\lambda x)$. Then the condition $X(L) = 0$ means that $D\sin(\lambda L) = 0$, but taking $D = 0$ leads to $X(x) = 0$ and $u(x, t) = 0$, which is of no use. We conclude that $\sin(\lambda L) = 0$ and so $\lambda L = k\pi$ for some integer k . The choice $k = 0$ again yields $X(x) = 0$, so discard that choice. Also, there is nothing gained by taking $k < 0$ since the sine function is odd (this merely changes the sign of X , and this can be absorbed into D). All in all we must take $\lambda L = k\pi$ for some integer k with $k \geq 1$, or $\lambda = k\pi/L$, just as for the heat equation. Then $X(x) = D\sin(k\pi x/L)$ for an integer $k \geq 1$.

The main (and dramatic) difference between this situation and the heat equation is that $T(t)$ is no longer a decaying exponential. In this case with $\lambda = k\pi/L$ it follows from (8.110) that $T(t) = A\cos(ck\pi t/L) + B\sin(ck\pi t/L)$. This yields a family of solutions $u_k(x, t) = T(t)X(x)$ to the wave equation, one for each integer $k \geq 1$, given by

$$u_k(x, t) = a_k \cos(ck\pi t/L) \sin(k\pi x/L) + b_k \sin(ck\pi t/L) \sin(k\pi x/L) \quad (8.111)$$

where we have lumped together the product of arbitrary constants AD and BD into a_k and b_k respectively, and indexed them by k (since they may change for each $u_k(x, t)$). For each integer $k \geq 1$ and each choice of constants a_k and b_k the function $u_k(x, t)$ satisfies the wave equation with boundary conditions $u_k(0, t) = 0$ and $u_k(L, t) = 0$.

Superposition and Obtaining the Initial Data

The wave equation is linear, and so finite sums of solutions are again solutions. Additional solutions can be constructed by using a superposition of the basic solutions u_k in (8.111), as

$$\begin{aligned} u(x, t) &= \sum_{k=1}^n u_k(x, t) \\ &= \sum_{k=1}^n (a_k \cos(ck\pi t/L) \sin(k\pi x/L) + b_k \sin(ck\pi t/L) \sin(k\pi x/L)) \end{aligned} \quad (8.112)$$

for some value of n (which will eventually be allowed to approach infinity). The goal now is to choose the coefficients a_k and b_k for $1 \leq k \leq n$ to best fit the initial data, or reproduce the initial data perfectly as $n \rightarrow \infty$. Not surprisingly, this involves Fourier series expansions of the initial data.

Evaluating $u(x, 0)$ with u as in (8.112) yields

$$u(x, 0) = \sum_{k=1}^n a_k \sin(k\pi x/L).$$

We want this sum to approximate f , and we know how to do this based on the analysis of Section 8.2.4: the a_k should be chosen as the Fourier sine coefficients for $f(x)$, namely

$$a_k = \frac{2}{L} \int_0^L f(x) \sin(k\pi x/L) dx. \quad (8.113)$$

This choice for the a_k will minimize the quantity $\|f - u(x, 0)\|_2$ and provide the best approximation to the initial position data in the sense of the L^2 distance. Moreover, if f is piecewise continuous then $\|f - u(x, 0)\|_2$ will approach zero as $n \rightarrow \infty$.

The b_k coefficients are determined by the initial velocity. First use (8.112) to compute

$$\frac{\partial u}{\partial t}(x, t) = \sum_{k=1}^n \frac{ck\pi}{L} (-a_k \sin(ck\pi t/L) \sin(k\pi x/L) + b_k \cos(ck\pi t/L) \sin(k\pi x/L))$$

Then

$$\frac{\partial u}{\partial t}(x, 0) = \sum_{k=1}^n \frac{ck\pi}{L} b_k \sin(k\pi x/L) \quad (8.114)$$

In order to best approximate $g(x)$ (that is, minimize $\|\partial u(x, 0)/\partial t - g\|_2$) we need to choose the coefficient on $\sin(k\pi x/L)$ on the right in (8.114) to match the Fourier sine coefficients for g , so that

$$\frac{ck\pi}{L} b_k = \frac{2}{L} \int_0^L g(x) \sin(k\pi x/L) dx.$$

Note that b_k itself is not the sine coefficient of g ; $ck\pi b_k/L$ is the sine coefficient. Solving for b_k yields

$$b_k = \frac{2}{\pi ck} \int_0^L g(x) \sin(k\pi x/L) dx. \quad (8.115)$$

This choice for b_k minimizes $\|\partial u(x, 0)/\partial t - g\|_2$. If g is piecewise continuous then $\|\partial u(x, 0)/\partial t - g\|_2$ approaches zero as $n \rightarrow \infty$.

It can be shown that if $f(x)$ is twice continuously-differentiable and $g(x)$ once continuously-differentiable then taking n to infinity in (8.112) and setting

$$u(x, t) = \sum_{k=1}^{\infty} (a_k \cos(ck\pi t/L) \sin(k\pi x/L) + b_k \sin(ck\pi t/L) \sin(k\pi x/L)) \quad (8.116)$$

with the a_k chosen according to (8.113) and the b_k chosen according to (8.115) yields a solution to the wave equation. The solution satisfies the boundary conditions (8.105) as well as the initial conditions (8.106) and (8.107). Even if f and g are not this differentiable we obtain a solution in a certain generalized sense that we won't go into here.

■ **Example 8.15** Consider the wave equation (8.104) on the interval $0 \leq x \leq 2$ (so $L = 2$) with wave speed $c = 3$, boundary conditions given by (8.105), and initial position $u(x, 0) = f(x)$ with $f(x) = x(2-x)$, initial velocity given by $\frac{\partial u}{\partial t}(x, 0) = g(x)$ with $g(x) = 4\sin(2\pi x)$. The Fourier sine coefficients for $f(x)$ on $0 \leq x \leq 2$ can be found analytically using (8.113) and are $a_k = 16(1 - (-1)^k)/(k^3\pi^3)$, so $a_k = 0$ when k is even and $a_k = 32/(k^3\pi^3)$ when k is odd. From (8.115) it follows that $b_4 = 2/(3\pi)$ and all other $b_k = 0$. From (8.116) the solution to the wave equation is

$$u(x, t) = \frac{2\sin(6\pi t)\sin(2\pi x)}{3\pi} + \frac{32}{\pi^3} \sum_{k=1, \text{ odd}}^{\infty} \frac{\cos(3k\pi t/2)\sin(k\pi x/L)}{k^3}.$$

The graph of $u(x, t)$ on the interval $0 \leq x \leq 2$ is shown in at times $t = 0, 0.3, 0.6$, and 0.9 in Figure 8.24. ■

■ **Example 8.16** Consider the vibrating string of a guitar. As you can show in the project “Strung Out” of Section 8.5.3, if a string with linear mass density λ and at a tension of T (units of force) vibrates vertically then the string’s vertical displacement $u(x, t)$ approximately obeys the wave equation (8.104) with wave speed $c = \sqrt{T/\lambda}$.

Reading Exercise 8.4.8 Verify that if T has the dimension of force and λ has the dimension of mass per length then c has the dimension of velocity or speed.

To illustrate, the “A” string on an acoustic guitar might have a typical linear density of $\lambda = 3.5 \times 10^{-3}$ kg per meter. Suppose the tension in this string is 70 newtons. In this case we have

$$c = \sqrt{\frac{70}{3.5 \times 10^{-3}}} \approx 141.42$$

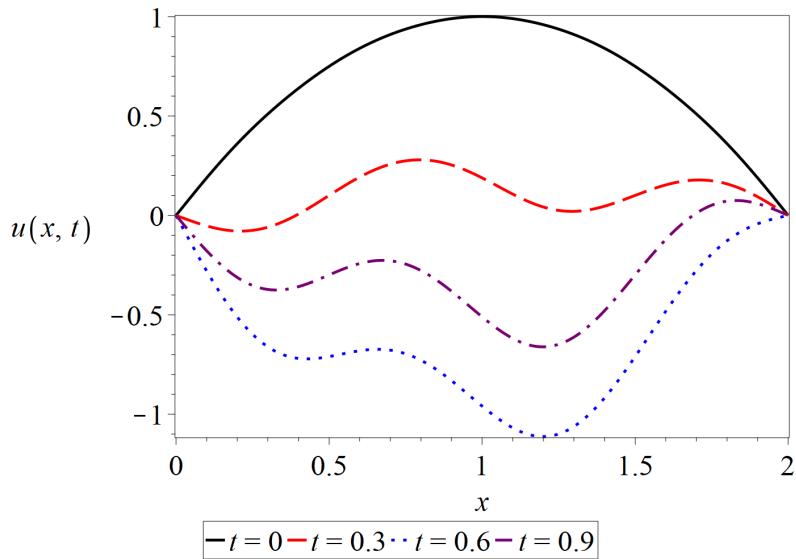


Figure 8.24: Solution to wave equation on $0 \leq x \leq 2$ with wave speed $c = 3$, boundary data $u(0, t) = u(2, t) = 0$, initial data $u(x, 0) = x(2-x)$ and $\frac{\partial u}{\partial t}(x, 0) = 4 \sin(2\pi x)$, at times $t = 0, 0.3, 0.6$, and 0.9 .

meters per second. Suppose the string is plucked, that is, pulled to a given displacement and released with zero initial velocity at time $t = 0$. A typical acoustic guitar string has a length of $L = 0.66$ meters. Suppose the initial displacement of the string is given by

$$f(x) = 0.01 - 0.03|x - 0.33|(1 - e^{-20|x - 0.33|})$$

(see left panel of Figure 8.25; this f might approximate the string if it is plucked at the halfway point, to a maximum displacement of one centimeter). We take $g(x) = 0$. From (8.116) the solution is

$$\begin{aligned} u(x, t) \approx & 0.00854 \cos(673.16t) \sin(4.76x) - 0.00094 \cos(2019.5t) \sin(14.28x) \\ & + 0.0032 \cos(3365.8t) \sin(23.80x) - \dots, \end{aligned}$$

to a few terms. The first term on the right corresponds to a radial frequency of 673.16, or $673.16/(2\pi) \approx 107.14$ Hz. This is the **fundamental frequency** of this string. The next two terms correspond to frequencies of about 321.4 and 535.7 Hz, three and five times the fundamental. These are **harmonics** of the fundamental frequency. The harmonics here are all odd multiples of the fundamental due to the symmetry of the initial displacement $f(x)$, but in general there can be harmonics that are both odd and even multiples of the fundamental frequency. The position of the guitar string at several times is shown in Figure 8.25.

Reading Exercise 8.4.9 The “A” string on a guitar should be tuned so that the fundamental is 110 Hz, so this instrument is out of tune. How should the tension in the string be adjusted to accomplish this? Assume $L = 0.66$ meters and $\lambda = 3.5 \times 10^{-3}$ kg per meter. Of course, you wouldn’t actually measure the tension, you’d adjust the tuning peg until the right frequency is obtained.

■

8.4.5 The Wave Equation on the Real Line

The solution to the wave equation on the domain $-\infty < x < \infty$ is also relatively easy to obtain. This is a one-dimensional analog of the wave equation that governs the propagation of light in three dimensions, though here you can think of a string of infinite (or at least indefinite) length.

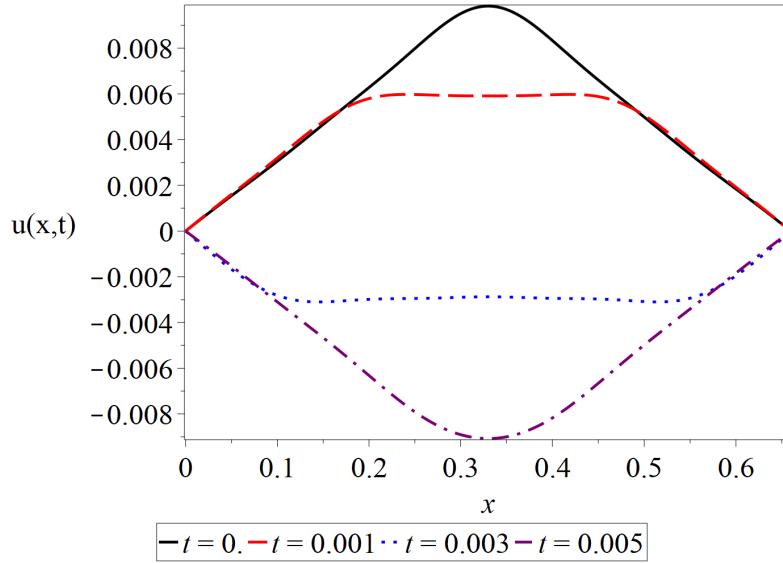


Figure 8.25: Solution to wave equation on $0 \leq x \leq 0.66$ (shape of guitar string) at time $t = 0, 0.001, 0.003$, and 0.005 .

We already know that left- and right-moving waves, which are solutions to the advection equation with velocity $\pm c$, both satisfy the wave equation (8.104), so we will seek a solution to (8.104) with initial position $u(x,0) = f(x)$ and initial velocity $\frac{\partial u}{\partial t}(x,0) = g(x)$ by using an ansatz that consists of a superposition of right- and left-moving waves in the form

$$u(x,t) = \phi(x - ct) + \psi(x + ct) \quad (8.117)$$

where ϕ and ψ are to be determined from the initial data. The initial condition $u(x,0) = f(x)$ in conjunction with (8.117) means ϕ and ψ have to satisfy

$$\phi(x) + \psi(x) = f(x). \quad (8.118)$$

Use (8.117) to compute $\frac{\partial u}{\partial t}(x,t) = -c\phi'(x-ct) + c\psi'(x+ct)$ so that $\frac{\partial u}{\partial t}(x,0) = g(x)$ forces

$$-c\phi'(x) + c\psi'(x) = g(x). \quad (8.119)$$

Equations (8.118) and (8.119) will allow us to determine ϕ and ψ .

To find ϕ and ψ , antidifferentiate both sides of (8.119) with respect to x to obtain

$$-c\phi(x) + c\psi(x) = G(x) \quad (8.120)$$

where $G(x)$ is any antiderivative for $g(x)$. Note that G is only determined up to an additive constant. Solve (8.118) and (8.120) algebraically for ϕ and ψ to obtain

$$\begin{aligned} \phi(x) &= \frac{1}{2}f(x) - \frac{1}{2c}G(x) \\ \psi(x) &= \frac{1}{2}f(x) + \frac{1}{2c}G(x). \end{aligned} \quad (8.121)$$

It's now convenient to define a specific choice for the antiderivative $G(x)$, namely

$$G(x) = \int_0^x g(z) dz$$

so that $G(0) = 0$. In this case (8.121) becomes

$$\begin{aligned}\phi(x) &= \frac{1}{2}f(x) - \frac{1}{2c} \int_0^x g(z) dz. \\ \psi(x) &= \frac{1}{2}f(x) + \frac{1}{2c} \int_0^x g(z) dz.\end{aligned}\tag{8.122}$$

If we substitute $\phi(x)$ and $\psi(x)$ from (8.122) into $u(x, t) = \phi(x - ct) + \psi(x + ct)$ (this is the ansatz (8.117)) we find that

$$\begin{aligned}u(x, t) &= \phi(x - ct) + \psi(x + ct) \\ &= \frac{1}{2}f(x - ct) - \frac{1}{2c} \int_0^{x-ct} g(z) dz + \frac{1}{2}f(x + ct) + \frac{1}{2c} \int_0^{x+ct} g(z) dz \\ &= \frac{1}{2}(f(x - ct) + f(x + ct)) + \frac{1}{2c} \left(\int_{x-ct}^0 g(z) dz + \frac{1}{2c} \int_0^{x+ct} g(z) dz \right) \\ &= \frac{1}{2}(f(x - ct) + f(x + ct)) + \frac{1}{2c} \int_{x-ct}^{x+ct} g(z) dz.\end{aligned}\tag{8.123}$$

In the transition from the second to third line in (8.123) we reversed the limits $z = 0$ to $z = x - ct$ in the first integral involving g and compensated by changing the sign of the integral. The transition from the third line to the last line combined the two integrals on the right using the usual rules from integral calculus. We thus have a fairly explicit solution to (8.104) on $-\infty < x < \infty$ with initial data $u(x, 0) = f(x)$ and $\frac{\partial u}{\partial t}(x, 0) = g(x)$. The solution provided by the last line in (8.123) is called the **D'Alembert solution** to the wave equation.

Examples

■ **Example 8.17** The situation when $g(x) = 0$ is exceptionally simple, for then (8.123) yields

$$u(x, t) = \frac{1}{2}(f(x - ct) + f(x + ct)).$$

This corresponds to two waves, one moving to the right at speed c , the other moving to the left. Both are the same shape as $f(x)$, but only half the amplitude. To illustrate, suppose that $c = 1$ and the initial data are $f(x) = e^{-x^2}$ and $g(x) = 0$ for all x . The solution in this case is

$$u(x, t) = \frac{1}{2}e^{-(x-t)^2} + \frac{1}{2}e^{-(x+t)^2}.$$

Figure 8.26 shows the solution at times $t = 0$, $t = 2$, and $t = 4$. The initial displacement $f(x)$ splits into two parts that move in opposite directions at speed 1. ■

■ **Example 8.18** Consider the case of Example 8.17 with $c = 1$, initial position $f(x) = e^{-x^2}$, but with initial velocity $g(x) = x/(1+x^2)^2$. The solution in this case is, from (8.123),

$$\begin{aligned}u(x, t) &= \frac{1}{2}e^{-(x-t)^2} + \frac{1}{2}e^{-(x+t)^2} + \frac{1}{2} \int_{x-t}^{x+t} \frac{xdx}{(1+x^2)^2} \\ &= \frac{1}{2}e^{-(x-t)^2} + \frac{1}{2}e^{-(x+t)^2} + \frac{2xt}{((x-t)^2+1)((x+t)^2+1)}\end{aligned}$$

after working the integral. Figure 8.27 shows the solution at times $t = 0$, $t = 2$, and $t = 5$. ■

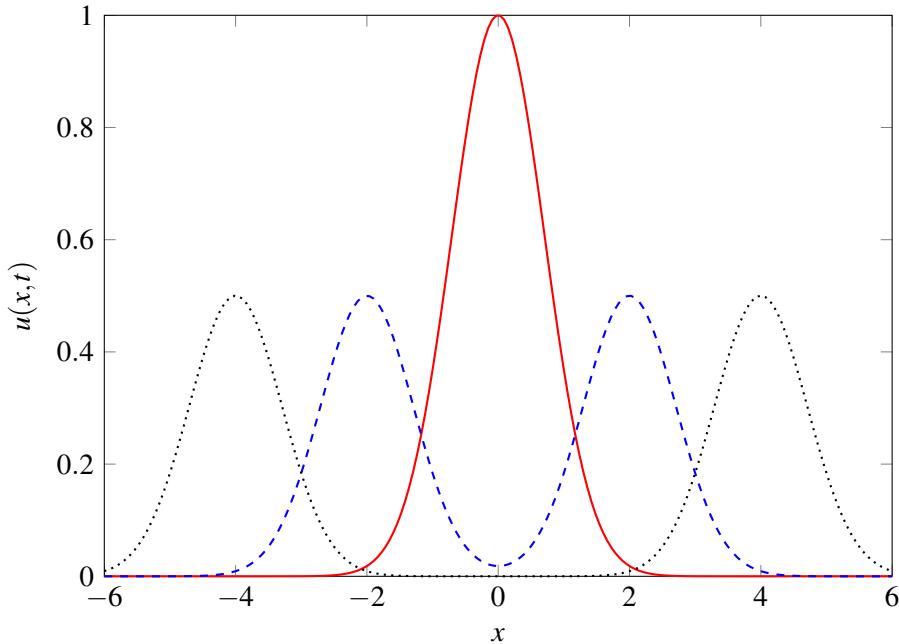


Figure 8.26: Solution to the wave equation with initial position $f(x) = e^{-x^2}$ and initial velocity $g(x) = 0$, at time $t = 0$ (solid red), $t = 2$ (dashed blue), and $t = 4$ (dotted black).

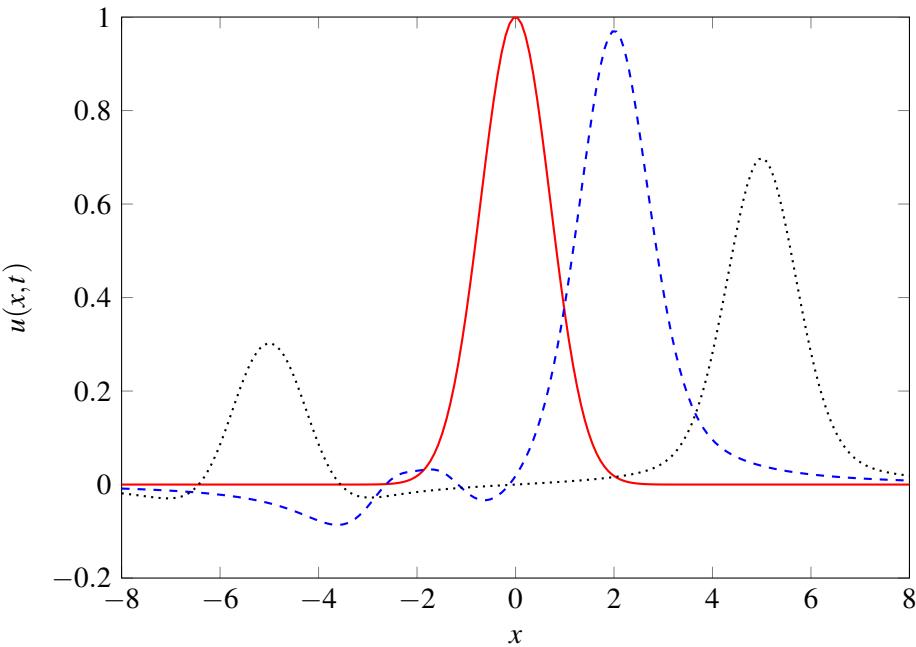


Figure 8.27: Solution to the wave equation with wave speed $c = 1$, initial position $f(x) = e^{-x^2}$, and initial velocity $g(x) = x/(1+x^2)^2$, at time $t = 0$ (solid red), $t = 2$ (dashed blue), and $t = 5$ (dotted black).

The Light Cone and Causality

Refer to Figure 8.28. Here the x (spatial) axis is the horizontal axis and the t axis (time) is vertical, so upward movement is forward in time and downward movement is backward in time, the same setting as Figure 8.22 for the advection equation. The point (x_0, t_0) in this figure is some

fixed location in space and time. The shaded yellow region inside the lines $x + ct = x_0 + ct_0$ and $x - ct = x_0 - ct_0$ is called the *backward light cone* for (x_0, t_0) . If you look at the D'Alembert solution (8.123) for the value of $u(x_0, t_0)$, you will see that the solution $u(x_0, t_0)$ is synthesized out of data from $x = x_0 - ct_0$ to $x = x_0 + ct_0$ along the $t = 0$ axis (the horizontal axis), which is where the initial data lives. Specifically, to compute $u(x_0, t_0)$ using (8.123), we integrate g from $x = x_0 - ct_0$ to $x = x_0 + ct_0$ and combine this with the average the value of f at $x = x_0 - ct_0$ and $x = x_0 + ct_0$ to form $u(x_0, t_0)$. This is the basis of the **principle of causality** for the wave equation. Because the solution at $x = x_0, t = t_0$ is determined only by data along the x axis from $x = x_0 - ct$ to $x = x_0 + ct$, the values of f and g on the x axis outside the backward light cone of (x_0, t_0) have no effect on the solution at (x_0, t_0) .

There's nothing special about initial conditions at $t = 0$, either; we could just as well have taken initial conditions at $t = 1$, in which only the initial data on the horizontal line $t = 1$ that lie inside the backward light cone are relevant to determining the solution at (x_0, t_0) . More generally, only events that occur inside the backward light cone of (x_0, t_0) can affect the solution at (x_0, t_0) . Thus, for example, whatever occurs at the point (x_1, t_1) in Figure 8.28 cannot affect the solution at (x_0, t_0) , for the light black line that connects (x_1, t_1) to (x_0, t_0) represents information propagating faster than c . (Be careful: this line has a slope that is shallower than the sides of the light cone, but because x is horizontal and t is vertical this is faster than c .) The principle of causality for the wave equation is sometimes stated "information cannot travel faster than c ."

The flip side of the coin is that the value of $u(x_0, t_0)$ at (x_0, t_0) can influence the solution to the wave equation only at those points in the forward light cone of (x_0, t_0) , as illustrated in Figure 8.29, for a point (x_2, t_2) lies in the forward light cone of (x_0, t_0) if and only if (x_0, t_0) lies in the backward light cone for (x_2, t_2) .

Reading Exercise 8.4.10 What would the backward light cone for (x_0, t_0) look like if c is close to zero? What would the backward light cone look like if c were large (limits to infinity)?

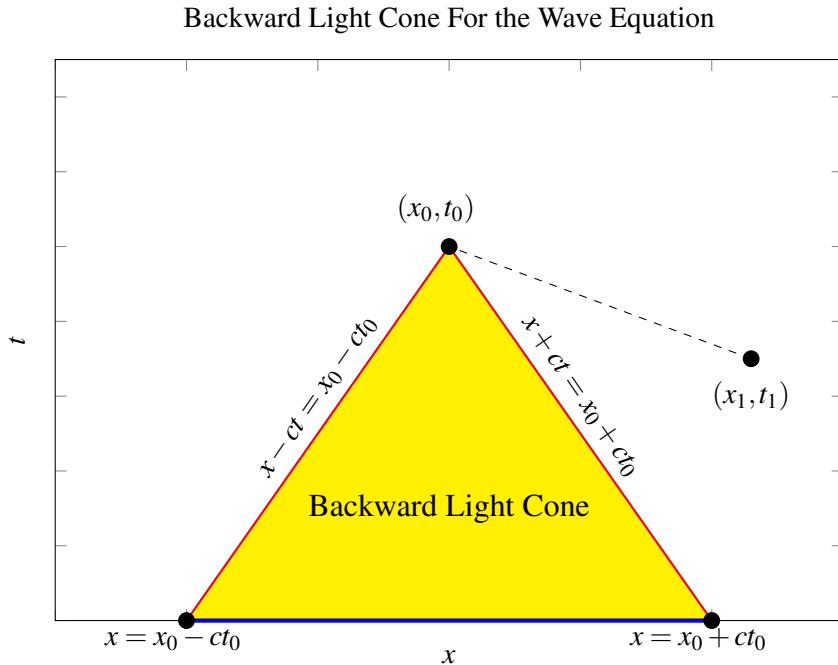


Figure 8.28: The backward light cone for the point (x_0, t_0) .

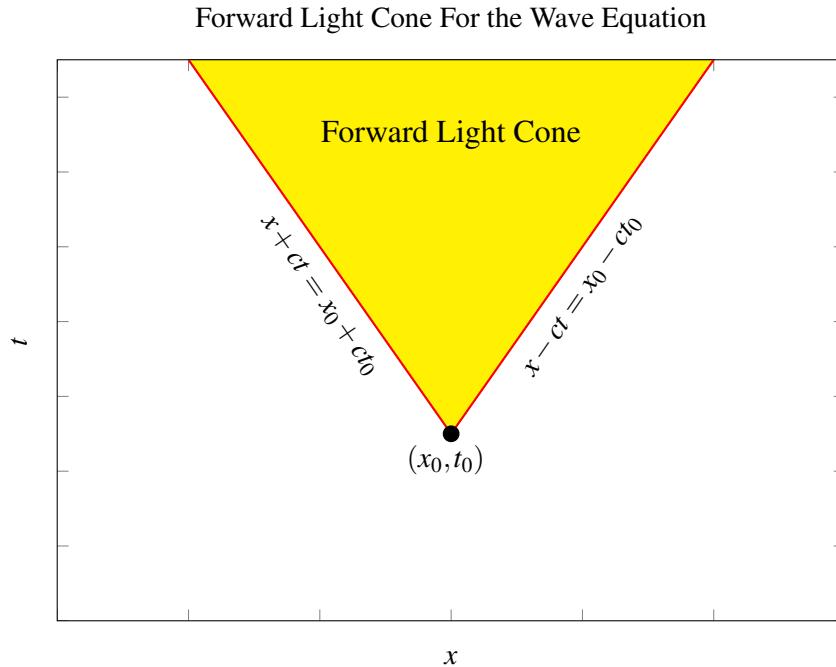


Figure 8.29: The forward light cone for the point (x_0, t_0) .

8.4.6 Exercises

Exercise 8.4.1 In each case below solve the advection equation (8.98) with the given value of c and initial data $\rho(x, 0) = f(x)$. Verify that your solution in fact satisfies (8.98), and plot the solution at times $t = 0, 2, 5$ on a suitable interval for x .

- (a) $c = 2, f(x) = x/(x^2 + 1)$
- (b) $c = -2, f(x) = x/(x^2 + 1)$
- (c) $c = 0, f(x)$ unspecified. Provide a physical interpretation (and recall Reading Exercise 8.4.4).

Exercise 8.4.2 Find a Fourier series solution of the form (8.116) to the wave equation on the given interval $0 \leq x \leq L$ with the given initial conditions $u(x, 0) = f(x)$, $\frac{\partial u}{\partial t}(x, 0) = g(x)$ and boundary conditions $u(0, t) = u(L, t) = 0$. Compute the solution to N terms as indicated, and plot the solution as a function of x at the given times.

- (a) $0 \leq x \leq 2, c = 1, f(x) = \sin(\pi x), g(x) = 0, N = 3$ terms. Plot at $t = 0, 0.4, 0.8$.
- (b) $0 \leq x \leq 2, c = 1, f(x) = \sin(\pi x), g(x) = 3 \sin(2\pi x), N = 5$ terms. Plot at $t = 0, 0.4, 0.8$.
- (c) $0 \leq x \leq 4, c = 2, f(x) = 0, g(x) = 2 - |x - 2|, N = 5$ terms. Plot at $t = 0, 0.5, 1.0, 2.3$.
- (d) $0 \leq x \leq 10, c = 2, f(x) = e^{-(x-5)^2}, g(x) = 4(x-5)e^{-(x-5)^2}, N = 10$ terms. Plot at $t = 0, 1, 2, 3, 5, 7, 8$.

Exercise 8.4.3 In each case below a second-order differential operator D involving $d^2/dt^2, d/dt$, and I (the identity operator, $I(f) = f$ for a function $f(t)$) is given. Factor the operator into two first-order differential operators P_1 and P_2 and find a solution to each corresponding first-order ODE $P_k(u) = 0$, $k = 1, 2$. Show that these solutions also satisfy the relevant homogeneous second-order ODE $D(u) = 0$.

- (a) $D = d^2/dt^2 + 9d/dt + 8I$
- (b) $D = d^2/dt^2 + 4d/dt + 4I$
- (c) $D = d^2/dt^2 + 9I$ (complex solutions allowed)

Exercise 8.4.4 Suppose that $\rho(x, t)$ satisfies the wave equation with speed c for $-\infty < x < \infty$ and with initial condition $\rho(x, 0) = f(x)$, and that ρ represents a right-moving wave. Show that the initial velocity data for $\rho(x, t)$ must be $\frac{\partial \rho}{\partial t}(x, 0) = -cf'(x)$. Hint: as a right-moving wave, $\rho(x, t)$ must satisfy the advection equation with speed c .

Exercise 8.4.5 Find a second order linear PDE for a function $u(x, t)$ in which left-moving waves that move at speed 2 are solutions and right-moving waves of speed 5 are solutions. Hint: compose the relevant differential operator for the advection equation that has left-moving waves of speed 2 as solutions with the operator for the advection equation that has right-moving waves of speed 5 as solutions. Verify that $u(x, t) = f(x + 2t) + g(x - 5t)$ does in fact satisfy the PDE you find for any choice of f and g .

Exercise 8.4.6 Suppose a string of density λ is held at tension T , so the wave speed in this string is $c = \sqrt{T/\lambda}$, as remarked in Example 8.16 (and you can show in the Project “Strung Out” of Section 8.5.3). Suppose the string is of length L with boundary conditions $u(0, t) = u(L, t) = 0$. Assume all units are standard metric or SI. Show that the lowest frequency (with respect to time) that can appear in the wave equation solution (8.116) is $\pi\sqrt{T/\lambda}/L$ radians per second, or $\frac{\sqrt{T/\lambda}}{2L}$ hertz. This is the fundamental frequency of this string.

Exercise 8.4.7 The highest-pitch string on a standard acoustic guitar is the “E” string. Such a string might have a density of about $\lambda = 3.1 \times 10^{-4}$ kg per meter. Suppose this string is tensioned to $T = 58.69$ newtons and has a length of $L = 0.66$ meters. Assume the string obeys the wave equation with boundary conditions $u(0, t) = u(L, t) = 0$.

- (a) What is the wave speed c in this string? What is the lowest frequency term (in hertz) that can appear in the solution (8.116) to the wave equation here? Hint: see Exercise 8.4.6. The “E” string should vibrate at a fundamental frequency of 329.63 hz; is this string in tune?
- (b) Suppose the initial displacement of the string is $u(x, 0) = f(x)$ with

$$f(x) = \begin{cases} 0, & x < 0 \\ 0.04545x, & 0 \leq x < 0.22 \\ 0.015 - 0.02273x, & 0.22 \leq x \leq 0.66 \\ 0, & x \geq 1 \end{cases}$$

(This approximates the string being plucked 1/3 of the way along its length, to a displacement of 0.01 meter.) Assume the initial velocity is zero, so $\frac{\partial u}{\partial t}(x, 0) = 0$. Work out a

5-term Fourier expansion for the string's motion using (8.116). Plot the string position for $0 \leq x \leq 0.66$ at times $t = 0, 0.0005, 0.0008$, and 0.0012 .

Exercise 8.4.8

- (a) Verify that if $u(x, t)$ satisfies the wave equation then so does the function $w(x, t) = u(x, -t)$. In this sense we say that the wave equation is *time reversible*.
- (b) Verify that if $u(x, t)$ satisfies the heat equation then $w(x, t) = u(x, -t)$ does not satisfy the heat equation, but rather satisfies

$$\frac{\partial w}{\partial t} + \alpha \frac{\partial^2 w}{\partial x^2} = 0. \quad (8.124)$$

Thus the heat equation is not time reversible. Equation (8.124) is known as the **backward heat equation**.

Exercise 8.4.9 In formulating the wave equation (8.104) we composed the differential operators $\partial/\partial t - c\partial/\partial x$ and $\partial/\partial t + c\partial/\partial x$ to arrive at the wave equation, with the observation that if $u(x, t)$ satisfies either advection equation then u will satisfy the wave equation. But we could have simply multiplied these advection equations, as

$$\left(\frac{\partial u}{\partial t} - c \frac{\partial u}{\partial x} \right) \left(\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} \right) = \left(\frac{\partial u}{\partial t} \right)^2 - c^2 \left(\frac{\partial u}{\partial x} \right)^2.$$

Be careful: the right side above involves the first derivatives squared, not second derivatives. If u satisfies the advection equation with wave velocity $\pm c$ then u should satisfy the PDE

$$\left(\frac{\partial u}{\partial t} \right)^2 - c^2 \left(\frac{\partial u}{\partial x} \right)^2 = 0. \quad (8.125)$$

- (a) Verify the if $u(x, t) = \phi(x - ct)$ or $u(x, t) = \phi(x + ct)$ for some differentiable function ϕ then u satisfies (8.125).
- (b) What drawback does (8.125) present as a model of wave motion? Hint: does superposition hold?

Exercise 8.4.10 This exercise is devoted to modeling the situation in which a pollutant is transported along a conduit (for example, a river or canal) at velocity c but decays over time, so the total amount of pollutant is not conserved. We then show how to solve the resulting PDE that results, a simple variation on the advection equation.

In Exercise 8.1.7 you derived (8.38), a modified version of the continuity equation that holds when stuff is not conserved. Suppose this is the case for the situation in Section 8.4.1, in which the constitutive relation (8.97) holds. We'll assume, as in that section, that the stuff is a pollutant being carried by the water. The pollutant will be measured on a per mass basis, so ρ has the dimension of mass per length and q has the dimension of mass per time.

- (a) Argue that if $r(x, t)$ denotes the rate at which the pollutant is being created ($r > 0$) or

destroyed ($r < 0$) on a mass per length per unit time basis then $\rho(x, t)$ satisfies the PDE

$$\frac{\partial \rho}{\partial t} + c \frac{\partial \rho}{\partial x} = r(x, t).$$

- (b) Suppose that the pollutant in any specific short length of the conduit is being destroyed at a rate proportional to the amount present in that short length (maybe the pollutant is radioactive and decays in time, or perhaps is chemically broken down over time). Argue that taking $r(x, t) = -k\rho(x, t)$ is a reasonable model here, where k is some positive constant with the dimension of reciprocal time. Verify that this relation is dimensionally consistent.
- (c) Argue that based on parts (a) and (b) the function ρ should satisfy the PDE

$$\frac{\partial \rho}{\partial t} + c \frac{\partial \rho}{\partial x} = -k\rho. \quad (8.126)$$

The goal is to solve this PDE with initial data $\rho(x, 0) = f(x)$.

- (d) The PDE (8.126) can be solved by using the characteristic curves $x - ct = x_0$ defined in (8.100), where x_0 is the x -intercept (when $t = 0$). Specifically, consider the value of ρ along any such characteristic $x = x_0 + ct$, which is given by $\rho(x_0 + ct, t)$ (where we've used $x = x_0 + ct$ in $\rho(x, t)$). For notational clarity, let $\phi(t) = \rho(x_0 + ct, t)$ denote the value of ρ on such a characteristic, as a function of t . Use (8.126) to show that $\phi(t)$ satisfies $d\phi/dt = -k\phi$. Solve this for $\phi(t)$ using the fact that $\phi(0) = \rho(x_0, 0) = f(x_0)$.
- (e) Conclude that the solution to (8.126) is given by $\rho(x, t) = f(x - ct)e^{-kt}$. Verify this by substituting this form for ρ into (8.126).
- (f) With $f(x) = e^{-x^2}$, $c = 1$, and $k = 0.25$, write out $\rho(x, t)$ and graph ρ as a function of x on the interval $-10 \leq x \leq 10$ at times $t = 0, 2, 5, 8$. Do the results seem reasonable for a pollutant being transported at speed 1 and decaying over time?

8.5 Modeling Projects

8.5.1 Project: It's a Blast (Furnace)!

In this project we will perform an analysis of a simplified version of the problem posed in Section 8.1.1. You should begin by re-reading that section.

We will first make two simplifications to the general problem of determining the thickness of the refractory lining of a furnace. First, the full problem is three-dimensional and the furnace wall/refractory lining may be a complex shape. We will instead examine a one-dimensional version of the problem that may approximate the behavior of the three-dimensional case. Also, although the temperature of the furnace may vary over time, when the furnace is at a steady-state condition of operation the temperature should not depend on time, to good approximation. This simplified model will serve as a *proof of concept*, to illustrate that it might in fact be possible to determine the thickness of the furnace wall or lining from exterior temperature measurements (or perhaps, demonstrate that even in a simple setting it cannot be done).

A One-Dimensional Model

To motivate a one-dimensional version of the problem, consider Figure 8.30. Here we delineate a hypothetical cylindrical cross-section D of the furnace wall and refractory lining; on one end lies the exterior of the furnace and at the other lies the interior molten charge of the furnace, with the central axis of D labeled as x . Assume that in the region of the furnace near the cylindrical cross-section D the flow of heat energy is horizontally outward from the interior to the exterior of

the furnace, parallel to the x axis, and that the temperature of the wall and refractory lining depends only on the horizontal position x . As a result no heat energy will flow over the lateral boundary of D , and the temperature of the wall and lining will be only a function of x and t .

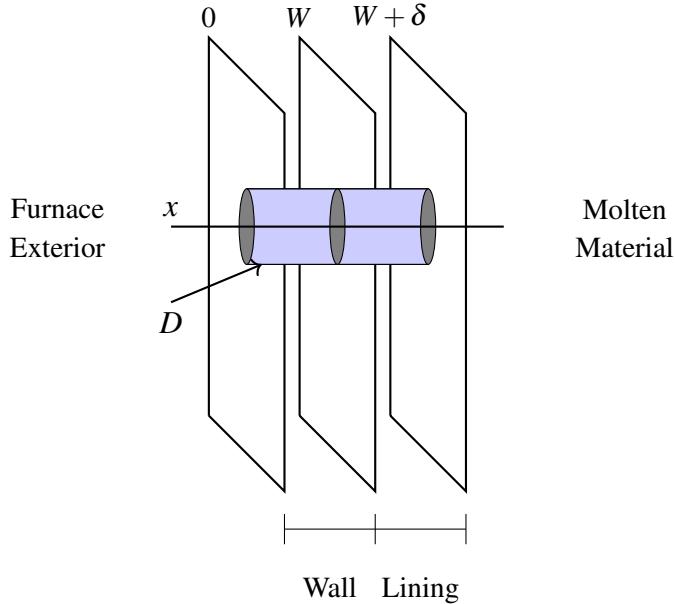


Figure 8.30: A cylindrical cross-section D (shaded green) of the furnace wall and refractory lining.

This setup is the basis for a one-dimensional model for the flow of heat through the furnace wall and lining. Let us take $x = 0$ as the outer wall of the furnace, in contact with the outside air, $x = W$ as the location of the wall/refractory lining interface, and $x = W + \delta$ as the location of the interface between the lining and the interior of the furnace. It is the value of δ that we wish to deduce, from data collected at $x = 0$.

The material that makes up the furnace wall and the material that makes up the refractory lining will have different physical properties. Let us use c_W for the specific heat of the material that makes up the wall and c_L for the material that makes up the lining. Similarly let k_W and k_L denote the thermal conductivities of the wall and lining, respectively, and λ_W and λ_L denote the linear density of these materials through the conduit D (these linear densities would depend on the cross-sectional area of D , but it won't matter). Though the wall and lining materials may differ in their physical properties, assume that the wall and lining regions themselves are homogeneous (the material parameters are constant in each). With these assumptions the temperature $u(x, t)$ throughout D satisfies the heat equation, in the wall ($0 < x < W$), and in the lining ($W < x < W + \delta$), in the form

$$\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = 0 \quad (8.127)$$

where from (8.18) we have diffusivity $\alpha_W = \frac{k_W}{c_W \lambda_W}$ in the wall and $\alpha_L = \frac{k_L}{c_L \lambda_L}$ in the lining.

Steady-State Heat Flow

As mentioned, we will assume that the furnace is in steady-state operation, with all associated physical quantities independent of time, or at least changing very slowly. As a consequence, the temperature $u(x, t)$ is independent of time, that is, the temperature is a function $u(x)$. The same conclusion holds for the thermal energy density $\rho(x)$ and the heat flux $q(x)$.

Modeling Exercise 5.1.1 Use (8.127) to argue that the temperature $u(x)$ satisfies $u''(x) = 0$ in the wall $0 < x < W$ and in the lining $W < x < W + \delta$, and that u is a linear function in each region

(recall Reading Exercise 8.1.8 and Exercise 8.1.6). Note that the value of the diffusivity in each region is irrelevant in this case.

The Wall-Lining Interface

From Modeling Exercise 5.1.1 the temperature $u(x)$ is linear in the wall and in the refractory lining. However, u need not be the same linear function in each region; this is due to the abrupt change in the material properties across the interface at $x = W$, as will be shown. Suppose that $u(x) = m_W x + b_W$ in the wall region $0 < x < W$ for some constants m_W and b_W , and $u(x) = m_L x + b_L$ in the lining region $W < x < W + \delta$ for some constants m_L and b_L . As a piecewise function, the temperature throughout the furnace wall and lining is

$$u(x) = \begin{cases} m_W x + b_W, & 0 \leq x < W \\ m_L x + b_L, & W \leq x < W + \delta \end{cases} \quad (8.128)$$

The value of the constants m_W, b_W, m_L , and b_L are constrained by some physics.

First, the temperature function $u(x)$ must be continuous over the interface at $x = W$. This follows from the fact that in our model of heat flow, thermal energy flows in proportion to the temperature gradient. If the temperature is discontinuous at $x = W$ this indicates that there is a nonzero temperature drop or rise in an arbitrarily small interval around $x = W$, and this corresponds to an unbounded temperature gradient and heat flux q at $x = W$, which is not physically possible. We thus enforce continuity of u at $x = W$ in (8.128).

Modeling Exercise 5.1.2 Requiring continuity of $u(x)$ at $x = W$ means that

$$\lim_{x \rightarrow W^-} u(x) = \lim_{x \rightarrow W^+} u(x).$$

Show that this forces the condition $m_W W + b_W = m_L W + b_L$.

There is another condition that must hold at the interface. The continuity equation $\frac{\partial \rho}{\partial t} + \frac{\partial q}{\partial x} = 0$ holds for $0 < x < W$ and $W < x < W + \delta$ where ρ is thermal energy density in the wall; ρ is related to the temperature u according to (8.12). But if the problem is steady-state this means that $\frac{\partial \rho}{\partial t} = 0$ and it follows that $\frac{\partial q}{\partial x} = 0$ in each of $0 < x < W$ and $W < x < W + \delta$, so that $q(x)$ must be constant in each region. Moreover, q must be constant across the interface at $x = W$. To see this suppose $q = q^-$ for $x < W$ and $q = q^+$ for $x > W$ and consider a small interval I of the form $W - \epsilon < x < W + \epsilon$ around $x = W$. The net heat flux into the interval I is $q^- - q^+$ (q^- enters on the left, q^+ leaves on the right). If $q^- \neq q^+$ then there would be a net heat flux into or out of I over time, and this is inconsistent with steady-state operation; the region I would be heating up or cooling down. It follows that q is constant on the whole interval $0 < x < W + \delta$.

Modeling Exercise 5.1.3 In the wall we have $q(x) = -k_W u'(x)$ and in the lining $q(x) = -k_L u'(x)$. Show that if q is constant for $0 < x < W + \delta$ then

$$-k_W m_W = -k_L m_L.$$

Based on Modeling Exercises 5.1.2 and 5.1.3, the constants m_W, b_W, m_L , and b_L in (8.128) have to satisfy the conditions

$$\begin{aligned} m_W W + b_W &= m_L W + b_L \\ -k_W m_W &= -k_L m_L. \end{aligned} \quad (8.129)$$

Let's begin to fill in some values for the relevant material constants. If the wall of the furnace is made of steel, a typical value for k_W in this region would be $k_W \approx 25$ watts per meter per degree Celsius. For the refractory lining a value of $k_L \approx 0.5$ watts per meter per degree Celsius is more in the ballpark (typical of firebrick). These values can vary with temperature, but we won't model this. A typical value for the wall thickness is W is $W = 0.15$ meters. The value for δ can vary, but we'll use $\delta = 0.25$ meters for now. See references [68, 12] for ranges of these parameter values. With these choices the second equation in (8.129) becomes $-25m_W = -0.5m_L$ from which we can immediately deduce that

$$m_L = 50m_W. \quad (8.130)$$

With (8.130) and $W = 0.15$, the first equation in (8.129) yields $0.15m_W + b_W = 7.5m_W + b_L$ or

$$b_W = 7.35m_W + b_L. \quad (8.131)$$

Equations (8.130) and (8.131) impose two conditions on the set m_W, m_L, b_W , and b_L and leave two free variables in the solution (8.128). We need two additional conditions to determine all of the constants in (8.128) and these come from boundary conditions at the interior and exterior of the furnace.

The Boundary Conditions

A furnace generally operates at a controlled, fixed internal temperature, so at the inner wall ($x = W + \delta$) the temperature may be considered known. Let us suppose that this temperature is 1600 degrees Celsius (see [68]). This is one boundary condition.

The second boundary condition comes from the physics at $x = 0$, the outer wall. This stems from the loss of heat through the walls of the furnace to the ambient environment. Let's take a simple model in which the wall of the furnace loses heat in a Newton-cooling fashion: the heat flux out of the wall (which is proportional to $-u'(0)$) is proportional to the difference between the wall and ambient temperature. Suppose the ambient temperature is $T_A = 30$ degrees Celsius; the Newton-cooling assumption leads to a boundary condition of the form

$$\frac{du}{dx}(0) = k_0(u(0) - T_A) \quad (8.132)$$

for some positive constant k_0 . We will take $k_0 = 0.7$ as a start.

The Forward Problem

With given boundary conditions, as well as specified values for W , δ , and the thermal conductivity in each region, the problem of determining the temperature profile $u(x)$ in (8.128) is called the **forward problem** or the **direct problem**. It's the traditional thing to do with a differential equation—find the solution.

Modeling Exercise 5.1.4 Make use of (8.130), (8.131), the condition $u(W + \delta) = 1600$, and (8.132) with the specified parameters values to solve for m_L, m_W, b_L , and b_W . Then use (8.128) to show that $u(x)$ is given approximately by

$$u(x) = \begin{cases} 111.52x + 189.31, & 0 \leq x < 0.15 \\ 5575.8x - 630.34, & 0.15 \leq x < 0.4 \end{cases}$$

Plot $u(x)$ on the range $0 \leq x \leq 0.4$. Is the plot consistent with $u''(x) = 0$ in each region $0 < x < W$ and $W < x < W + \delta$? Is $u(x)$ continuous through $x = W$? What the external temperature of the furnace wall? What is the hottest temperature for the steel portion of the furnace wall?

The Inverse Problem

Reading Exercises 5.1.1 to 5.1.4 show that if all essential parameters are known, namely W and δ (to specify the geometry), k_W and k_L (material parameters) and the boundary condition parameters and data, we can compute the temperature $u(x)$ at any point in the wall. But the problem of determining whether the refractory lining of the furnace is thinning is the problem of determining the value of δ , the thickness of the refractory lining, which we do not know and cannot measure directly. We have to infer δ from data concerning $u(x)$ —that's the real problem.

Let's suppose that in addition to knowing the Newton-cooling boundary condition at $x = 0$ and the internal temperature at $x = W + \delta$ (even if δ itself is not known), we have an additional piece of data: the value of $u(0)$, the temperature of the outer wall. This is something that can be readily measured. If we now treat δ as an unknown, but use the additional piece of information that $u(0) = u_0$ (here u_0 is the measured temperature) can we estimate δ ? This is an example of an **inverse problem** in which one uses information about the solution to a DE in order to estimate some unknown parameter that appears in the differential equation. It is also a parameter estimation problem as in Section 3.4.

For an inverse problem the essential questions of interest are:

- **Uniqueness** Can the unknown of interest be uniquely identified from the data at hand? If not, what additional data or assumptions would make that possible?
- **Reconstruction** Is there a constructive procedure for estimating the unknown quantity?
- **Stability** How sensitive is any estimate of the unknown to noise in the data or modeling assumptions?

Analysis of the Inverse Problem

Modeling Exercise 5.1.5 The goal here is to treat δ as unknown and infer it from an additional measurement of u . We'll assume that we know that $W = 0.15$, $k_1 = 25$, $k_2 = 0.5$ as before, as well as $k_0 = 0.7$ and $T_A = 30$. We'll also suppose that we measure $u(0) = 200$. The value of δ is considered unknown.

- (a) Redo the solution procedure of Modeling Exercise 5.1.4, but treat δ as an unknown, so now $u(x)$ will depend on δ . Show that the requirement that $u(0) = 200$ leads to the equation $\frac{45.804+30\delta}{\delta+0.0316} = 200$ (or something equivalent).
- (b) Solve the equation of part (a) for δ ; what is the thickness of the refractory lining? Plot $u(x)$ for $0 \leq x \leq W + \delta$.

Modeling Exercise 5.1.6 Suppose that the external temperature is $u(0) = u_0$, where u_0 is unspecified. Modify the equation from part (a) of Modeling Exercise 5.1.5 appropriately. Solve the equation for δ in terms of u_0 , and argue that we can always determine the lining thickness δ from a measurement of u_0 (at least for $u_0 > 30$). This addresses the uniqueness question above.

Modeling Exercise 5.1.7 Suppose that in Modeling Exercise 5.1.5 we erroneously measure $u(0) = 205$ (instead of $u(0) = 200$). How far off is our estimate of δ ? What if $u(0) = 195$? This gets at the sensitivity issue—we hope that small errors in our data result in small errors in our estimates.

Modeling Exercise 5.1.8 Suppose that the minimum safe value for the lining thickness is $\delta = 0.15$, with all other parameters as in Modeling Exercise 5.1.5. What is the maximum value for the external temperature $u(0)$ that can be allowed before shutting down the furnace for repair?

Conclusion

Of course, a real furnace is three-dimensional, the heat flow is time-dependent, and many facets of the operation have not been modeled (actively cooled walls as noted, and the precise boundary condition on the outer face are but two examples). All of this would have to be accounted for if one wanted to actually use these ideas for furnace operation. But as noted, a simplified model can still

give valuable intuition, by illuminating what variables might be most important, and whether what we seek to do can be done at all.

For a version of this problem in which the full time-dependent heat equation is used, see [131] (work done by undergraduates) and [102, 44] for more information on this problem.

8.5.2 Project: Finding Polluters

Background and Source Localization

As mentioned in Example 8.13, **source localization** is a common problem in applied mathematics. The goal is to locate the source of some substance or energy by using data that may be obtained far from the source itself. One example is *acoustic source localization* in which data from microphones can be used to locate a sound source, for example, a gunshot; see [74], though the idea goes back to at least World War I [103]. Another example is locating a radio source from remote measurements; see [27]. Locating sources of electrical activity in the brain from electroencephalogram (EEG) data is a common problem in medical imaging; see [54].

In this project we revisit Example 8.13 with a view toward determining the location of a pollutant source in a body of water from measurements of the pollutant concentration far from the source. See [86, 84, 104] for an examination some specific scenarios, or [82] for a survey and many additional references on this problem. We will assume that the body of water is still, so that the pollutant merely diffuses, but many models include a transport term similar to the advection equation.

A Pollutant Model

You should begin by reviewing Example 8.13 in Section 8.3.4. Assume that a mass A of pollutant is dumped into a waterway at time $t = 0$ and at some unknown location $x = x_0$. Let $\rho(x, t)$ denote the pollutant density or concentration on a mass per length basis in this one-dimensional stretch of water spanned by $0 \leq x \leq L$. Assume that the pollutant spreads only by diffusion so (8.77) holds for some diffusivity α . We also assume that no water (and hence no pollutant) flows in or out of the ends, so that $\frac{\partial \rho}{\partial x}(0, t) = \frac{\partial \rho}{\partial x}(L, t) = 0$. These are like insulating boundary conditions for the heat equation.

Modeling Exercise 5.2.1 If at time $t = 0$ a polluter dumps A mass units of pollutant into this waterway at location $x = x_0$ then we have initial condition $\rho(x, 0) = A\delta(x - x_0)$ (this is (8.78)). Show that the pollutant concentration at time $t > 0$ is given by

$$\rho(x, t) = \frac{A}{L} + \sum_{j=1}^{\infty} a_j e^{-\alpha\pi^2 j^2 t/L^2} \cos(j\pi x/L) \quad (8.133)$$

where

$$a_j = \frac{2}{L} \int_0^L A\delta(x - x_0) \cos(j\pi x/L) dx = \frac{2A}{L} \cos(j\pi x_0/L). \quad (8.134)$$

The Localization Problem: Uniqueness

The goal now is to use concentration data collected somewhere in the waterway to determine the source location of the pollutant and how much was dumped. We will assume that the data collected is $\rho(0, t)$ for times $t > 0$. In practice this data would consist of measurements of $\rho(0, t)$ at finitely many times, say $t = t_1, t_2, \dots, t_n$. But in order to understand the limitations of what we can do, assume for the moment that $\rho(0, t)$ can be measured for all $t > 0$ with perfect precision. The length L of this waterway is known, as is α , the diffusivity of the pollutant in water.

Under these ideal circumstances we can show that determining A and x_0 from the data $\rho(0, t)$ is theoretically possible.

Modeling Exercise 5.2.2 Use (8.133) to argue that we can determine the amount of pollutant A that was dumped by computing

$$\lim_{t \rightarrow \infty} \rho(0, t).$$

Modeling Exercise 5.2.3 Use (8.133) and the result of Modeling Exercise 5.2.2 to argue that we can determine the location x_0 at which the pollutant was dumped, by computing

$$\lim_{t \rightarrow \infty} e^{\alpha\pi^2 t/L^2} (\rho(0, t) - A/L).$$

Hint: show this is a_1 in (8.134), and that x_0 can be recovered from a_1 uniquely.

The Localization Problem: Data Collection and Sensitivity

Based on Modeling Exercises 5.2.2 and 5.2.3, with perfect data $\rho(0, t)$ for $t > 0$ we can recover the amount and location of pollutant that was dumped. But we can only take data for a finite time window. If we know the pollutant dump occurred at $t = 0$, when should we start collecting data, and when should we stop? If we start collecting data too soon (right at $t = 0$) the pollutant will not have diffused to $x = 0$ and all concentrations we measure will be close to zero. If we wait a very long time then Reading Exercise 5.2.2 shows that $\rho(0, t) \approx A$ no matter what x_0 is, so we will be able to deduce A but will have no information about x_0 . There should be some optimal time window for collecting data, in which the data contains as much information about A and x_0 as possible.

A simple numerical experiment gives some insight as to when concentration data $\rho(0, t)$ should be collected. In what follows let's fix $L = 10$ and $\alpha = 1$. Let's assume that data will be collected on some time interval $t_i \leq t \leq t_f$ for initial and final times t_i and t_f , respectively. Suppose that the pollutant source is suspected to be near some point $x = x_0$. Our goal is to estimate x_0 as precisely as possible. Let $d_0(t)$ be the concentration $\rho(0, t)$ at $x = 0$ when the source is at $x = x_0$. Let $x = x_1$ be another potential dump location that is near x_0 , and let $d_1(t)$ be the concentration $\rho(0, t)$ at $x = 0$ when the source is at $x = x_1$. In order to reliably determine where the pollutant was dumped, we'd like to choose t_i and t_f so that $d_0(t)$ and $d_1(t)$ differ as much as possible on the interval $t_i \leq t \leq t_f$, so that x_0 and x_1 can be distinguished from each other.

Modeling Exercise 5.2.4 Use (8.133) to write out a 20-term (at least) solution for $\rho(x, t)$ with initial amount $A = 1$ and $x_0 = 5$ on the interval $0 \leq x \leq 10$ with diffusivity $\alpha = 1$, then plot $d_0(t) = \rho(x, t)$ on the range $0 \leq t \leq 50$. Also write out a 20-term (at least) solution with $A = 1$ and source at $x_1 = 4$ and plot $d_1(t) = \rho(0, t)$ for this case, also on the range $0 \leq t \leq 50$ (the behavior of $d_0(t)$ and $d_1(t)$ might be a bit anomalous near $t = 0$). You should find that $d_0(t)$ and $d_1(t)$ differ the most for $1 \leq t \leq 30$, although this is a bit subjective. Why does it make physical sense that the solutions should differ little when t is close to zero? Why do the solutions differ little when t is large?

Repeat this same comparison when $x_0 = 2$ versus $x_1 = 3$, or $x_0 = 8$ versus $x_1 = 9$. What time window yields the greatest difference in the measured data?

Based on Modeling Exercise 5.2.4, if the pollution source is not too far from the data collection at $x = 0$ then we can expect the data at $\rho(0, t)$ on the time interval $1 \leq t \leq 30$ to contain information about the location of the source. But with data confined to such an interval the results from Modeling Exercises 5.2.2 and 5.2.3 are mostly useless, as they involve limits as t approaches infinity. Instead, we can try a least squares approach.

Modeling Exercise 5.2.5 Let $L = 10$ and $\alpha = 1$. Suppose we take data at times $t = 5, 10, 15, 20, 25$, and 30 ; let d_j denote the measurement of $\rho(0, 5j)$, that is, $d_1 = \rho(0, 5), d_2 = \rho(0, 10), \dots, d_6 =$

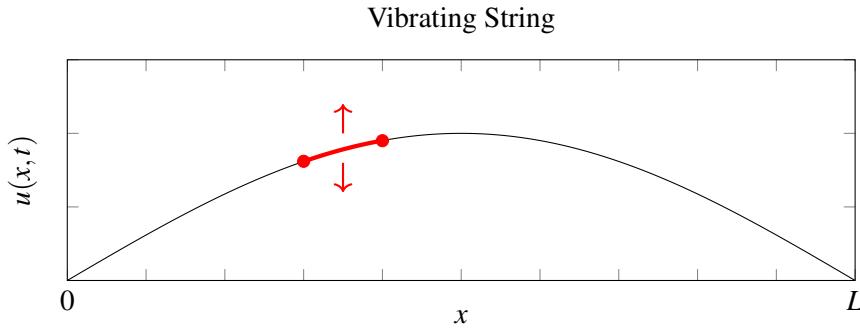


Figure 8.31: Snapshot at time t of a vibrating string spanning $0 \leq x \leq L$ and typical short element of the string (element shown in red/thick).

$\rho(0, 30)$. Consider A , the amount of pollutant, and x_0 , the location at which it was dumped, to be unknown. Suppose the data are

$$d_1 = 0.135, d_2 = 0.153, d_3 = 0.149, d_4 = 0.146, d_5 = 0.143, d_6 = 0.141$$

to three significant figures. Use this to estimate x_0 and A . You can try a guess and check approach or form an appropriate sum of squares (treating x_0 and A as the unspecified parameters to be estimated) and minimize.

Modeling Exercise 5.2.6 Suppose the data can only be measured to two significant figures. Round the data from Modeling Exercise 5.2.5 appropriately and estimate A and x_0 from this rounded data. Repeat when the data is accurate to only one significant figure. Do the estimates of A and x_0 seem to be sensitive to finite precision or other error in the data?

Modeling Exercise 5.2.7 Suppose the data is taken instead at later times, $t = 30, 35, 40, 45, 50, 55$, and yields

$$d_1 = 0.141, d_2 = 0.140, d_3 = 0.139, d_4 = 0.139, d_5 = 0.138, d_6 = 0.138$$

where d_1 corresponds to $t = 30$, d_2 to $t = 35$, etc. Can you explain why the data is nearly constant? Use this data to estimate A and x_0 . Explain why we should expect the estimate of A to be accurate, but probably not x_0 .

8.5.3 Project: Strung Out

In this project we use basic physics to show that to good approximation the wave equation governs the motion of a string that vibrates transversely to its axis in a vertical plane. Consider a string that spans an interval $0 \leq x \leq L$ along the x axis as illustrated in Figure 8.31. The ends of the string are fixed to the x axis at the ends $x = 0$ and $x = L$. Let $u(x, t)$ denote the vertical displacement of the vibrating string from the x axis at time t . Since the ends are fixed we have $u(0, t) = u(L, t) = 0$ for all times t . Assume the string has a uniform linear density of λ mass units per length unit and that the string is under a constant tension T throughout its length. The equilibrium position of the string is thus $u(x, t) = 0$.

Assumptions

Our analysis will focus on a typical short piece of the string, a *string element*, also shown in Figure 8.31. We will make use of various approximations in this analysis. We will assume that

- The amplitude of the string's vibration is small (close to zero displacement, in a sense to be specified).

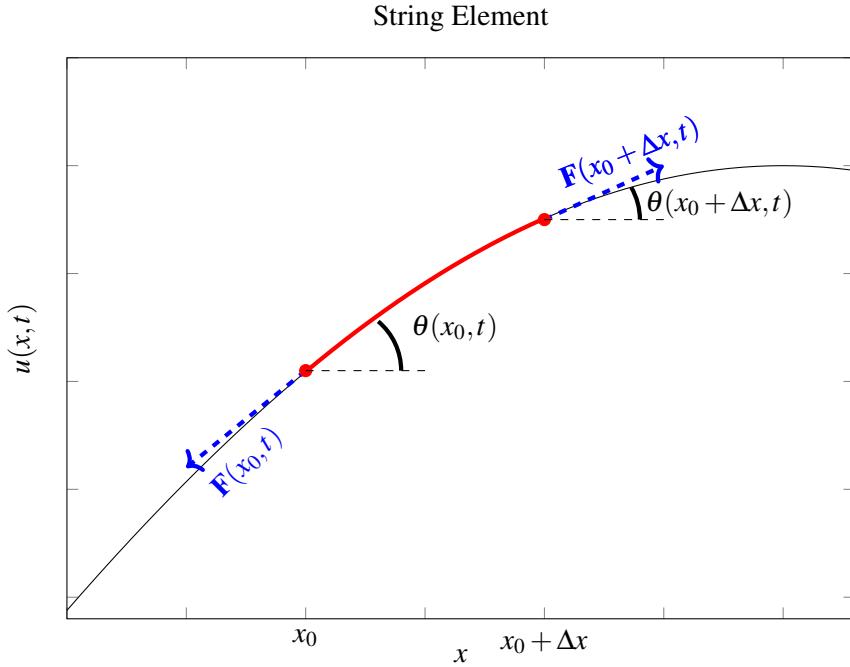


Figure 8.32: String element spanning $x = x_0$ to $x = x_0 + \Delta x$ and forces acting on the element at time t .

- The slope $\partial u / \partial x$ of the string also remains close to zero at all times.
- The motion of the string element is purely vertical. As a consequence, if the string element stretches from $x = x_0$ to $x = x_0 + \Delta x$ in the horizontal direction when the string is at equilibrium, this element still stretches from $x = x_0$ to $x = x_0 + \Delta x$ when the string is in motion.
- The tension T in the vibrating string is constant (to good approximation) in space and time.

Approximations and Analysis

The analysis focuses on the application of Newton's second law $\mathbf{F} = m\mathbf{a}$ to the string element, in conjunction with the above assumptions.

Modeling Exercise 5.3.1 Let m denote the mass of the string element. Argue that if the slope of the string $\partial u / \partial x$ remains close to 0 then the length of the string element is well-approximated by Δx and the mass m of the string element is approximately $m = \lambda \Delta x$.

Modeling Exercise 5.3.2 Justify that if the string motion is purely vertical then the acceleration vector \mathbf{a} for the string element's motion in Newton's second law should be $\mathbf{a} = \langle 0, \partial^2 u / \partial t^2 \rangle$.

The last ingredient in applying Newton's second law to the string element in Figure 8.31 is an accounting of the net force acting on the string element. Refer to Figure 8.32, which provides a close-up view of the element. The dashed blue vectors labeled $\mathbf{F}(x_0, t)$ and $\mathbf{F}(x_0 + \Delta x, t)$ at each end of the element represent the forces acting on the relevant ends of the string element at time t due to the tension in the string. These forces are tangential to the string at each point. The quantities $\theta(x_0, t)$ and $\theta(x_0 + \Delta x, t)$ indicate the angle that the string makes at that point with respect to the horizontal, with the conventional counterclockwise orientation in radians. The net force \mathbf{F}_{net} on the string is

$$\mathbf{F}_{net} = \mathbf{F}(x_0 + \Delta x, t) + \mathbf{F}(x_0, t). \quad (8.135)$$

In what follows let us suppress the dependence of the various quantities on the time variable t ,

and so write $\mathbf{F}(x_0)$ instead of $\mathbf{F}(x_0, t)$ or $\frac{\partial u}{\partial t}(x_0)$ instead of $\frac{\partial u}{\partial t}(x_0, t)$, since all quantities below are evaluated at a common fixed time.

Modeling Exercise 5.3.3 Argue that since the tension in the string is T and is constant in time and space we have (refer to Figure 8.32)

$$\begin{aligned}\mathbf{F}(x_0) &= -T \langle \cos(\theta(x_0)), \sin(\theta(x_0)) \rangle \\ \mathbf{F}(x_0 + \Delta x) &= T \langle \cos(\theta(x_0 + \Delta x)), \sin(\theta(x_0 + \Delta x)) \rangle,\end{aligned}$$

then use (8.135) to show that the net force on the string element is

$$\begin{aligned}\mathbf{F}_{net} &= \mathbf{F}(x_0 + \Delta x) + \mathbf{F}(x_0) \\ &= T \langle \cos(\theta(x_0 + \Delta x)) - \cos(\theta(x_0)), \sin(\theta(x_0 + \Delta x)) - \sin(\theta(x_0)) \rangle.\end{aligned}\tag{8.136}$$

If $f(x)$ is any suitably differentiable function then a Taylor expansion at $x = x_0$ shows that

$$f(x_0 + \Delta x) = f(x_0) + \frac{df}{dx}(x_0)\Delta x + O((\Delta x)^2)$$

where $O((\Delta x)^2)$ means terms proportional to $(\Delta x)^2$ and/or higher powers of Δx . Since Δx will be close to zero in our analysis, $(\Delta x)^2$ is negligible compared to Δx and we may drop the $O((\Delta x)^2)$ on the right in part (a), then move the $f(x_0)$ to the left side to obtain an approximation

$$f(x_0 + \Delta x) - f(x_0) \approx \frac{df}{dx}(x_0)\Delta x.\tag{8.137}$$

Modeling Exercise 5.3.4

- (a) Apply (8.137) to the function $f(x) = \sin(\theta(x))$ to justify the approximation

$$\sin(\theta(x_0 + \Delta x)) - \sin(\theta(x_0)) \approx \cos(\theta(x_0)) \frac{\partial \theta}{\partial x}(x_0) \Delta x.\tag{8.138}$$

where we write $\frac{\partial \theta}{\partial x}$ since θ depends on x and t , even if the t dependence is suppressed.

- (b) Apply (8.137) to the function $f(x) = \cos(\theta(x))$ to justify the approximation

$$\cos(\theta(x_0 + \Delta x)) - \cos(\theta(x_0)) \approx -\sin(\theta(x_0)) \frac{\partial \theta}{\partial x}(x_0) \Delta x.\tag{8.139}$$

We can make some additional reasonable approximations. First, we are assuming that the string maintains a small slope as it vibrates, so $\theta(x)$ is close to zero at all times, and for all x . The usual Taylor expansion for the cosine function shows that $\cos(\theta) = 1 - \theta^2/2 + O(\theta^4)$ and if $\theta \approx 0$ we have a good approximation that $\cos(\theta) \approx 1$. In this case we may write (8.138) as

$$\sin(\theta(x_0 + \Delta x)) - \sin(\theta(x_0)) \approx \frac{\partial \theta}{\partial x}(x_0) \Delta x.\tag{8.140}$$

We also have the Taylor approximation $\sin(\theta) = \theta + O((\theta)^3)$ and if $\theta \approx 0$ this yields the approximation $\sin(\theta) \approx \theta$ (perhaps familiar from elementary calculus). Then (8.139) can be approximated as

$$\cos(\theta(x_0 + \Delta x)) - \cos(\theta(x_0)) \approx -\theta(x_0) \frac{\partial \theta}{\partial x}(x_0) \Delta x.\tag{8.141}$$

Modeling Exercise 5.3.5 Use (8.140) and (8.141) in (8.136) to justify the approximation

$$\mathbf{F}_{\text{net}} = T\Delta x \frac{\partial \theta}{\partial x}(x_0) \langle -\theta(x_0), 1 \rangle \quad (8.142)$$

for the net force on the string element.

The vector $\langle -\theta(x_0), 1 \rangle$ that appears on the right in (8.142) has a first component, $-\theta(x_0)$, that should be very close to zero if the string's slope remains small while it vibrates. We will thus ignore this horizontal component of the force and focus solely on the vertical component $T\Delta x \frac{\partial \theta}{\partial x}(x_0)$. Note this is entirely consistent with our assumption that the motion of the string element is purely vertical.

In Newton's second law $\mathbf{F}_{\text{net}} = m\mathbf{a}$ we drop the first (horizontal) component of each vector and match the second (vertical) components. From (8.142) for the second component of \mathbf{F} and from Modeling Exercises 5.3.1 and 5.3.2 we find $\lambda\Delta x \partial^2 u / \partial t^2 = T\Delta x \frac{\partial \theta}{\partial x}(x_0)$ or

$$\lambda \frac{\partial^2 u}{\partial t^2}(x_0) = T \frac{\partial \theta}{\partial x}(x_0). \quad (8.143)$$

One last approximation is needed. From Figure 8.32 we see that

$$\theta(x) = \arctan\left(\frac{\partial u}{\partial x}(x)\right).$$

If we differentiate both sides above with respect to x we find that

$$\begin{aligned} \frac{\partial \theta}{\partial x} &= \frac{\partial^2 u / \partial x^2}{1 + (\partial u / \partial x)^2} \\ &\approx \frac{\partial^2 u}{\partial x^2} \end{aligned} \quad (8.144)$$

if we again make use of the assumption that $\partial u / \partial x$ remains close to zero, so the denominator on the right in the first line of (8.144) is approximately 1.

Modeling Exercise 5.3.6 Make use of (8.144) in (8.143) to justify the wave equation

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0 \quad (8.145)$$

where $c^2 = T/\lambda$ as a PDE that must be obeyed by the function $u(x, t)$ that describes the vertical displacement of the string of linear density λ and under tension T . Does $c = \sqrt{T/\lambda}$ seem reasonable? For example, is it intuitively consistent with how wave speed should behave with respect to the tension T ? If the string density λ increases, would you expect this to increase or decrease the wave speed?

Adding Damping

Modeling Exercise 5.3.7 Suppose the string experiences frictional forces as it vibrates. Specifically, suppose the string element of Figure 8.31 or Figure 8.32 experiences a vertical force that is proportional and opposed to its vertical velocity. It also makes sense that this force should be proportional to the length of the element, Δx . Argue for a frictional term of the form

$$\mathbf{F}_{\text{frictional}} = \langle 0, -\mu \frac{\partial u}{\partial t} \Delta x \rangle$$

to be added to the right side of (8.142), for some nonnegative constant μ . Then redo the chain of reasoning from (8.142) to (8.145) to show that string displacement should obey

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} + \beta \frac{\partial u}{\partial t} = 0 \quad (8.146)$$

where $c^2 = T/\lambda$ and $\beta = \mu/\lambda$. This is known as the **damped wave equation**.

Modeling Exercise 5.3.8 Suppose the string with position $u(x,t)$ that obeys (8.146) has fixed ends, so $u(0,t) = u(L,t) = 0$ for all $t > 0$.

- (a) Argue that if $u(x,t) = X(x)T(t)$ is a separable solution to (8.146) then

$$\begin{aligned} X''(x) + \alpha^2 X(x) &= 0, \quad 0 < x < L \\ T''(t) + \beta T'(t) + c^2 \alpha^2 T(t) &= 0, \quad t > 0 \end{aligned}$$

for some constant α .

- (b) Further argue that if the separable solution in part (a) satisfies $u(0,t) = u(0,L) = 0$ then $\alpha = k\pi/L$ for some integer k and $X(x) = C \sin(k\pi x)$ for some constant C . Show also that $T(t)$ then satisfies

$$T''(t) + \beta T'(t) + \frac{c^2 k^2 \pi^2}{L^2} T(t) = 0. \quad (8.147)$$

- (c) Take $L = 5$, $c = 3$, and $\beta = 1$. Show that (8.147) is underdamped for each $k = 1, 2, 3, \dots$ and for each k find a real-valued general solution to (8.147) of the form

$$T(t) = A e^{-\gamma t} \cos(\omega_k t) + B e^{-\gamma t} \sin(\omega_k t)$$

for constants γ and ω_k (ω_k will depend on k , but γ doesn't). Here A and B are arbitrary constants.

- (d) Based on parts (b) and (c) explain why

$$u(x,t) = \sum_{k=1}^{\infty} (a_k e^{-\gamma t} \cos(\omega_k t) \sin(k\pi x/L) + b_k e^{-\gamma t} \sin(\omega_k t) \sin(k\pi x/L))$$

should provide a general solution to the damped wave equation (8.146) with boundary data $u(0,t) = u(L,t) = 0$ for any choice of constants a_k and b_k .

- (e) Mimic the procedure in which we used (8.113) and (8.115) in (8.116), in this case to find a solution to the damped wave equation with initial data $u(x,0) = f(x)$ and $\frac{\partial u}{\partial t}(x,0) = g(x)$. Specifically, how should a_k and b_k be chosen here? Hint: b_k is not quite the same.
(f) Solve the damped wave equation with initial data $f(x) = \sin(\pi/5)$ and $g(x) = 5 \sin(2\pi x/5)$; in this case the Fourier expansions for f and g have only one term each. Plot the solution at times $t = 0, 2, 4$, and 8 . Does the solution seem reasonable for a damped string?

8.5.4 Project: Frequency Analysis of Signals

In this project we consider how Fourier series methods can be used to analyze the frequencies present in time-dependent signals like audio data, the motion of a spring-mass system, or sunspot data. By determining what frequencies make up a signal we can glean information about the underlying physical process that produced the signal, or alter the signal to attain some goal, for example, remove noise.

Audio Signals

Consider an audio signal recorded by a microphone. Such a signal can be represented mathematically by a function $f(t)$, where t denotes time, which for this project we will assume is measured in seconds unless otherwise noted. The function f itself would quantify the sound intensity at the microphone, as pressure variations in the air relative to some reference pressure (although these pressure variations are turned into a time-varying voltage and routed to some other equipment like a computer).

Frequency Analysis

Suppose the audio signal $f(t)$ is defined on an interval $0 \leq t \leq T$ and is piecewise smooth. By using the methods of Section 8.2 the function $f(t)$ can be expanded into a Fourier cosine series (using t as the independent variable instead of x) as

$$f(t) = \frac{c_0}{2} + \sum_{k=1}^{\infty} c_k \cos(k\pi t/T) \quad (8.148)$$

where the Fourier cosine coefficients c_k are given by

$$c_k = \frac{2}{T} \int_0^T f(t) \cos(k\pi t/T) dt. \quad (8.149)$$

Here is an essential observation in the analysis that follows: for $k > 0$ the function $\cos(k\pi t/T)$ is periodic with period $P = 2T/k$ seconds (you should check this). That is, $\cos(k\pi t/T)$ is a cosine wave with frequency $k/(2T)$ hertz, so the right hand side of (8.148) is a superposition of cosine waveforms with frequencies $0, 1/(2T), 2/(2T), 2/(3T), \dots$ hertz. The expansion (8.148) shows how to synthesize $f(t)$ as a superposition of these cosine waveforms. The value of c_k as given by (8.149) indicates exactly how much of the cosine waveform at frequency $k/(2T)$ is needed.

Modeling Exercise 5.4.1 Consider the function $f(t) = t(2-t)^2 + 3 \cos(3\pi t/2)$ on the interval $0 \leq t \leq 2$. Plot this signal for $0 \leq t \leq 2$. Compute c_k for $0 \leq k \leq 10$ and then plot the pairs $(k/(2T), |c_k|)$ for $0 \leq k \leq 10$. What frequency dominates this signal? Reconcile this with the formula for f .

The cosine expansion formulas (8.148) and (8.149) give us the ability to decompose a signal $f(t)$ into its constituent frequencies and determine how much of each frequency is present in the signal. The same analysis can be done with the sine expansion (8.51)-(8.53), or using a combinations of sines and cosines as in Exercise 8.2.6, or most commonly using complex exponentials as in Exercise 8.2.10. This kind of analysis is at the heart of a great deal of modern technology. Examples include analyzing the vibration of structures, nuclear magnetic resonance (NMR) analysis, CT scans in medical imaging, almost all types of electronic audio and image processing, and as we've already seen, solving differential equations. In the remainder of this project we'll apply this analysis to some real signals.

Sampling

Unlike $f(t)$ in Modeling Exercise 5.4.1, signals are not presented to the computer as formulas, but rather as data. Specifically, the time-varying signal $f(t)$, as a voltage, would be measured to some precision at periodic time intervals, and this data would be stored in the computer. For example, audio signals are often measured at intervals of $1/44100$ th of a second. The process of measuring a continuously-varying function $f(t)$ at specified times is called **sampling**. An illustration is given in Figure 8.33 in which the function $f(t)$ from Modeling Exercise 5.4.1 is sampled on the time interval $0 \leq t \leq 2$ at times $t = 0.1, 0.3, 0.5, \dots, 1.9$. In that figure the signal is sampled 5 times per second; we say that the **sampling rate** is 5 Hz. It's clear that sampling a signal results in some loss of information about the signal. The sampling rate should be high enough that the sampled data

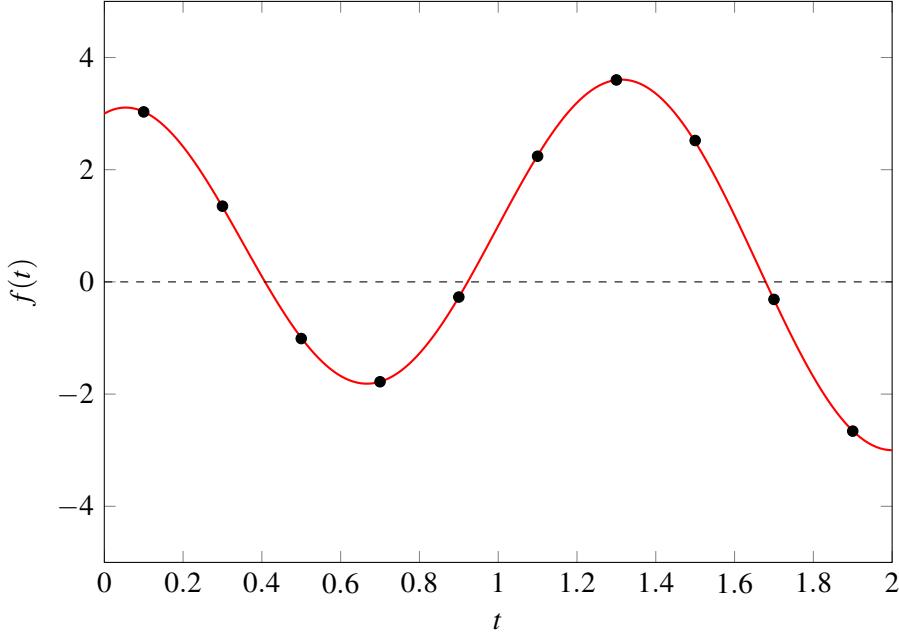


Figure 8.33: Signal $f(t) = t(2-t)^2 + 3\cos(3\pi t/2)$ (red curve) and $f(t)$ sampled at times $t = 0.1, 0.3, \dots, 1.9$ (black dots).

is a faithful representation of the underlying signal $f(t)$, but not so high that a lot of unnecessary information is collected. The appropriate sampling rate depends on the application.

Computing the Fourier cosine coefficients c_k using (8.149) requires doing an integral, but if we have a sampled version of $f(t)$ instead a formula for the function itself, how do we compute c_k ? We reach back to integral calculus and the notion of numerical integration. It will be convenient here to use the midpoint rule. Suppose the signal $f(t)$ exists on an interval $0 \leq t \leq T$ and is sampled at n times t_0, t_1, \dots, t_{n-1} where

$$t_j = \frac{(j+1/2)T}{n} \quad (8.150)$$

for $j = 0, 1, \dots, n-1$. (It will be easier to index from 0 to $n-1$ rather than 1 to n .) Let $f_j = f(t_j)$ denote the sampled value of f at time $t = t_j$. This scheme involves sampling at time intervals of length T/n seconds, so the sampling rate here is n/T Hz. This is the situation illustrated in Figure 8.33 with $T = 2$ and $n = 10$.

To approximate the value of c_k from sampled values of f we use the midpoint rule to approximate the integral on the right in (8.149). Refer to Figure 8.34, in which a graphical depiction of the midpoint rule for the signal in Figure 8.33 is shown, applied to compute the integral $\int_0^2 f(t) dt$. This would allow us to estimate c_0 . The midpoint rule evaluates the function of interest at each t_j from (8.150) to compute the height of the approximating rectangle (shown as shaded/yellow in Figure 8.34), multiplies by the base width (in this case T/n), and then adds up the results. For an arbitrary function $g(t)$ the midpoint rule approximation is

$$\int_0^T g(t) dt \approx \frac{T}{n} \sum_{j=0}^{n-1} g(t_j). \quad (8.151)$$

Applying the midpoint rule (8.151) to the integral on the right in (8.149) by using $g(t) =$

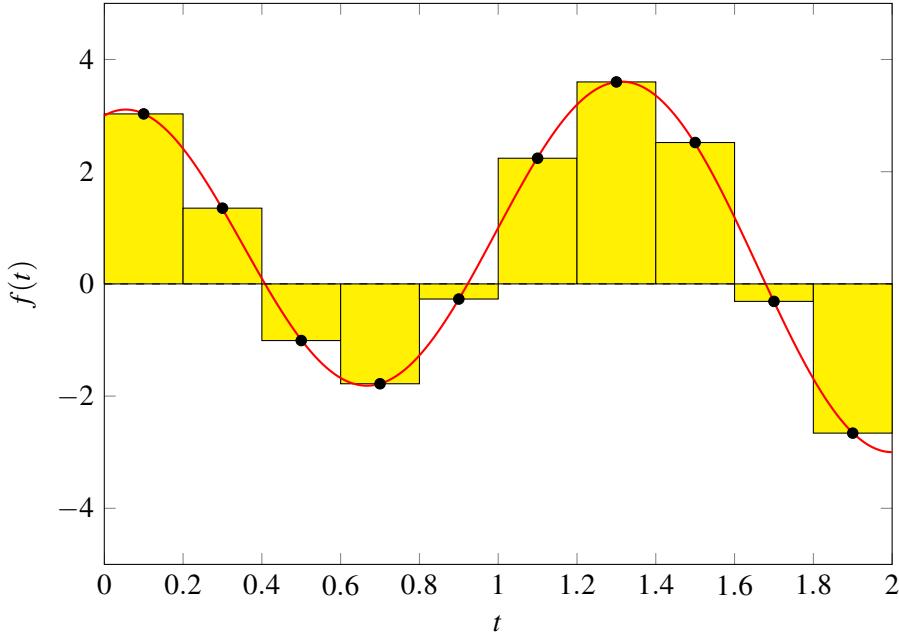


Figure 8.34: Signal $f(t) = t(2-t)^2 + 3\cos(7\pi t/2)$ (red curve) and signal sampled at times $t = 0.1, 0.3, \dots, 1.9$ (black dots), with midpoint rule approximation to $\int_0^2 f(t) dt$ (sum of the yellow rectangle areas).

$f(t) \cos(k\pi t/T)$ yields an approximation $c_k \approx C_k$ where

$$\begin{aligned} C_k &= \frac{2}{T} \frac{T}{n} \sum_{j=0}^{n-1} f(t_j) \cos(k\pi t_j/T) \\ &= \frac{2}{n} \sum_{j=0}^{n-1} f_j \cos(\pi(j+1/2)k/n) \end{aligned} \quad (8.152)$$

with $f_j = f(t_j)$. This allows us to use the data f_j to decompose the sampled signal into a sum of the basic sampled cosine waveforms and so estimate the frequency content of the underlying signal $f(t)$.

Modeling Exercise 5.4.2 Let $f(t) = t(2-t)^2 + 3\cos(3\pi t/2)$ on the interval $0 \leq t \leq 2$ (this is $f(t)$ from Modeling Exercise 5.4.1). Take $n = 4$ in (8.152), so that from (8.150) we have $t_0 = 0.25, t_1 = 0.75, t_2 = 1.25$, and $t_3 = 1.75$.

- (a) Compute f_0, f_1, f_2 , and f_3 to at least four significant figures, then use (8.152) to compute each of C_0, C_1, C_2 , and C_3 . Compare to the true values obtained from (8.149).
- (b) Compute C_4, C_5, \dots, C_{16} . Show that in this case $C_4 = 0$, while $C_5 = -C_3, C_6 = -C_2, C_7 = -C_1$, and $C_8 = -C_0$. Also observe that $C_9 = C_7, C_{10} = C_6, \dots, C_{15} = C_1$, and finally $C_{16} = C_0$.

Modeling Exercise 5.4.2 illustrates a more general phenomena. If a signal f is sampled at n points then we can estimate $c_k \approx C_k$ for $0 \leq k \leq n-1$ with some reliability. But $C_n = 0$ is always true, so it is not likely to be a good estimate of c_n unless c_n just happens to be close to zero. And the C_k for $k \geq n$ can be computed from the C_k in the range $0 \leq k \leq n-1$, and so may have nothing to do with the correct value of the corresponding coefficient. In general, when given sampled signal values f_0, \dots, f_{n-1} , we can use the sampled data to estimate the Fourier cosine coefficients c_k with some reliability only in the range $0 \leq k \leq n-1$.

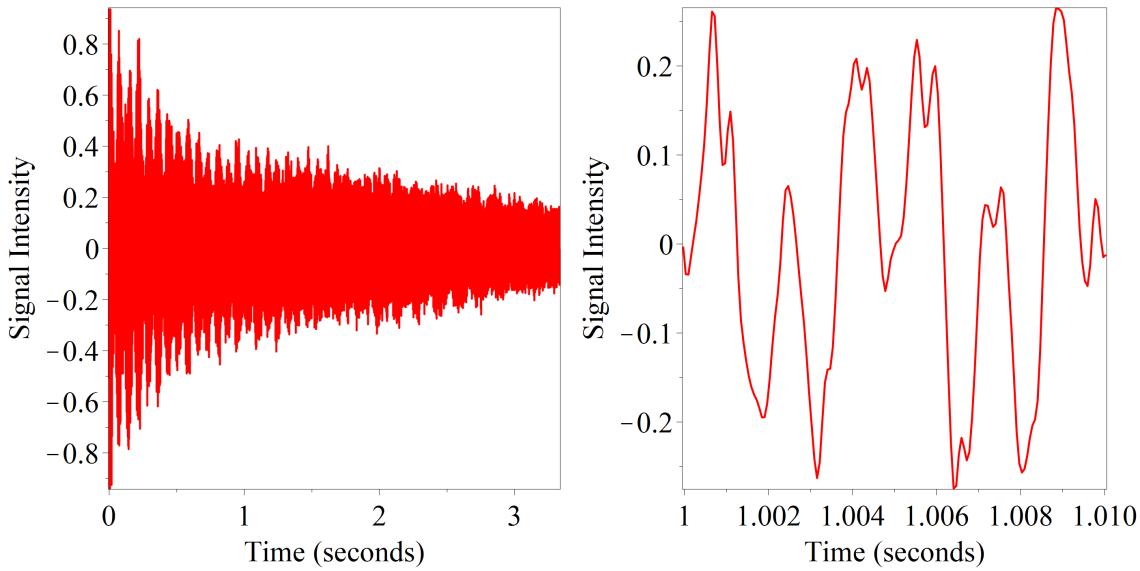


Figure 8.35: Left panel: a plot of the gong audio data for $0 \leq t \leq 3.3315$ seconds. Right panel: a plot of the gong audio data on the time interval $1.0 \leq t \leq 1.01$ seconds.

Analysis of an Audio Signal

In the supplied code for the text at [8] are contained three sampled signals, each stored in an Excel worksheet, as well as template analysis code in Maple, Mathematica, Matlab, and Sage. One is an audio signal, a recording of the sound of a round metal dinner gong that hangs in the author’s kitchen. The other comes from the spring-mass project “Parameter Estimation with Second-Order ODEs” from Section 4.6.3. The last data set concerns annual sunspot activity over the past several hundred years, obtained from <https://wwwbis.sidc.be/silso/datafiles>.

Let’s consider the analysis of the gong sound. You may wish to refer to the appropriate code as you read. The sample gong audio signal consists of $n = 53304$ samples in the time interval $0 < t < T$ with $T = 3.3315$ seconds, a sampling rate of $r = 16$ kHz. The time-signal pairs (t_j, f_j) for $0 \leq j \leq 53303$ are shown in the left panel of Figure 8.35 on the entire time range, although with so many data points it’s hard to see detail. The right panel in Figure 8.35 zooms in one the interval $1 \leq t \leq 1.01$ seconds. There is also an audio file “gong_sound.wav” that you can play to hear the sound.

From the $n = 53304$ sampled values for the gong audio signal we can use (8.152) to compute the C_k , estimates of the Fourier cosine coefficients c_k in (8.149), for each k in the range $0 \leq k \leq n - 1$. The only roadblock is that the computation for each C_k involves a sum of n terms and a corresponding number of multiplications—a lot of numerical operations, comparable to $2n^2 \approx 5.68 \times 10^9$. To facilitate this we will make use of a built-in (or supplied) command called “dct” or some variation thereof. The dct command is based on an algorithm called the **fast Fourier transform** (commonly abbreviated as **FFT**), one of the most important algorithms developed in the past century. See [26, 25] for more information. The dct command reduces the computation from something proportional to n^2 to something proportional to $n \log(n)$, which is a huge savings in computation and time. The details of the FFT algorithm and computations do not concern us though, only the results—the C_k .

In the left panel of Figure 8.36 we show a plot of the magnitude $|C_k|$ versus frequency $k/(2T)$ for $0 \leq k \leq n - 1$. Recall that C_k estimates c_k , the amount of the $\cos(k\pi t/T)$ waveform with frequency $k/(2T)$ Hz present in the signal f . The highest frequency estimated is thus $(n - 1)/(2T)$; since $T = n/r$ where $r = 16000$ is the sampling rate we see that this highest frequency is $\frac{n-1}{2(n/r)} =$

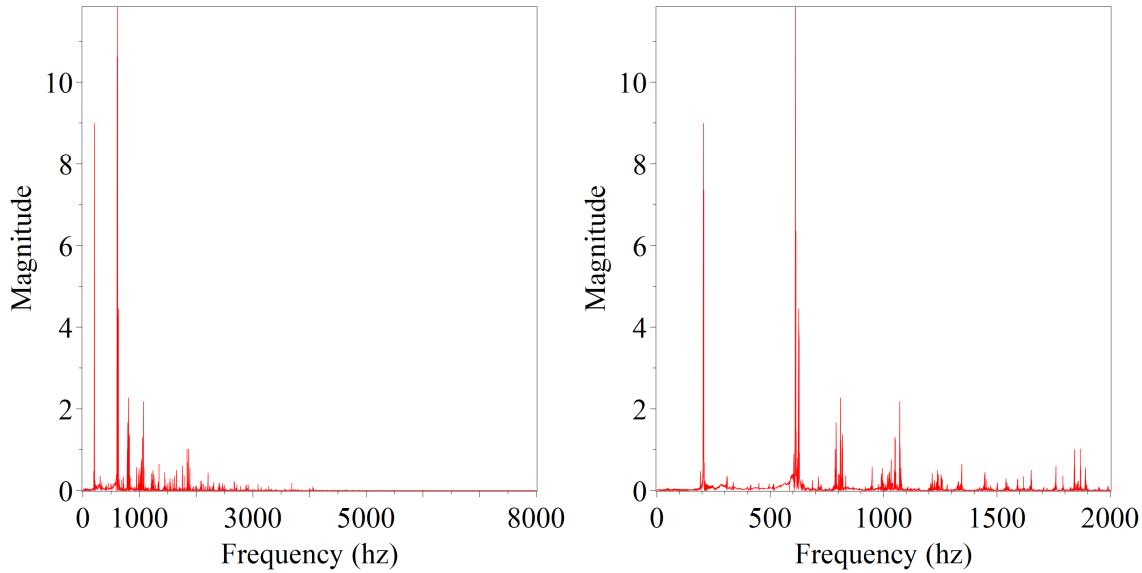


Figure 8.36: Left panel: a plot of the frequency-amplitude pairs $(k/(2T), |C_k|)$ for $0 \leq k \leq n - 1$ (frequency range 0 to 8 kHz). Right panel: same, but for $0 \leq k \leq 13335$ (frequency range 0 to 2 kHz).

$\frac{(n-1)r}{2n} \approx r/2 = 8000$ Hz here. More generally the highest frequency we can estimate is about $r/2$, half the sampling rate. This is called the **Nyquist frequency**.

The left panel in Figure 8.36 makes it quite clear that certain frequencies, those corresponding to large values for $|C_k|$, dominate this signal, though the frequencies above about 3000 Hz seem to contribute little. The right panel in Figure 8.36 focuses on the range 0 to 2000 Hz. We can see a very large peak around 611 Hz, and another around 206 Hz. Other smaller peaks occur at 811 Hz, 1072 Hz, and elsewhere. This type of analysis lets us pick apart a signal into its constituent frequencies and identify the most dominant. In many cases this can be used to reveal important physical parameters about the underlying physical process, as you will see in Modeling Exercise 5.4.3.

At this point we should point out that the `dct` command in the various software packages actually uses the sampled data f_0, \dots, f_{n-1} and returns the quantities

$$\begin{aligned} C_0 &= \frac{\sqrt{1}}{\sqrt{n}} \sum_{j=0}^{n-1} f_j \\ C_k &= \frac{2}{\sqrt{n}} \sum_{j=0}^{n-1} f_j \cos(\pi(j+1/2)k/n) \text{ for } 1 \leq k \leq n-1. \end{aligned} \tag{8.153}$$

This is a minor variation on (8.152). For $k \geq 1$ the C_k have been scaled by \sqrt{n} , while C_0 has been scaled by $\sqrt{n}/2$. This has little bearing on the type of analysis we do, since each coefficient is still proportional to the amount of $\cos(k\pi t/T)$ wave form in the signal. The scaling in (8.153) is the conventional definition of the **discrete cosine transform** (DCT), hence the command name “`dct`.” The DCT takes in sampled data f_0, \dots, f_{n-1} and returns coefficients C_0, \dots, C_{n-1} . It’s easy to see that the coefficient C_0 is \sqrt{n} times the mean value of the signal samples.

Application to Spring-Mass Parameter Estimation

Modeling Exercise 5.4.3 You should begin by reading/rereading everything in the project “Parameter Estimation with Second-Order ODEs” prior to Modeling Exercise 6.3.1 in Section 4.6.3. In

brief, the goal here is to use sampled position data to analyze the oscillatory motion of the spring-mass system in that project. Although this is not an audio signal, precisely the same principles apply. We will make use of the Maple, Mathematica, Matlab, or Sage script/worksheet entitled “dct_analysis”. Although our analysis indexes the samples f_j from $j = 0$ to $j = n - 1$, most software indexes them from $j = 1$ to $j = n$. The C_k will also be indexed from $k = 1$ to $k = n$. Keep this minor annoyance in mind, but the worksheets take this alternate indexing into account.

- (a) Load in the data from the Excel data file “spring_mass_data_fourier.xls”. This data consists of measurements of the mass position every 0.02 seconds, a sampling rate of $r = 50$ Hz. There are $n = 1460$ data points spanning $T = 29.2$ seconds. Plot the data. Note that the data is not centered vertically on position $y = 0$, something we had to adjust for in Section 4.6.3. It will no longer be an explicit concern.
- (b) Compute the DCT of the sampled signal to produce the coefficients C_k . Because the sampling rate is 50 Hz, the coefficient C_n corresponds to a frequency of about 25 Hz. Plot the DCT coefficient magnitudes. The coefficient C_0 (indexed in the software as $C[1]$) should be much larger in magnitude than all the other C_k . This is because the sampled data has a mean value that is not zero. Redo the plot starting with C_1 or C_2 , to exclude C_0 . Based on the plot, what is the dominant frequency in the vibration of this mass, in hertz?
- (c) Given that a lightly-damped spring mass system with mass m and spring constant k oscillates with a radial frequency close to $\omega = \sqrt{k/m}$, use the dominant frequency from part (b) to estimate ω . The mass m in this system was measured to be 0.2 kg. Use this and the estimated value for ω to estimate the spring constant k .

Analysis of Sunspot Data

Modeling Exercise 5.4.4 The number of sunspots that appear on the surface of the sun each year has been observed to vary periodically for centuries and data concerning this phenomena goes back to at least the 1600s. This variation is now known to be an effect of the sun’s periodic reversal of its magnetic field and is called the **solar activity cycle**. Let us examine some data related to this phenomena.

- (a) Load in the data from the Excel data file “sunspots.xlsx”, available at the text website. This data is from [96]. The data consists of the average daily number of observed sunspots on the surface of the sun for each year from 1700 to 2020, so we have $n = 321$ sampled data points with a sampling rate of $r = 1$ (units are reciprocal years) spanning a time interval of $T = 320$ years. Plot the data; you can consider the first data point at the year 1700 as time $t = 0$.
- (b) Compute the DCT of the sampled signal to produce the coefficients C_k . To what frequency does C_n correspond (in units of “per year”)? Plot the DCT coefficient magnitudes. The coefficient C_0 should have a much larger magnitude than all the others since the sampled data has a fairly large mean value. To better see the magnitude of the DCT coefficients, redo the plot starting with C_1 or C_2 , to exclude C_0 . Based on the plot, what is the dominant frequency in the sun’s solar activity cycle? What is the period of this cycle?

8.5.5 Project: It’s All Relative

The year 1905 was Albert Einstein’s *annus mirabilis*, his “year of miracles” in which he published four landmark physics papers: an examination of Brownian motion, the development of the special theory of relativity, the derivation of his famous mass-energy equivalence $E = mc^2$, and an analysis of the photoelectric effect. This last paper is the primary work that won him the 1921 Nobel Prize in physics. It is special relativity that concerns us in this project.

To set the stage, see Figure 8.37, in which we consider the situation in one space dimension. The observer standing on the foundational gray slab (“Observer 1”) measures spatial positions using an x coordinate, with $x = 0$ at some fixed location, say directly beneath this observer, and $x > 0$ to the right. Observer 1 measures time using a variable t . Another observer, depicted by the

person on the cart moving to the right at speed v (“Observer 2”), measures all positions using a spatial coordinate y that is centered on the middle of the cart, with $y > 0$ to the right. Observer 2 measures time using a variable s . The set of all possible positions in space and moments in time is called **spacetime**. Observer 1 standing on the foundational slab thus quantifies the location and time of any event in spacetime using x and t , respectively. Observer 2 on the cart uses y and s .

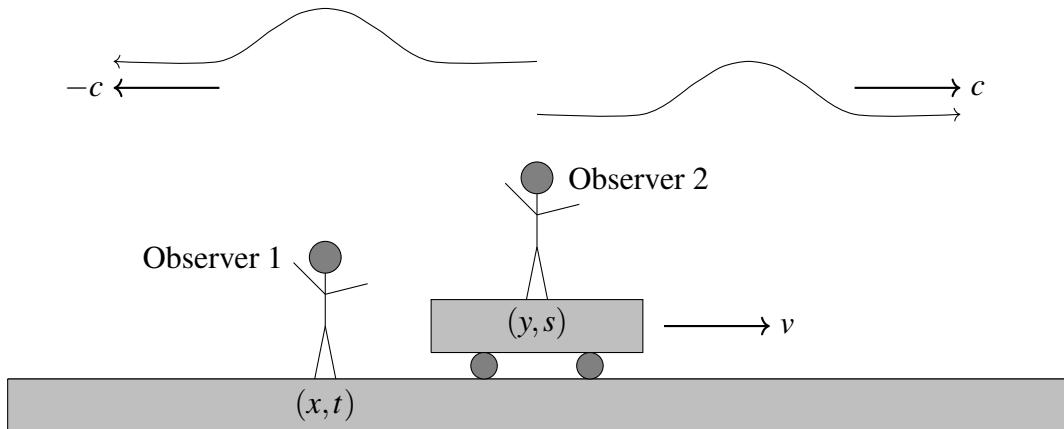


Figure 8.37: Two observers, one with spacetime coordinate system (x, t) (“Observer 1”), the other with coordinate system (y, s) (“Observer 2”). The (x, t) observer is stationary with respect to the gray slab. Overhead, waves propagate at speed c to the left and right, at least with respect to Observer 1.

The Galilean World

Let’s suppose the observers synchronize watches so that $t = 0$ corresponds to $s = 0$, and that both observers use the same units for time, e.g., seconds, as well as the same unit for distance, e.g., meters. In our usual and intuitive framework for how time passes this means that $s = t$ for all s and t . We also suppose that the cart is moving in such a manner that the spatial coordinate $y = 0$ coincides with $x = 0$ at time $t = s = 0$. Physically, both observers are in the same position at time $s = t = 0$.

Modeling Exercise 5.5.1 We’ve already decided that $s = t$, so the observers will register the same time for any given event in spacetime, and that when $s = t = 0$ their coordinate systems coincide, so $x = y$. Argue that the relation between x and y at other times is

$$y = x - vt \text{ or } x = y + vs. \quad (8.154)$$

The relation (8.154) is called a **Galilean transformation** between these coordinate systems.

Suppose that, as depicted above the heads of the observers in Figure 8.37, waves propagate to the left or right at speed c as seen by Observer 1, in accordance with the wave equation. The quantity c here does not yet need to be the speed of light. The waves could be in a string stretched overhead. As we’ve seen, any superposition of such waves again satisfies the wave equation. Let $u(x, t)$ denote the height of the wave at coordinates (x, t) as measured by Observer 1. Observer 2 will measure the same height, but at the corresponding coordinates (y, s) given by (8.154). To avoid future confusion, let us use $w(y, s)$ to denote the height of the wave at coordinates (y, s) as measured by Observer 2, so $u(x, t) = w(y, s)$ if (x, t) and (y, s) are in the relation (8.154).

Modeling Exercise 5.5.2 Observer 1 sees waves that obey the wave equation with wave speed c , and so from Observer 1's perspective the wave displacement obeys

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0.$$

Based on the above discussion we have $u(x, t) = w(y, s) = w(x - vt, t)$, where we have used (8.154) and the fact that $s = t$. Use $u(x, t) = w(x - vt, t)$ and the chain rule to show that

$$\begin{aligned}\frac{\partial u}{\partial x}(x, t) &= \frac{\partial w}{\partial y}(y, s) \\ \frac{\partial^2 u}{\partial x^2}(x, t) &= \frac{\partial^2 w}{\partial y^2}(y, s) \\ \frac{\partial u}{\partial t}(x, t) &= -v \frac{\partial w}{\partial y}(y, s) + \frac{\partial w}{\partial s}(y, s) \\ \frac{\partial^2 u}{\partial t^2}(x, t) &= v^2 \frac{\partial^2 w}{\partial y^2}(y, s) - 2v \frac{\partial^2 w}{\partial y \partial s}(y, s) + \frac{\partial^2 w}{\partial s^2}(y, s).\end{aligned}\tag{8.155}$$

Then use (8.155) and the fact that u satisfies the wave equation for all x and t to conclude that $w(y, s)$ satisfies

$$\frac{\partial^2 w}{\partial s^2} - 2v \frac{\partial^2 w}{\partial y \partial s} + (v^2 - c^2) \frac{\partial^2 w}{\partial y^2} = 0\tag{8.156}$$

for all y and s .

Observer 1 sees a world in which the motion of waves is governed by the usual wave equation with solutions that are a superposition of waves moving to the left and right with wave speed c . But Observer 2 sees waves governed by equation (8.156). Note that (8.156) coincides with the usual wave equation only when $v = 0$.

Modeling Exercise 5.5.3 Show that we can factor the second-order differential operator on the left in (8.156) into two first-order differential operators, as

$$\frac{\partial^2}{\partial s^2} - 2v \frac{\partial^2}{\partial y \partial s} + (v^2 - c^2) \frac{\partial^2}{\partial y^2} = \left(\frac{\partial}{\partial s} - (v - c) \frac{\partial}{\partial y} \right) \left(\frac{\partial}{\partial s} - (v + c) \frac{\partial}{\partial y} \right).\tag{8.157}$$

(See also Exercise 8.4.5.) Conclude that a solution to either of the advection equations

$$\begin{aligned}\frac{\partial w}{\partial s} + (c - v) \frac{\partial w}{\partial y} &= 0 \\ \frac{\partial w}{\partial s} - (c + v) \frac{\partial w}{\partial y} &= 0\end{aligned}\tag{8.158}$$

is a solution to (8.157) (as well as any superposition). As we know from Section 8.4.2, solutions to the first equation in (8.158) are waves moving to the right at speed $c - v$ (though if $v > c$, the wave will appear to move left). Solutions to the second equation in (8.158) are waves moving the left at speed $c + v$ (or to the right at speed $-c - v$). Why does each of these conclusions make intuitive sense? Hint: consider the case in which $v = c$, or in which $v = -c$. What should Observer 2 see in each case?

The Relativistic World

In a Galilean world the appearance of the wave equation varies between observers in different frames of reference, like Observers 1 and 2 above. Maxwell's equations, developed over the course of the 19th century, established that electromagnetic waves, like light, should obey the wave equation with speed c , the speed of light. If a Galilean transformation governed the coordinate transformation between two different frames of reference then the speed of light would also seem to vary according to the relative motion of the observers through space. In the scenario above Observer 1, who seems to be in a privileged position of absolute rest, sees light propagate at speed c in both directions. Observer 2 sees the speed of light vary according to which direction the wave is moving. Unfortunately, experimental evidence in the late 19th century, for example, the Michelson-Morley experiment, indicated that the speed of light always appears as a fixed value, regardless of observer motion. This means that a Galilean transformation relating the coordinate systems (x, t) and (y, s) cannot be correct in our universe. We need a transformation in which both observers see light waves obeying the wave equation and always moving at speed c .

To address this issue Lorentz posited a relation between (x, t) and (y, s) of the form

$$\begin{aligned} s &= (t - vx/c^2)/\gamma \\ y &= (x - vt)/\gamma \end{aligned} \tag{8.159}$$

where c is the speed of light in a vacuum and $\gamma = \sqrt{1 - v^2/c^2}$ is the **Lorentz factor**, a constant that depends on v and c (in particular, the ratio v/c). Equations (8.159) are known as the **Lorentz transformation**. We can also solve for x and t in terms of y and s as

$$\begin{aligned} t &= (s + vy/c^2)/\gamma \\ x &= (y + vs)/\gamma. \end{aligned} \tag{8.160}$$

Note that v in (8.159) is replaced by $-v$ in (8.160), since the coordinate frames are moving in opposite directions. The transformations here are arranged so that $(x, t) = (0, 0)$ corresponds to $(y, s) = (0, 0)$. One assumption that should be mentioned is that we assume the coordinate systems (x, t) and (y, s) are **inertial**, that is, neither is undergoing acceleration. See [55] for more details.

Einstein put the Lorentz transformation on a coherent foundation that rested on rather general physical principles, for example, that there is no preferred frame of reference, and that the speed of light is the same in all reference frames. This forms the basis of the special theory of relativity. Others, for example Poincaré, had been pursuing similar ideas. The peculiar relationships (8.159) and (8.160) that transform between coordinate systems leave the wave equation unchanged. Observers in any inertial coordinate system see waves that travel at speed c , as you can verify in Modeling Exercise 5.5.4.

Modeling Exercise 5.5.4

- (a) Suppose the $u(x, t)$ is a solution to the advection equation $\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0$ (of course then u also satisfies the wave equation). This right-moving wave quantified by u is what Observer 1 sees, and we know from Section 8.4.2 that such a wave can be described as

$$u(x, t) = \phi(x - ct)$$

for some function ϕ . According to (8.160) Observer 2 sees a wave given by

$$w(y, s) = u(x, t) = \phi(x - ct) = \phi((y + vs)/\gamma - (cs + vy/c)/\gamma). \tag{8.161}$$

Use (8.161) along with the chain rule to show that w satisfies the advection equation

$$\frac{\partial w}{\partial s} + c \frac{\partial w}{\partial y} = 0.$$

(It won't matter what ϕ is.) Thus Observer 2 also sees a wave moving to the right at speed c .

- (b) Repeat part (a) but with u quantifying a wave moving to the left at speed c , so start with $\frac{\partial u}{\partial t} - c \frac{\partial u}{\partial x} = 0$. Show that Observer 2 also see a wave moving to the left at speed c .
- (c) Conclude that since every solution to the wave equation consists of a superposition of left- and right-moving waves (that is, $u(x, t) = \phi(x + ct) + \psi(x - ct)$) Observer 2 will also always see a superposition of waves that move to the left or right at speed c , regardless of v . That is, Observer 2 will see waves that obey the wave equation with speed c .

The Lorentz transformation is an unavoidable consequence of the observed properties of light and the requirement that there is no preferred frame of reference for observers, but the Lorentz transformation has a number of strange consequences. For example, observers in motion relative to each other will measure the passage of time at different rates, or that the dimensions of an object may appear to change (contract) depending on its motion relative to an observer, or that the notion of two events occurring simultaneously makes no sense. See [55] for an exploration of these ideas.

Modeling Exercise 5.5.5 Show that if $|v| \ll c$ ($|v|$ is much less than c , or as $v \rightarrow 0$) the Lorentz transformation (8.159) becomes the Galilean transformation (8.154) (along with $s \approx t$).

A. Complex Numbers

This appendix offers a brief introduction to the essentials of complex numbers necessary for analyzing differential equations.

A.1 Motivation and Definition

Solvability of Quadratic Equations

Equations of the form $ax + b = 0$ where a and b are real numbers with $a \neq 0$ are ubiquitous in mathematics and easily solvable for x , with unique solution $x = -b/a$. Quadratic equations $ax^2 + bx + c = 0$ in which a, b , and c are real numbers with $a \neq 0$ are also common and can be solved by using the **quadratic formula**. The solutions are $x = r_1$ and $x = r_2$ where

$$r_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a},$$
$$r_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

The numbers r_1 and r_2 are called the roots of the quadratic polynomial $ax^2 + bx + c$. If $b^2 - 4ac > 0$ then these roots are real numbers and $r_1 \neq r_2$. Moreover the factorization

$$ax^2 + bx + c = a(x - r_1)(x - r_2)$$

holds. In the case that $b^2 - 4ac = 0$ we obtain $r_1 = r$ and $r_2 = r$ where $r = -\frac{b}{2a}$; this is the situation in which the quadratic equation has a double root at $x = r$. The factorization $ax^2 + bx + c = a(x - r)^2$ holds. In each case, $b^2 - 4ac > 0$ and $b^2 - 4ac = 0$, the quadratic polynomial $ax^2 + bx + c$ has two roots, if a double root is counted as two roots (that happen to equal each other).

The case in which $b^2 - 4ac < 0$ presents a difficulty because $\sqrt{b^2 - 4ac}$ has no meaning as a real number. (If $v = \sqrt{b^2 - 4ac}$ then $v^2 = b^2 - 4ac < 0$, but the basic rules of arithmetic for the real numbers forbid $v^2 < 0$ for any real number v .) In this case the quadratic equation $ax^2 + bx + c = 0$ has no solutions in the real numbers. In order to solve such quadratic equations we must enlarge our universe of numbers.

Definition of Complex Numbers

An example of a quadratic equation that has no real solutions is $x^2 + 1 = 0$ or

$$x^2 = -1. \quad (\text{A.1})$$

But it turns out to be of enormous value to declare that the solutions to this equation are $\sqrt{-1}$ and $-\sqrt{-1}$, much as the equation $x^2 = 2$ has solutions $\sqrt{2}$ and $-\sqrt{2}$. Rather than writing $\sqrt{-1}$ it is conventional to define $i = \sqrt{-1}$, so the solutions to (A.1) are $x = i$ and $x = -i$. (In electrical engineering it is common to use j and $-j$.) More generally, we define the set of complex numbers as an extension of the real numbers as follows.

Definition A.1.1 — Complex Numbers. A complex number is a quantity z of the form $z = x + yi$ where x and y are real numbers, with the convention that $i^2 = -1$.

The set of complex numbers is denoted by the symbol \mathbb{C} and sometimes referred to as the **complex plane**, since each complex number $x + yi$ can be identified in a natural way with the point (x, y) . Every real number x can also be considered as a complex number in a natural way, namely as $x = x + 0i$. Complex numbers of the form $0 + yi$ in which $x = 0$ are said to be purely imaginary. As obvious as it may seem, the complex numbers $z = x + yi$ and $w = u + vi$ are equal and we write $z = w$ precisely when $u = x$ and $v = y$, so that w and z have the same real and imaginary parts.

The advantage of working in the complex numbers is that every quadratic polynomial will have two roots, either real and distinct or a double root as above, or both complex. More generally, any n th-degree polynomial will turn out to have n roots in the complex plane, as will be discussed shortly.

Terminology and Notation

Real and Imaginary Parts

If $z = x + yi$ then the real number x called the **real part** of z and the real number y is called the **imaginary part** of z . The notation $x = \operatorname{Re}(z)$ and $y = \operatorname{Im}(z)$ is common. Thus, for example, $z = 2 + 4i$ is a complex number with $\operatorname{Re}(z) = 2$ and $\operatorname{Im}(z) = 4$. Note that the imaginary part of z does not include the i , but only the real coefficient 4 in front of i .

Modulus and Conjugate

The **modulus** of a complex number $z = x + yi$ is the real number

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

Thus, for example, if $z = 2 + 4i$ then $|z| = \sqrt{2^2 + 4^2} = \sqrt{20} = 2\sqrt{5}$. In the special case that $z = x$ is a real number it follows that $|z| = |x|$, so for real numbers the modulus is just the absolute value. The modulus provides a way to quantify the size of a complex number, much as absolute value does for real numbers. It is easy to check that $|z| \geq 0$ always, and $|z| = 0$ if and only if $z = 0 + 0i$.

The **conjugate** of a complex number $z = x + yi$ is the complex number $\bar{z} = x - yi$. It's straightforward to check that a complex number z is real if and only if $z = \bar{z}$.

Reading Exercise A.1.1 If $z = -3 + 5i$, compute $\operatorname{Re}(z)$, $\operatorname{Im}(z)$, \bar{z} , and $|z|$.

A.2 Arithmetic with Complex Numbers

Addition and Subtraction

Let $z = a + bi$ and $w = c + di$ be complex numbers. The sum $z + w$ is defined as

$$z + w = (a + c) + (b + d)i$$

while the difference $z - w$ is defined as

$$z - w = (a - c) + (b - d)i.$$

Addition is easily seen to be commutative, that is, $z + w = w + z$. The complex number $0 + 0i$ is usually denoted by 0, and $0 + z = z + 0 = z$ for any complex number z .

Multiplication

Multiplication is defined as

$$\begin{aligned} zw &= (a + bi)(c + di) \\ &= ac + adi + bic + (bi)(di) \quad (\text{FOIL the previous product}) \\ &= ac + adi + bci + bdi^2 \\ &= (ac - bd) + (ad + bc)i \quad (\text{use } i^2 = -1, \text{ group terms}). \end{aligned} \tag{A.2}$$

In the computation above we made certain assumptions about real numbers commuting with i under multiplication, e.g., that $bic = bci$. In summary, the product zw is the complex number with real part $ac - bd$ and imaginary part $ad + bc$.

It is straightforward to verify that multiplication is commutative, so $zw = wz$. Moreover if z, w , and u are all complex numbers then multiplication distributes over addition, $u(z + w) = uz + uw$. It is worth noting that if $z = a + bi$ then the conjugate of z is $\bar{z} = a - bi$ and one can easily verify that $z\bar{z} = a^2 + b^2$ is a nonnegative real number that equals $|z|^2$.

Division

Let's first consider the problem of reciprocating a complex number $w = a + bi$. The goal is to express the real and imaginary parts of $1/w = 1/(a + bi)$ as explicitly as possible. To do this let $z = 1/w$ so that $wz = 1$. Suppose $z = x + yi$ so that $wz = 1$ becomes $(a + bi)(x + yi) = 1 + 0i$ (writing $1 + 0i$ for 1). Use the definition (A.2) to compute the product $(a + bi)(x + yi) = (ax - by) + (ay + bx)i$, set this equal to $1 + 0i$, then match real and imaginary parts to find

$$ax - by = 1 \quad \text{and} \quad ay + bx = 0.$$

This is two equations in real unknowns x and y , with solution $x = a/(a^2 + b^2)$, $y = -b/(a^2 + b^2)$. The reciprocal $1/(a + bi)$ can thus be expressed explicitly as

$$\frac{1}{a + bi} = \frac{a}{a^2 + b^2} - \frac{b}{a^2 + b^2}i. \tag{A.3}$$

With $w = a + bi$ it's not hard to check that this can also be written as $1/w = \bar{w}/|w|^2$.

The quotient z/w of two complex numbers $z = c + di$ and $w = a + bi$ can be computed as $z/w = z(1/w) = (c + di)(1/(a + bi))$, the product of z and the reciprocal of w as defined in (A.3). A bit of algebra then leads to

$$\frac{z}{w} = \frac{ac + bd}{a^2 + b^2} + \frac{ad - bc}{a^2 + b^2}i. \tag{A.4}$$

This provides a definition for complex division. If $w = a + bi$ is not zero then $|w|^2 = a^2 + b^2 > 0$ so the quotient z/w as given by (A.4) is defined.

With these definitions of addition, subtraction, multiplication, and division, the basic arithmetic properties of the complex numbers are essentially identical to those that you are familiar with for the real numbers.

Reading Exercise A.2.1 Let $w = 1 + 2i$ and $z = 2 - i$. Compute each of $z + w$, $z - w$, zw , and z/w .

Square Roots

If $a > 0$ is a real number then $\sqrt{-a}$ can be defined as a complex number as follows. Let $z = x + yi$; if $z = \sqrt{-a}$ then $z^2 = -a$, which means $(x + yi)^2 = -a$. After expanding $(x + yi)^2$ we obtain $(x^2 - y^2) + 2xyi = -a + 0i$, with the placeholder $0i$ to emphasize that the imaginary part of $-a$ is zero. Matching the real and imaginary parts of z^2 and $-a$ leads to the equations

$$x^2 - y^2 = -a \quad \text{and} \quad 2xy = 0$$

for x and y , the real and imaginary parts of z . We require each of x and y to be a real number. To solve this system, first note that the equation $2xy = 0$ means that either $x = 0$ or $y = 0$. If $y = 0$ then $x^2 - y^2 = -a$ becomes $x^2 = -a$, which has no real solution. If $x = 0$ then $x^2 - y^2 = -a$ becomes $-y^2 = -a$ or $y^2 = a > 0$ with two real solutions $y = \sqrt{a}$ and $y = -\sqrt{a}$. In summary, if $z = x + yi$ satisfies $z^2 = -a$ then $z = i\sqrt{a}$ or $z = -i\sqrt{a}$. Either of these choices might have claim to being called the square root of $-a$, but the conventional choice is $i\sqrt{a}$. Thus if $a > 0$ is a real number we define

$$\sqrt{-a} = i\sqrt{a}.$$

This is similar to the situation with square roots for positive real numbers: if $a > 0$ then the convention is that \sqrt{a} is the positive solution to $x^2 = a$.

More generally one can define the square root for any complex number $w = a + bi$, by solving $z^2 = w$. If $z = x + yi$ this comes down to solving $x^2 - y^2 = a$ and $2xy = b$ for x and y , both real. There are always two solutions for z , unless $w = 0$ (in which case $x = y = 0$ is the only solution, so $z = 0$). These solutions differ only in sign. Which solution is chosen as $\sqrt{a+bi}$ must be specified, although we won't need to concern ourselves with this more general case.

A.3 Exponentiation of Complex Numbers

Definition Using Taylor Series

Complex numbers can be exponentiated. Let us begin by defining e^{bi} where b is a real number. Recall from basic calculus that if y is any real number then e^y has a Taylor series based at $y = 0$,

$$1 + y + y^2/2! + y^3/3! + \cdots + y^k/k! + \cdots = \sum_{k=0}^{\infty} \frac{y^k}{k!} \tag{A.5}$$

where $k! = k(k-1)(k-2)\cdots 3 \cdot 2 \cdot 1$ and $0! = 1$. The series in (A.5) converges to e^y for any real number y .

This can be used to define e^{bi} , by inserting $y = bi$ into (A.5). It can be shown that the resulting series with complex terms

$$1 + bi + (bi)^2/2! + (bi)^3/3! + \cdots + (bi)^k/k! + \cdots$$

converges to some complex number, for any real b . This complex number is what we will call e^{bi} , so we define

$$e^{bi} = 1 + bi + (bi)^2/2! + (bi)^3/3! + \cdots + (bi)^k/k! + \cdots \tag{A.6}$$

although the series on the right hand side isn't very explicit. In the next section we make it so.

Euler's Formula

The series on the right in (A.6) can be made more explicit by noting that $i^2 = -1$, $i^3 = -i$, $i^4 = 1$, $i^5 = i$, and so on. In short, the powers of i run through the sequence $1, i, -1, -i$, and then repeat with a cycle of length four. Based on this the series in (A.6) can be written as

$$e^{bi} = 1 + bi - b^2/2! - ib^3/3! + b^4/4! + ib^5/5! + \dots$$

Regroup terms (which is permitted for this infinite sum) to obtain

$$e^{bi} = \underbrace{\left(1 - b^2/2! + b^4/4! - b^6/6! + \dots\right)}_{\text{Re}(e^{bi})} + \underbrace{\left(b - b^3/3! + b^5/5! + \dots\right)i}_{\text{Im}(e^{bi})}. \quad (\text{A.7})$$

The series that defines the real part of e^{bi} on the right in (A.7) is just the Taylor series for $\cos(b)$ from elementary calculus, and the series that defines the imaginary part is $\sin(b)$. We then have

$$e^{bi} = \cos(b) + i \sin(b) \quad (\text{A.8})$$

for any real number. This is known as **Euler's formula**.

More generally we define e^z for any complex number $z = a + bi$ by imposing the property $e^{a+bi} = e^a e^{bi}$ (analogous to $e^{x+y} = e^x e^y$ for real numbers x and y) to obtain

$$e^{a+bi} = e^a e^{bi} = e^a (\cos(b) + i \sin(b)) = e^a \cos(b) + ie^a \sin(b). \quad (\text{A.9})$$

That is, the real part of e^{a+bi} is $e^a \cos(b)$ and the imaginary part is $e^a \sin(b)$.

With this definition complex exponentiation has many of the same algebraic properties as exponentiation for real numbers. In particular $e^0 = 1$, $e^{w+z} = e^w e^z$, and $e^{-z} = 1/e^z$.

Reading Exercise A.3.1 Compute $e^{i\pi}$, $e^{i\pi/2}$, and $e^{i\pi/4}$.

Sine and Cosine

From (A.8) with $-b$ in place of b it follows that

$$e^{-bi} = \cos(b) - i \sin(b)$$

since $\sin(-b) = -\sin(b)$. Adding this equation to (A.8) (left sides and right sides) and dividing by 2 shows that

$$\cos(b) = \frac{e^{bi} + e^{-bi}}{2}.$$

Subtracting the equations and dividing by $2i$ yields

$$\sin(b) = \frac{e^{bi} - e^{-bi}}{2i}.$$

So the familiar trigonometric functions can be expressed using complex exponential functions.

A.4 The Fundamental Theorem of Algebra

Polynomials and Roots

An n th-degree polynomial $p(z)$ over the complex numbers is a function

$$p(z) = a_0 + a_1 z + a_2 z^2 + \dots + a_n z^n$$

in which the a_k are complex numbers and $a_n \neq 0$. The a_i are called the **coefficients** of p . We considered linear and quadratic polynomials in Section A.1. A complex number α is a **root** of p if $p(\alpha) = 0$. A first-degree polynomial has exactly one root, $z = -a_0/a_1$. A quadratic polynomial $p(z) = a_0 + a_1z + a_2z^2$ typically has two roots, which may be obtained by using the quadratic formula

$$z = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_0a_2}}{2a_2},$$

though these roots coincide if $a_1^2 - 4a_0a_2 = 0$. In this case the root $z = -a_1/(2a_2)$ is called a root of **multiplicity 2**, or a **double root**.

The fundamental theorem of algebra states that an n th-degree polynomial always has n roots, when counted properly.

Theorem A.4.1 — Fundamental Theorem of Algebra. Let $p(z) = a_0 + a_1z + \dots + a_nz^n$ be an n th-degree polynomial, $n \geq 1$, with coefficients $a_k \in \mathbb{C}$ and $a_n \neq 0$. Then p factors as

$$p(z) = a_n(z - \alpha_1)(z - \alpha_2) \cdots (z - \alpha_n). \quad (\text{A.10})$$

for complex numbers $\alpha_1, \alpha_2, \dots, \alpha_n$, which are not necessarily distinct. The α_k are the roots of p .

■ **Example A.1** As an example, let $p(z) = 4z^3 - 8(1+i)z^2 + (-8+12i)z + (12-4i)$. The equation $p(z) = 0$ has three solutions, $z = 1$, $z = -1+i$, and $z = 2+i$. These are the roots of $p(z)$, and hence p factors as $p(z) = 4(z-1)(z-(-1+i))(z-(2+i))$. ■

Reading Exercise A.4.1 Compute the four solutions to $z^4 + 5z^2 + 6 = 0$. Hint: let $u = z^2$, so the equation is $u^2 + 5u^2 + 6 = 0$, a quadratic equation in u . Find both solutions u_1 and u_2 and then solve each of $z^2 = u_1$ and $z^2 = u_2$.

Multiplicity of Roots

As stated in Theorem A.4.1, the roots α_k are not necessarily distinct; any α_k may appear multiple times on the right side of (A.10). Let $\beta_1, \beta_2, \dots, \beta_r$ be the distinct roots for p , and m_k the number of times β_k appears as a root of p in (A.10). Then (A.10) can be written as

$$p(z) = (z - \beta_1)^{m_1}(z - \beta_2)^{m_2} \cdots (z - \beta_r)^{m_r}.$$

The integer m_k is called the **multiplicity of the root** β_k , and then $m_1 + m_2 + \dots + m_r = n$.

■ **Example A.2** To illustrate, suppose

$$p(z) = z^6 - 7z^5 + 17z^4 - 13z^3 - 10z^2 + 20z - 8.$$

The polynomial $p(z)$ factors as

$$p(z) = (z-1)^2(z+1)(z-2)^3.$$

Thus p has three distinct roots: 1 is a root of multiplicity 2, -1 is a root of multiplicity 1, and 2 is a root of multiplicity 3. Note that the sum of the multiplicities adds up to the degree of p , i.e., $2+1+3=6$. ■

Conjugation 2

Let us briefly revisit the notion of conjugation, to highlight some properties relevant to polynomials. Conjugation has some surprising algebraic properties. If $z = a + bi$ and $w = c + di$ then $\bar{z} = a - bi$ and $\bar{w} = c - di$. One can then verify that

1. $\overline{z+w} = \bar{z} + \bar{w}$.
2. $\overline{z-w} = \bar{z} - \bar{w}$.
3. $\overline{zw} = (\bar{z})(\bar{w})$.
4. $\overline{z/w} = \bar{z}/\bar{w}$.

In short, conjugation distributes over all of the standard arithmetic operations. Repeated application of $\overline{zw} = (\bar{z})(\bar{w})$ with $w = z$ shows that $\overline{z^n} = (\bar{z})^n$. Conjugation even distributes over exponentiation in that

$$\overline{e^z} = e^{\bar{z}}.$$

To show this let $z = a + bi$. Then from (A.9) $e^z = e^a \cos(b) + ie^a \sin(b)$, so

$$\overline{e^z} = e^a \cos(b) - ie^a \sin(b).$$

Also, $e^{\bar{z}} = e^{a-bi}$, which leads to

$$e^{\bar{z}} = e^a \cos(b) + ie^a \sin(-b) = e^a \cos(b) - ie^a \sin(b)$$

since $\sin(-b) = -\sin(b)$. The last two displayed equations show that $\overline{e^z} = e^{\bar{z}}$.

Reading Exercise A.4.2 If $z = 1 - i$ and $w = 3 + i$, compute \bar{z} , \bar{w} , then verify directly that $\overline{zw} = (\bar{z})(\bar{w})$, $\overline{z/w} = \bar{z}/\bar{w}$, and $\overline{e^z} = e^{\bar{z}}$.

Roots of Polynomials with Real Coefficients

Suppose $p(z) = a_0 + a_1z + \cdots + a_nz^n$ is an n th-degree polynomial in which the coefficients a_i are all real numbers. If α is a root of $p(z)$ then either α is real or part of a complex conjugate pair of roots, that is, $\bar{\alpha}$ must also be a root of $p(z)$. To see why this is true, start with $p(\alpha) = 0$ or

$$0 = a_0 + a_1\alpha + \cdots + a_n\alpha^n.$$

Conjugate both sides and then use the various properties of conjugation (the conjugate of the sum is the sum of the conjugates, the conjugate of the product is the product of the conjugates) to find that

$$\begin{aligned} 0 &= \overline{0} \\ &= \overline{a_0 + a_1\alpha + \cdots + a_n\alpha^n} \\ &= \overline{a_0} + \overline{a_1}(\bar{\alpha}) + \cdots + \overline{a_n}(\bar{\alpha})^n. \end{aligned}$$

But each a_k is real, so $\overline{a_k} = a_k$ and we have

$$0 = a_0 + a_1\bar{\alpha} + \cdots + a_n(\bar{\alpha})^n.$$

This last equation is precisely the statement that $p(\bar{\alpha}) = 0$ and so $\bar{\alpha}$ is a root of $p(z)$.

■ **Example A.3** Let $p(z) = z^5 - 8z^4 + 35z^3 - 80z^2 + 94z - 52$; note that p has real coefficients. The polynomial p has roots $z = 2$, $z = 1 \pm i$, and $z = 2 \pm 3i$. The root $z = 2$ is real, and the last four roots come in conjugate pairs. ■

Rational Functions

A rational function $r(z)$ is a function of the form

$$r(z) = \frac{p(z)}{q(z)}$$

where p and q are polynomials. We generally assume that p and q have no roots in common, for if each has $z = \alpha$ as a root then both the numerator and denominator of $r(z)$ contain a factor of $z - \alpha$ that can be cancelled. We will assume this has been done. The roots of the numerator $p(z)$ are called the **zeros** of the rational function $r(z)$. The roots of the denominator $q(z)$ are called the **poles** of $r(z)$. A rational function is undefined at its poles.

Our primary interest is the case in which the degree of p is strictly less than the degree of q .

A.5 Partial Fraction Decompositions over the Complex Numbers

Factoring Polynomials

The computations of partial fraction decompositions from calculus are conceptually a little simpler if we work with complex numbers, although this won't give any gains in computational ease or efficiency. The conceptual simplicity comes from the fact that any polynomial factors completely into linear pieces over the complex numbers, as in (A.10), and this makes partial fraction decompositions simpler in form. This isn't necessarily the case if we work in the real numbers.

■ **Example A.4** Consider factoring $p(z) = z^3 - z^2 + z - 1$ if we can only work with real numbers. In this case p factors as $p(z) = (z - 1)(z^2 + 1)$. But the quadratic piece $z^2 + 1$ does not factor, since it has no real roots. However if we work in the complex plane then p factors into linear pieces as

$$p(z) = (z - 1)(z - i)(z + i). \quad (\text{A.11})$$

■

Partial Fraction Decompositions

Suppose we are given a rational function $r(z) = p(z)/q(z)$ on which to perform a partial fraction decomposition, and assume that the degree of the polynomial $p(z)$ is strictly less than the degree of the polynomial $q(z)$. Also assume that $p(z)$ and $q(z)$ have no common roots. When q has one or more complex roots, working over the complex numbers can make partial fraction decompositions a little easier. We begin with an example.

■ **Example A.5** Let's find a partial fraction decomposition of the rational function $r(z) = \frac{2z^2 + 4z - 2}{z^3 - z^2 + z - 1}$. Here $r(z) = p(z)/q(z)$ with $p(z) = 2z^2 + 4z - 2$ and $q(z) = z^3 - z^2 + z - 1$. If we are confined to working with real numbers then we can factor the denominator as $q(z) = (z - 1)(z^2 + 1)$ as in Example A.4. The usual rules for partial fractions dictate that we should try a decomposition of the form

$$\frac{2z^2 + 4z - 2}{z^3 - z^2 + z - 1} = \frac{A}{z - 1} + \frac{Bz + C}{z^2 + 1}$$

for some constants A, B , and C . The next step is to find a common denominator, match powers of z in the numerator, and solve for A, B, C . Doing so here yields $A = 2$, $B = 0$, and $C = 4$ (computation omitted). Then

$$\frac{2z^2 + 4z - 2}{z^3 - z^2 + z - 1} = \frac{2}{z - 1} + \frac{4}{z^2 + 1}. \quad (\text{A.12})$$

After obtaining the decomposition (A.12) we are in a position to, for example, integrate $f(z)$, by integrating each piece in the decomposition, or perform an inverse Laplace transform.

But consider how this plays out when we can work with complex numbers. In this case the denominator $q(z)$ factors completely as $q(z) = (z - 1)(z + i)(z - i)$ with three distinct roots, and so we try an expansion of the form

$$r(z) = \frac{2z^2 + 4z - 2}{z^3 - z^2 + z - 1} = \frac{A_2}{z - 1} + \frac{B_2}{z - i} + \frac{C_2}{z + i}.$$

Obtaining a common denominator on the right (a bit of messy algebra required) yields

$$\begin{aligned} \frac{2z^2 + 4z - 2}{z^3 - z^2 + z - 1} &= \frac{A_2}{z - 1} + \frac{B_2}{z - i} + \frac{C_2}{z + i} \\ &= \frac{(A_2 + B_2 + C_2)z^2 + (i - 1)(B_2 + iC_2)z + (A_2 + i(C_2 - B_2))}{z^3 - z^2 + z - 1}. \end{aligned}$$

To make this match $r(z)$ we need to match the numerator $p(z) = 2z^2 + 4z - 2$ of $r(z)$, which leads to

$$A_2 + B_2 + C_2 = 2, \quad (i-1)(B_2 + iC_2) = 4, \quad A_2 + i(C_2 - B_2) = -2.$$

This is three linear equations in three complex unknowns and the solution is $A_2 = 2$, $B_2 = -2i$, and $C_2 = 2i$. Then

$$r(z) = \frac{2z^2 + 4z - 2}{z^3 - z^2 + z - 1} = \frac{2}{z-1} - \frac{2i}{z-i} + \frac{2i}{z+i}. \quad (\text{A.13})$$

Compare this to (A.12). ■

As you were warned, performing the computations over the complex numbers doesn't necessarily make the algebra easier, it merely assures that the partial fraction expansion has a simpler and more explicit dependence on the poles of $r(z)$. In (A.13) all denominators on the right are linear in z .

The General Recipe

More generally, suppose we have a rational function $r(z) = p(z)/q(z)$ in which the degree of p is strictly less than the degree of q . Suppose q has n distinct complex roots β_1, \dots, β_n where β_k has multiplicity m_k , and none of these is a root for p . Then $r(z)$ has a partial fraction expansion of the form

$$r(z) = \sum_{k=1}^{m_1} \frac{A_k}{(z-\beta_1)^k} + \sum_{k=1}^{m_2} \frac{B_k}{(z-\beta_2)^k} + \dots + \sum_{k=1}^{m_n} \frac{R_k}{(z-\beta_r)^k} \quad (\text{A.14})$$

for some (complex) coefficients $A_1, \dots, A_{m_1}, B_1, \dots, B_{m_2}, \dots, R_1, \dots, R_{m_n}$. Informally, if $q(z)$ contains a root α of multiplicity n then the partial fraction decomposition should contain a linear combination involving the terms

$$\frac{1}{z-\alpha}, \quad \frac{1}{(z-\alpha)^2}, \quad \dots, \quad \frac{1}{(z-\alpha)^n}.$$

■ **Example A.6** Let us compute the partial fraction decomposition for $r(z) = \frac{4z}{z^4 + 2z^2 + 1}$; here $r(z) = p(z)/q(z)$ with $p(z) = 4z$ and $q(z) = z^4 + 2z^2 + 1$. The denominator factors as

$$q(z) = (z^2 + 1)^2 = (z-i)^2(z+i)^2.$$

Thus i is root of $q(z)$, or a pole of $r(z)$, of multiplicity 2, and $-i$ as well. Neither is a root for $p(z)$ (the only root of p is $z=0$, which is a zero for $r(z)$). Based on (A.14) we try an expansion of the form

$$\frac{4z}{z^4 + 2z^2 + 1} = \frac{A_1}{z-i} + \frac{A_2}{(z-i)^2} + \frac{B_1}{z+i} + \frac{B_2}{(z+i)^2}.$$

Obtaining a common denominator on the right and matching powers of z^0, z^1, z^2, z^3 leads to four equations in four unknowns, A_1, A_2, B_1, B_2 (the algebra is omitted). The solution is $A_1 = 0, A_2 = -i, B_1 = 0, B_2 = i$, so

$$\frac{4z}{z^4 + 2z^2 + 1} = -\frac{i}{(z-i)^2} + \frac{i}{(z+i)^2}. \quad \blacksquare$$

Although the computations involved in a partial fraction expansion are still cumbersome, the concept embodied by (A.14) is important: the poles of $r(z)$ dictate the form of the partial fraction expansion, and this governs such things as the form of an antiderivative for $r(z)$ or the nature of the inverse Laplace transform of $r(z)$. This in turn may tell us a great deal about the physical application that gave rise to $r(z)$. In many applications the number and locations of the poles of r in (A.14) are of primary interest.

A.6 Additional Exercises

Exercise A.6.1 For each pair of complex numbers z and w , compute

- $\operatorname{Re}(z)$, $\operatorname{Im}(z)$, $\operatorname{Re}(w)$, and $\operatorname{Im}(w)$.
- $z + w$, $z - w$, zw , and z/w
- $|z|$, $|w|$, and $|zw|$. Verify that $|zw| = |z||w|$ (a general truth).
- \bar{z} , \bar{w} , and \bar{zw} . Verify that $(\bar{z})(\bar{w}) = \bar{zw}$.
- e^z , e^w , $e^z e^w$, and e^{z+w} using (A.9). Verify that $e^z e^w = e^{z+w}$. You may find the trigonometric identities $\sin(x+y) = \sin(x)\cos(y) + \cos(x)\sin(y)$ and $\cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y)$ helpful.

- (a) $z = 3 + 4i$, $w = 1 - i$
- (b) $z = 3$, $w = i$
- (c) $z = \pi i$, $w = 1 + \pi i/2$

Exercise A.6.2 Find all complex numbers $z = x + yi$ such that $z^2 = i$. Hint: expand z^2 in terms of x and y , match real and imaginary parts with $0 + i$, and solve for x and y (necessarily real numbers) using the resulting two equations. Make sure to verify that your solutions for z satisfy $z^2 = i$. According to Theorem A.4.1 there should be two solutions, since $z^2 = i$ is a quadratic equation.

Exercise A.6.3 For each polynomial $p(z)$ below, list the distinct roots of p and the multiplicity of each. Then comment on whether p has real-valued coefficients, without multiplying out the polynomial.

- (a) $p(z) = (z - 2)^3(z - i)(z + 3)^2(z + i)$
- (b) $p(z) = (z - (-1 - i))^2 z^7(z - i)^4$
- (c) $p(z) = (z^2 + 1)^{14}$

Exercise A.6.4 Find the roots of the polynomial $p(z) = z^5 - 2z^4 + 3z^3 - 2z^2 + 2z$. Verify that any complex roots are part of a conjugate pair of roots. Hint: $p(0) = 0$ and $p(i) = 0$.

Exercise A.6.5 For each rational function $r(z) = p(z)/q(z)$ below (some numerators and denominators are conveniently factored) find

- The poles and zeros of $r(z)$.
- A partial fraction decomposition based on the complete factorization of $q(z)$ over the complex numbers.

- (a) $r(z) = \frac{z^2 - 3z}{(z - 1)(z^2 + 4)}$
- (b) $r(z) = \frac{z^2 + 2z + 1}{(z - 1)(z - (1 - i))(z - (1 + i))}$
- (c) $r(z) = \frac{6z}{(z^2 + 1)(z^2 + 4)}$

B. Matrix Algebra

This appendix provides some background on the essential matrix algebra necessary to analyze linear systems of constant coefficient differential equations. For the most part we stick to two or three dimensions, especially for examples, but we point out the essential truths and techniques that work in any dimension.

B.1 Linear System of Equations

Linear systems of equations are among the most fundamental objects studied in mathematics and are central to applications. A **linear system of equations** is of the form

$$\begin{aligned} A_{1,1}x_1 + A_{1,2}x_2 + \cdots + A_{1,n}x_n &= b_1 \\ A_{2,1}x_1 + A_{2,2}x_2 + \cdots + A_{2,n}x_n &= b_2 \\ &\vdots \\ A_{m,1}x_1 + A_{m,2}x_2 + \cdots + A_{m,n}x_n &= b_m. \end{aligned} \tag{B.1}$$

Here the **coefficients** $A_{i,j}$ are real or complex numbers, as are the b_i . These coefficients are indexed so that $A_{i,1}, \dots, A_{i,n}$ correspond to the i th equation and $A_{i,j}$ is the coefficient of x_j in that equation. The system (B.1) contains m equations and the quantities x_1, x_2, \dots, x_n are unknowns to be found; they may be real or complex.

A linear system of equations may contain any number of equations and unknowns, but our interest is the special and common case in which $m = n$, so the number of equations equals the number of unknowns. In what follows we make this assumption unless otherwise noted.

■ **Example B.1** Consider the linear system

$$\begin{aligned} 2x_1 + x_2 &= 7 \\ 2x_1 - 4x_2 &= 2. \end{aligned}$$

This fits the mold of (B.1) with $m = n = 2$, $A_{1,1} = 2$, $A_{1,2} = 1$, $A_{2,1} = 2$, $A_{2,2} = -4$, $b_1 = 7$, and $b_2 = 2$. Straightforward algebra (subtract the second equation from the first) shows that the unique solution is $x_1 = 3$, $x_2 = 1$. ■

Matrix Notation

Matrix notation provides a conceptually simple and computationally efficient approach to analyzing linear systems of equations. A **matrix** is a rectangular array of numbers,

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,n} \end{bmatrix}.$$

The matrix \mathbf{A} has m **rows** and n **columns**, and is termed an $m \times n$ **matrix**. The row i column j entry of \mathbf{A} is $A_{i,j}$. Notice the similarity in the structure and indexing of \mathbf{A} and the linear system (B.1). This begins to indicate how matrices will be used to express linear systems.

The solution to a linear system like (B.1) consists of n numbers x_1, x_2, \dots, x_n . These numbers may be amalgamated into an n -dimensional vector $\mathbf{x} = \langle x_1, x_2, \dots, x_n \rangle \in \mathbb{R}^n$, where \mathbb{R}^n denotes n -dimensional Euclidean space. However, when matrices and vectors interact it will be convenient to write vectors in a column format, as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Thus an n -dimensional vector may be identified with an $n \times 1$ matrix.

Matrix-Vector Multiplication

It is possible to multiply a vector by a matrix, and this notion of multiplication is designed so that a linear system like (B.1) can be expressed concisely. Let \mathbf{A} be an $m \times n$ matrix and let \mathbf{x} be an n -dimensional vector. The product \mathbf{Ax} is defined to be the m -dimensional vector

$$\mathbf{Ax} = \begin{bmatrix} A_{1,1}x_1 + A_{1,2}x_2 + \cdots + A_{1,n}x_n \\ A_{2,1}x_1 + A_{2,2}x_2 + \cdots + A_{2,n}x_n \\ \vdots \\ A_{m,1}x_1 + A_{m,2}x_2 + \cdots + A_{m,n}x_n \end{bmatrix}. \quad (\text{B.2})$$

More precisely, if $\mathbf{y} = \mathbf{Ax} \in \mathbb{R}^m$ then \mathbf{y} has i th component

$$y_i = \sum_{j=1}^n A_{i,j}x_j$$

for $1 \leq i \leq m$.

■ **Example B.2** If

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 2 & -4 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} -1 \\ 3 \end{bmatrix}$$

then

$$\mathbf{Ax} = \begin{bmatrix} 2 & 1 \\ 2 & -4 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \end{bmatrix} = \begin{bmatrix} (2)(-1) + (1)(3) \\ (2)(-1) + (-4)(3) \end{bmatrix} = \begin{bmatrix} 1 \\ -14 \end{bmatrix}.$$

■

Reading Exercise B.1.1 Compute the matrix-vector product \mathbf{Ax} if

$$\mathbf{A} = \begin{bmatrix} -3 & 3 \\ 1 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} 2 \\ -4 \end{bmatrix}.$$

With this notation and definition for matrix-vector multiplication, the left sides of the equations in (B.1) can be amalgamated into the components of the product \mathbf{Ax} . The system (B.1) can then be expressed very compactly as

$$\mathbf{Ax} = \mathbf{b} \tag{B.3}$$

where \mathbf{b} is the m -dimensional vector (or $m \times 1$ matrix) with i th component b_i .

■ **Example B.3** The linear system of Example B.1 can be written in the form (B.3) by defining

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 2 & -4 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 7 \\ 2 \end{bmatrix}.$$

■

Reading Exercise B.1.2 Formulate the linear system of equations $3x_1 - x_2 + x_3 = 6$, $x_1 - x_2 = 3$, and $-2x_1 + x_2 + 4x_3 = 0$ (variables x_1, x_2 , and x_3) as $\mathbf{Ax} = \mathbf{b}$, by writing out \mathbf{A} and \mathbf{b} explicitly. Verify that the vector $\mathbf{x} = \langle 1, -2, 1 \rangle$ is a solution by computing \mathbf{Ax} and comparing to \mathbf{b} .

Matrix-vector multiplication is linear, that is,

$$\mathbf{A}(c_1\mathbf{x} + c_2\mathbf{y}) = c_1\mathbf{Ax} + c_2\mathbf{Ay}$$

for any vectors \mathbf{x}, \mathbf{y} and scalars c_1, c_2 . We use the vector $\mathbf{0}$ to indicate the vector with all components equal to zero, with whatever dimension is appropriate to the computation at hand.

When the system (B.3) has the same number of equations as unknowns the resulting matrix \mathbf{A} has the same number of rows and columns and is called a **square** matrix. This is the case on which we will focus in the next section.

Solvability of Linear Systems

Linear systems of equations may have a unique solution, infinitely many solutions, or no solutions. If a system $\mathbf{Ax} = \mathbf{b}$ has at least one solution vector \mathbf{x} we say that the system is **consistent**. If there are no solutions the system is **inconsistent**.

The following theorem is usually proved in a linear algebra class using the method of Gaussian elimination.

Theorem B.1.1 — Solvability of Linear Systems. Let \mathbf{A} be an $m \times n$ matrix. Then one of the following holds.

1. For any choice of $\mathbf{b} \in \mathbb{R}^m$ the system $\mathbf{Ax} = \mathbf{b}$ is consistent and has a unique solution \mathbf{x} in \mathbb{R}^n , or
2. For any choice of $\mathbf{b} \in \mathbb{R}^m$ the system $\mathbf{Ax} = \mathbf{b}$ is either inconsistent or is consistent with infinitely many solutions in \mathbb{R}^n .

Reading Exercise B.1.3

- (a) Formulate the linear system of equations $x_1 - 2x_2 = 3$, $2x_1 - 4x_2 = 6$ as $\mathbf{Ax} = \mathbf{b}$ by writing out \mathbf{A} and \mathbf{b} explicitly. Show that any vector \mathbf{x} of the form $\mathbf{x} = \langle 3 + 2t, t \rangle$ is a solution, for any choice of t , so this system is consistent and has infinitely many solutions.
- (b) Formulate the linear system of equations $x_1 - 2x_2 = 3$, $2x_1 - 4x_2 = 5$ as $\mathbf{Ax} = \mathbf{b}$ by writing out \mathbf{A} and \mathbf{b} explicitly. Show that this system is inconsistent.

B.2 Matrix Algebra

Throughout the remainder of this appendix all matrices will be square. The entries of an $n \times n$ matrix \mathbf{A} of the form $A_{i,i}$ for $i = 1$ to $i = n$ are called the **diagonal** entries of \mathbf{A} .

First let us note that any matrix \mathbf{A} can be multiplied by a scalar c , by multiplying each entry in \mathbf{A} by c . That is, the row i column j entry of $c\mathbf{A}$ is $cA_{i,j}$. Also, $n \times n$ matrices \mathbf{A} and \mathbf{B} can be added in an obvious fashion: the row i column j entry of $\mathbf{A} + \mathbf{B}$ is $A_{i,j} + B_{i,j}$.

Matrices of the proper dimensions can also be multiplied, which we now discuss.

Matrix Multiplication

Let \mathbf{A} and \mathbf{B} be $n \times n$ matrices. The product \mathbf{AB} is defined as an $n \times n$ matrix as follows. First, let us view the matrix \mathbf{B} as consisting of n column vectors $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$, each with n components, concatenated so that

$$\mathbf{B} = [\mathbf{b}_1 \mid \mathbf{b}_2 \mid \cdots \mid \mathbf{b}_n]$$

where the vertical symbols $|$ delineate the columns of \mathbf{B} . The product \mathbf{AB} is defined as

$$\mathbf{AB} = [\mathbf{Ab}_1 \mid \mathbf{Ab}_2 \mid \cdots \mid \mathbf{Ab}_n].$$

That is, the product \mathbf{AB} is computed by multiplying each column \mathbf{b}_j of \mathbf{B} by \mathbf{A} using the definition of matrix-vector multiplication (B.2).

■ **Example B.4** Let

$$\mathbf{A} = \begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 3 & 0 \\ -1 & 7 \end{bmatrix}.$$

Then (using the $|$ symbol to temporarily separate the columns of \mathbf{B} into vectors) we find

$$\begin{aligned} \mathbf{AB} &= \begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ -1 & 7 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 3 & | & 0 \\ -1 & | & 7 \end{bmatrix} \\ &= \begin{bmatrix} (1)(3) + (-1)(-1) & | & (1)(0) + (-1)(7) \\ (2)(3) + (3)(-1) & | & (2)(0) + (3)(7) \end{bmatrix} \\ &= \begin{bmatrix} 4 & | & -7 \\ 3 & | & 21 \end{bmatrix} \\ &= \begin{bmatrix} 4 & -7 \\ 3 & 21 \end{bmatrix}. \end{aligned}$$

■

Reading Exercise B.2.1 Compute \mathbf{BA} with \mathbf{A} and \mathbf{B} as in Example B.4. Hint: it's not the same as \mathbf{AB} .

Properties of Matrix Multiplication

For any $n \times n$ matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and n -dimensional vector \mathbf{x} the following properties hold.

1. **Associativity:** $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$ (both sides are $n \times n$ matrices) and $(\mathbf{AB})\mathbf{x} = \mathbf{A}(\mathbf{Bx})$ (both sides are n -dimensional vectors).
2. **Distributivity and Linearity:** $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$, $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$, and $(\mathbf{A} + \mathbf{B})\mathbf{x} = \mathbf{Ax} + \mathbf{Bx}$.

One property that matrix multiplication does not enjoy is commutativity, as illustrated by Reading Exercise B.2.1. In general

$$\mathbf{AB} \neq \mathbf{BA}.$$

Reading Exercise B.2.2 With \mathbf{A} and \mathbf{B} as in Example B.4 and

$$\mathbf{C} = \begin{bmatrix} -1 & 3 \\ 2 & 0 \end{bmatrix}$$

compute each product $(\mathbf{AB})\mathbf{C}$ and $\mathbf{A}(\mathbf{BC})$, and verify they are the same.

The Identity Matrix

The $n \times n$ **identity matrix** \mathbf{I} is defined as

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

The identity matrix has all entries equal to zero except the diagonal entries (entries of the form $I_{j,j}$), which all equal 1. The specific size of the matrix, $n \times n$, is not explicitly indicated by the notation \mathbf{I} but is usually clear from context. When necessary the matrix can be subscripted with an n , so \mathbf{I}_n indicates the $n \times n$ identity matrix.

The $n \times n$ identity matrix has the property that $\mathbf{IA} = \mathbf{AI}$ for any $n \times n$ matrix \mathbf{A} . The identity matrix is thus analogous to the number 1 with regard to multiplication in the real (or complex) numbers.

Reading Exercise B.2.3 Write out the 2×2 , 3×3 , and 4×4 identity matrices explicitly. Use \mathbf{I}_2 and the matrix \mathbf{C} from Example B.2.2 to compute $\mathbf{I}_2\mathbf{C}$ and \mathbf{CI}_2 .

Matrix Inversion

Recall from basic algebra that for any nonzero real number a there is a unique real number c such that $ac = ca = 1$. We conventionally call c the **reciprocal** or the **multiplicative inverse** of a and write $c = 1/a$ or $c = a^{-1}$. The notion of reciprocal is useful when solving equations like $ax = b$: multiply both sides of $ax = b$ by a^{-1} to obtain

$$a^{-1}(ax) = a^{-1}b. \quad (\text{B.4})$$

But since $a^{-1}(ax) = (a^{-1}a)x = 1x = x$, (B.4) collapses to $x = a^{-1}b$ or $x = b/a$. In this computation we used the associativity of real multiplication in writing $a^{-1}(ax) = (a^{-1}a)x$ and the fact that $1x = x$ for any x . The ability to reciprocate nonzero real numbers shows that we can solve any equation $ax = b$ when $a \neq 0$, and the solution x is unique.

Something similar can be done to solve a linear system $\mathbf{Ax} = \mathbf{b}$. It is not usually the most computationally efficient method, especially for large systems, but it is conceptually important. If you examine the computation in (B.4) you will see that, with \mathbf{A} replacing a , \mathbf{x} replacing x , and \mathbf{b} replacing b , the computation could proceed in exactly the same fashion with \mathbf{I} playing the role of 1, if we can find a matrix \mathbf{C} with the property that

$$\mathbf{CA} = \mathbf{I}.$$

It is also essential to make use of the associativity of matrix multiplication if the comparable computation is to be valid.

Such a matrix \mathbf{C} , if it exists, is called the **inverse matrix** or just **inverse** of \mathbf{A} , and is written \mathbf{A}^{-1} (never as $1/\mathbf{A}$). If \mathbf{A}^{-1} exists then we say that \mathbf{A} is invertible. For a general 2×2 matrix

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad (\text{B.5})$$

the inverse is

$$\mathbf{A}^{-1} = \frac{1}{D} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

where $D = ad - bc$, provided $ad - bc \neq 0$. This can be verified by computing $\mathbf{A}^{-1}\mathbf{A}$, which equals the 2×2 identity matrix. For a general 3×3 matrix

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \quad (\text{B.6})$$

the inverse is

$$\mathbf{A}^{-1} = \frac{1}{D} \begin{bmatrix} ei - fh & ch - bi & bf - ce \\ fg - di & ai - cg & cd - af \\ dh - eg & bg - ah & ae - bd \end{bmatrix}$$

where $D = aei - afh - dbi + gbf + dch - gce$, provided $D \neq 0$. Again, this can be verified by computing $\mathbf{A}^{-1}\mathbf{A}$. There is a general procedure for computing \mathbf{A}^{-1} for a square matrix of any size, provided the inverse exists, but for larger matrices the formulas are even more complicated. Happily, we rarely need to actually compute \mathbf{A}^{-1} ; the inverse is primarily a conceptual tool. The 2×2 case might be worth memorizing.

It can be shown that if \mathbf{A} is invertible then there is only one inverse for \mathbf{A} , that is, the inverse is unique. Matrices that are not invertible are said to be **singular**.

Some Properties of Matrix Inverses

Matrix inverses have some important properties. First, as alluded to above, not all matrices are invertible. In the 2×2 case for a matrix \mathbf{A} of the form (B.5), \mathbf{A}^{-1} exists if and only if the quantity $D = ad - bc$ is nonzero, while in the 3×3 case (B.6) the matrix \mathbf{A}^{-1} exists if and only if $D = aei - afh - dbi + gbf + dch - gce$ is nonzero. We'll say more about this quantity momentarily.

Reading Exercise B.2.4 Compute \mathbf{A}^{-1} where

$$\mathbf{A} = \begin{bmatrix} -1 & 3 \\ 2 & 1 \end{bmatrix}.$$

Verify that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, and also that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

Reading Exercise B.2.5 Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

Show that \mathbf{A}^{-1} has no inverse, by showing that the equation

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

is not satisfied for any choice of a, b, c , and d .

Reading Exercise B.2.4 illustrates an important point: although matrix multiplication is not commutative, it is true that any invertible matrix \mathbf{A} commutes with its own inverse, that is, if $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ then $\mathbf{AA}^{-1} = \mathbf{I}$ and vice versa. It is also the case that if \mathbf{A} and \mathbf{B} are both invertible then the product \mathbf{AB} is also invertible, and

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}. \quad (\text{B.7})$$

Notice that the right hand side of (B.7) is the product of the inverses but in reverse order.

Implications for Linear Systems of Equations

If an $n \times n$ matrix \mathbf{A} is invertible then the linear system of equations $\mathbf{Ax} = \mathbf{b}$ has a unique solution for any choice of \mathbf{b} , and this can be found by following the same pattern as in (B.4): multiply both sides of $\mathbf{Ax} = \mathbf{b}$ by \mathbf{A}^{-1} to obtain $\mathbf{A}^{-1}(\mathbf{Ax}) = \mathbf{A}^{-1}\mathbf{b}$ or

$$\mathbf{A}^{-1}(\mathbf{Ax}) = (\mathbf{A}^{-1}\mathbf{A})\mathbf{x} = \mathbf{Ix} = \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

We made use of the associativity of matrix multiplication in writing $\mathbf{A}^{-1}(\mathbf{Ax}) = (\mathbf{A}^{-1}\mathbf{A})\mathbf{x}$, as well as the properties of the identity matrix. Thus when \mathbf{A} is invertible the unique solution to $\mathbf{Ax} = \mathbf{b}$ is

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \quad (\text{B.8})$$

One drawback to using (B.8) for solving a linear system is that computing \mathbf{A}^{-1} for large matrices takes at least twice as much computation as solving $\mathbf{Ax} = \mathbf{b}$ using more direct methods. Still, the inverse is a useful conceptual and theoretical tool.

One useful conclusion we can draw for invertible matrices is this: if \mathbf{A} is invertible then the only solution to $\mathbf{Ax} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$, where $\mathbf{0}$ denotes the zero vector.. This is easy to see by using (B.8) with $\mathbf{b} = \mathbf{0}$. The converse also turns out to be true, as embodied by the next theorem. Again, one would prove this in a typical linear algebra course.

Theorem B.2.1 — Solvability of Linear Systems. A square matrix is invertible if and only if the unique solution to $\mathbf{Ax} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$.

Determinants

The quantities $D = ad - bc$ for the 2×2 matrix \mathbf{A} in (B.5) and $D = aei - afh - dbi + gbf + dch - gce$ for the 3×3 matrix \mathbf{A} in (B.6) are of special significance. As mentioned, the relevant matrices are invertible if and only if D is not zero. The quantity D is called the **determinant** of the matrix \mathbf{A} and usually written as $\det(\mathbf{A})$ or $|\mathbf{A}|$. More generally, one can define $\det(\mathbf{A})$ for any $n \times n$ matrix. The determinant of \mathbf{A} is a scalar quantity formed from the elements of \mathbf{A} and can be thought of as a function that accepts a matrix as input and returns a scalar.

We'll discuss how to compute $\det(\mathbf{A})$ momentarily, but first it's worth noting that the determinant has the following properties:

1. An $n \times n$ matrix \mathbf{A} is invertible if and only if $\det(\mathbf{A})$ is not zero.
2. For the $n \times n$ identity matrix \mathbf{I} we have $\det(\mathbf{I}) = 1$.
3. $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$ for any $n \times n$ matrices \mathbf{A} and \mathbf{B} .

From this point on we will use $|\mathbf{A}|$ for the determinant of an $n \times n$ matrix \mathbf{A} . The classic method for computing $|\mathbf{A}|$ is called **expansion by minors** (also **cofactor expansion** or **Laplace expansion**). The computation can proceed in many ways, but we will present just one, and illustrate with an application to a 4×4 matrix.

Let \mathbf{A} be an $n \times n$ matrix with row i , column j entry $A_{i,j}$. For any choice of i and j , let $\tilde{\mathbf{A}}_{i,j}$ be the $(n-1) \times (n-1)$ matrix formed by striking out all entries in row i of \mathbf{A} as well as all entries in

row j , then assembling the remaining elements into an $(n - 1) \times (n - 1)$ matrix in the natural way. Thus for example if

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 4 & 3 \\ 0 & 3 & -2 & 7 \\ 1 & 1 & -1 & 2 \\ 3 & -2 & 0 & 5 \end{bmatrix} \quad (\text{B.9})$$

then to compute $\tilde{\mathbf{A}}_{1,3}$ we first strike out all entries in row 1 and column 3 of \mathbf{A} , which might be depicted as

$$\tilde{\mathbf{A}}_{1,3} = \begin{bmatrix} - & - & * & - \\ 0 & 3 & | & 7 \\ 1 & 1 & | & 2 \\ 3 & -2 & | & 5 \end{bmatrix}.$$

We form $\tilde{\mathbf{A}}_{1,3}$ by assembling the remaining entries into a 3×3 matrix

$$\tilde{\mathbf{A}}_{1,3} = \begin{bmatrix} 0 & 3 & 7 \\ 1 & 1 & 2 \\ 3 & -2 & 5 \end{bmatrix}. \quad (\text{B.10})$$

The matrix $\tilde{\mathbf{A}}_{1,3}$ is called the **cofactor matrix** of the row 1, column 3 entry of \mathbf{A} .

The computation of the determinant of a matrix \mathbf{A} can be accomplished by computing the determinants of a number of smaller cofactor matrices. One approach is this: For each element $A_{1,j}$, $1 \leq j \leq n$ in the first row of \mathbf{A} , compute the corresponding cofactor matrix $\tilde{\mathbf{A}}_{1,j}$. The determinant of \mathbf{A} can then be computed as the alternating sum

$$\begin{aligned} |\mathbf{A}| &= A_{1,1}|\tilde{\mathbf{A}}_{1,1}| - A_{1,2}|\tilde{\mathbf{A}}_{1,2}| + A_{1,3}|\tilde{\mathbf{A}}_{1,3}| - \cdots + (-1)^{n-1}A_{1,n}|\tilde{\mathbf{A}}_{1,n}| \\ &= \sum_{j=1}^m (-1)^{j-1}A_{1,j}|\tilde{\mathbf{A}}_{1,j}|. \end{aligned} \quad (\text{B.11})$$

For example, for the 4×4 matrix \mathbf{A} in (B.9) we have

$$\begin{aligned} |\mathbf{A}| &= A_{1,1}|\tilde{\mathbf{A}}_{1,1}| - A_{1,2}|\tilde{\mathbf{A}}_{1,2}| + A_{1,3}|\tilde{\mathbf{A}}_{1,3}| - A_{1,4}|\tilde{\mathbf{A}}_{1,4}| \\ &= (1) \left| \begin{bmatrix} 3 & -2 & 7 \\ 1 & -1 & 2 \\ -2 & 0 & 5 \end{bmatrix} \right| - (2) \left| \begin{bmatrix} 0 & -2 & 7 \\ 1 & -1 & 2 \\ 3 & 0 & 5 \end{bmatrix} \right| + (4) \left| \begin{bmatrix} 0 & 3 & 7 \\ 1 & 1 & 2 \\ 3 & -2 & 5 \end{bmatrix} \right| - (3) \left| \begin{bmatrix} 0 & 3 & -2 \\ 1 & 1 & -1 \\ 3 & -2 & 0 \end{bmatrix} \right| \end{aligned} \quad (\text{B.12})$$

Each of the 3×3 determinants (four total) on the right in (B.12) can be computed using either the formula $D = aei - afh - dbi + gbf + dch - gce$ for the 3×3 matrix \mathbf{A} in (B.6), or the cofactor expansion approach. With the latter approach the determinant of each 3×3 cofactor matrix will require the computation of the determinant of three 2×2 matrices. The determinant of each such 2×2 matrix of the form (B.5) can be computed using $D = ad - bc$, or by performing cofactor expansion again, with the understanding that the determinant of a 1×1 matrix $\mathbf{A} = [a]$ is just $|\mathbf{A}| = a$.

This approach for computing the determinant of an $n \times n$ matrix is clearly laborious and requires on the order of $n! = n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1$ arithmetic operations, which is prohibitive if n is very large. The formula $\det(\mathbf{A}) = ad - bc$ for a 2×2 matrix is a quick and handy way to see if such a matrix is invertible and is worth knowing. For larger matrices more computationally efficient

methods exist and it's worth relying on software for the computation. In any case, the explicit computation of $|\mathbf{A}|$ is not usually an efficient method for checking whether \mathbf{A} is invertible. But we should emphasize that determinants have their uses. In this text, the Routh-Hurwitz theorem (Theorem 7.6.4 in Section 7.6.3) is one specific example.

Reading Exercise B.2.6 Compute the determinant of each 3×3 matrix on the right in (B.12), either by using expansion by minors on each matrix, or using the formula $D = aei - afh - dbi + gbf + dch - gce$ for the 3×3 matrix \mathbf{A} in (B.6). Use this to compute the determinant of the relevant 4×4 matrix \mathbf{A} in (B.9).

Reading Exercise B.2.7 Derive the formula $D = aei - afh - dbi + gbf + dch - gce$ for the 3×3 matrix \mathbf{A} in (B.6) by performing expansion by minors.

Reading Exercise B.2.8 Show that the determinant of an $n \times n$ diagonal matrix \mathbf{A} with diagonal elements $A_{i,i} = d_i$ is given by $|\mathbf{A}| = d_1 d_2 \cdots d_n$.

B.3 Eigenvalues and Eigenvectors

Definition

In Chapter 6 the notions of eigenvalues and eigenvectors for matrices are used to solve linear constant coefficient system of ODEs. In this section we provide the reader with a bit of background or review on the essentials of this topic.

Let \mathbf{A} be an $n \times n$ matrix with real or complex entries.

Definition B.3.1 — Eigenvalues and Eigenvectors. A nonzero n -dimensional vector \mathbf{v} is an **eigenvector** for \mathbf{A} and the scalar λ is the associated **eigenvalue** if

$$\mathbf{Av} = \lambda \mathbf{v}.$$

It is important to note that \mathbf{v} cannot be the zero vector, since $\mathbf{A}\mathbf{0} = \lambda\mathbf{0}$ is trivially true for any matrix \mathbf{A} and scalar λ , which is of little interest.

■ **Example B.5** Let

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}.$$

Then one can check that

$$\mathbf{A} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 4 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{A} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = -2 \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Thus the vector $\mathbf{v}_1 = \langle 1, 1 \rangle$ is an eigenvector for \mathbf{A} with eigenvalue $\lambda_1 = 4$, and $\mathbf{v}_2 = \langle 1, -1 \rangle$ is also an eigenvector for \mathbf{A} with eigenvalue $\lambda_2 = -2$. ■

Example B.5 shows that a matrix may have more than one eigenvector and eigenvalue. However, if \mathbf{v} is an eigenvector with eigenvalue λ then $c\mathbf{v}$ is also an eigenvector with the same eigenvalue λ for any nonzero scalar c . This follows from

$$\mathbf{A}(c\mathbf{v}) = c\mathbf{Av} = c(\lambda\mathbf{v}) = \lambda(c\mathbf{v}).$$

That is, $\mathbf{A}(c\mathbf{v}) = \lambda(c\mathbf{v})$, so $c\mathbf{v}$ (which is not $\mathbf{0}$ if $c \neq 0$) is an eigenvector with eigenvalue λ . Thus in Example B.5 the vector $2\mathbf{v}_1 = \langle 2, 2 \rangle$ is also an eigenvector for \mathbf{A} with eigenvalue $\lambda_1 = 4$, and so is $\pi\mathbf{v}_1 = \langle \pi, \pi \rangle$, and so on. If \mathbf{v} is an eigenvector for a matrix \mathbf{A} we do not count scalar multiples of \mathbf{v} as separate eigenvectors.

Reading Exercise B.3.1 Let

$$\mathbf{A} = \begin{bmatrix} -3 & 2 \\ -12 & 7 \end{bmatrix}.$$

Verify that $\mathbf{v}_1 = \langle 1, 2 \rangle$ is an eigenvector for \mathbf{A} with eigenvalue $\lambda_1 = 1$. Verify that $\mathbf{v}_2 = \langle 1, 3 \rangle$ is an eigenvector for \mathbf{A} ; what is the eigenvalue?

Computing Eigenvectors and Eigenvalues

Computing the eigenvectors and eigenvalues for a 2×2 matrix is straightforward, so we examine this case in detail first. This is the primary case of interest anyway and the method illuminates how one might go about computing these quantities for larger matrices. Then we'll point out why you should not generally use this method for 3×3 matrices or larger.

Computing Eigenvalues

The first step is to find the eigenvalues. Consider an $n \times n$ matrix \mathbf{A} (we are not specifying $n = 2$ yet). We seek a vector nonzero \mathbf{v} and a scalar λ such that $\mathbf{Av} = \lambda\mathbf{v}$. Write this last equation as

$$\mathbf{Av} - \lambda\mathbf{v} = \mathbf{0}. \quad (\text{B.13})$$

We'd like to factor out \mathbf{v} , that is, write $(\mathbf{A} - \lambda)\mathbf{v} = \mathbf{0}$, but the difference $\mathbf{A} - \lambda$ makes no sense since \mathbf{A} is a matrix (2×2 or larger) and λ is a scalar. Instead, write (B.13) as

$$\mathbf{Av} - \lambda\mathbf{I}\mathbf{v} = \mathbf{0}.$$

where \mathbf{I} is the identity matrix with the same dimensions as \mathbf{A} . The factorization

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0} \quad (\text{B.14})$$

is now valid, since \mathbf{A} and $\lambda\mathbf{I}$ have the same dimensions. Recall the \mathbf{v} is not the zero vector, and yet from (B.14) when \mathbf{v} is multiplied by the matrix $\mathbf{A} - \lambda\mathbf{I}$ the result must be $\mathbf{0}$. According to Theorem B.2.1 this means that $\mathbf{A} - \lambda\mathbf{I}$ is not invertible. Conversely, if $\mathbf{A} - \lambda\mathbf{I}$ is not invertible then by Theorem B.2.1 there is a nonzero vector \mathbf{v} that satisfies (B.14), and so satisfies $\mathbf{Av} = \lambda\mathbf{v}$. This yields an important conclusion:

The eigenvalues of \mathbf{A} are precisely those scalars λ such that the matrix $\mathbf{A} - \lambda\mathbf{I}$ is not invertible.

This observation can be used to easily compute the eigenvalues for a 2×2 matrix. We begin with an example.

■ **Example B.6** Let us compute the eigenvalues for the matrix \mathbf{A} from Example B.5 to illustrate. We have

$$\begin{aligned} \mathbf{A} - \lambda\mathbf{I} &= \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 - \lambda & 3 \\ 3 & 1 - \lambda \end{bmatrix}. \end{aligned}$$

The determinant provides a convenient way to find those λ such that $\mathbf{A} - \lambda\mathbf{I}$ is not invertible. Computing the determinant of $\mathbf{A} - \lambda\mathbf{I}$ shows that

$$\det(\mathbf{A} - \lambda\mathbf{I}) = (1 - \lambda)^2 - 9 = \lambda^2 - 2\lambda - 8.$$

The matrix $\mathbf{A} - \lambda \mathbf{I}$ is not invertible when

$$\lambda^2 - 2\lambda - 8 = 0, \quad (\text{B.15})$$

which is a quadratic equation in λ . The quadratic polynomial $\lambda^2 - 2\lambda - 8$ is called the **characteristic polynomial** for \mathbf{A} , and (B.15) is the **characteristic equation** for \mathbf{A} . The solutions to (B.15) are the eigenvalues of \mathbf{A} . Factoring or using the quadratic formula yields $\lambda = -2$ and $\lambda = 4$. ■

Computing Eigenvectors

After computing the eigenvalues, the next step is to find the eigenvector(s) for each eigenvalue. The matrix $\mathbf{A} - \lambda \mathbf{I}$ is not invertible when λ is an eigenvalue, and so by Theorem B.2.1 there must be a nonzero solution to \mathbf{v} to $(\mathbf{A} - \lambda \mathbf{I})\mathbf{v} = \mathbf{0}$. This is equivalent to $\mathbf{Av} = \lambda \mathbf{v}$. We consider each eigenvalue in turn, and solve $(\mathbf{A} - \lambda \mathbf{I})\mathbf{v} = \mathbf{0}$ to find a corresponding eigenvector \mathbf{v} .

Again, we proceed with an example.

■ **Example B.7** Let \mathbf{A} be the matrix from Examples B.5. The eigenvalues of \mathbf{A} were computed in Example B.6 and are $\lambda_1 = 4$ and $\lambda_2 = -2$. Let us find an eigenvector for λ_2 .

Begin by forming the matrix $\mathbf{A} - \lambda_2 \mathbf{I}$ or

$$\mathbf{A} - (-2)\mathbf{I} = \begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix}.$$

If we use x and y to denote the components of an eigenvector $\mathbf{v} = \langle x, y \rangle$ then $(\mathbf{A} - \lambda_2 \mathbf{I})\mathbf{v} = \mathbf{0}$ is equivalent to the equations $3x + 3y = 0$ and $3x + 3y = 0$. These equations are duplicates and it's easy to see both are equivalent to $y = -x$. Any nonzero vector $\mathbf{v} = \langle x, -x \rangle$ will be an eigenvector for \mathbf{A} with eigenvalue -2 . Choosing $x = 1$ and $y = -1$ yields the eigenvector \mathbf{v}_2 from Example B.5. ■

Reading Exercise B.3.2 Emulate the computation of Example B.7 to find an eigenvector \mathbf{v}_1 for the matrix \mathbf{A} in that example for the eigenvalue $\lambda_1 = 4$. Remember that \mathbf{v}_1 will be determined only up to a nonzero scalar multiple, so there are many answers, all multiples of each other.

Remark B.3.1 Here is a tip: if you've computed an eigenvalue λ , formed the matrix $\mathbf{A} - \lambda \mathbf{I}$, and found that the only solution to $(\mathbf{A} - \lambda \mathbf{I})\mathbf{v} = \mathbf{0}$ is to take $\mathbf{v} = \mathbf{0}$, then your eigenvalue is not correct. If λ is an eigenvalue then $(\mathbf{A} - \lambda \mathbf{I})\mathbf{v} = \mathbf{0}$ must possess a nonzero solution, since $(\mathbf{A} - \lambda \mathbf{I})$ is not invertible.

Let's consider one more 2×2 example that illustrates that complex eigenvalues and eigenvectors are possible, even if the matrix has real entries.

■ **Example B.8** Let us compute the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{bmatrix} -3 & 2 \\ -10 & 5 \end{bmatrix}.$$

First, the characteristic polynomial is

$$\det(\mathbf{A} - \lambda \mathbf{I}) = (-3 - \lambda)(5 - \lambda) - (2)(-10) = \lambda^2 - 2\lambda + 5.$$

The roots of the characteristic polynomial are the eigenvalues. From the quadratic formula these are $(2 \pm \sqrt{4 - 20})/2 = 1 \pm \sqrt{-4}$ or

$$\lambda_1 = 1 + 2i \quad \text{and} \quad \lambda_2 = 1 - 2i.$$

To find an eigenvector for λ_1 form

$$\mathbf{A} - (1 + 2i)\mathbf{I} = \begin{bmatrix} -4 - 2i & 2 \\ -10 & 4 - 2i \end{bmatrix}.$$

With $\mathbf{v} = \langle x, y \rangle$ the equation $(\mathbf{A} - \lambda_2 \mathbf{I})\mathbf{v} = \mathbf{0}$ is equivalent to the equations $(-4 - 2i)x + 2y = 0$, $-10x + (4 - 2i)y = 0$. In view of Remark B.3.1 these equations must possess a nonzero solution. These equations are in fact scalar multiples of each other: multiplying the first equation $(-4 - 2i)x + 2y = 0$ by $1 - 2i$ yields the second equation $-10x + (4 - 2i)y = 0$. The two equations are thus duplicates and so one can be discarded. If we work with $(-4 - 2i)x + 2y = 0$ we can choose either x or y freely, say $x = 1$, and then solve for $y = 2 + i$. Then $\mathbf{v}_1 = \langle 1, 2 + i \rangle$ is an eigenvector for \mathbf{A} with eigenvalue $1 + 2i$. Any nonzero multiple of \mathbf{v}_1 is also an eigenvector with this eigenvalue. Notice that \mathbf{v}_1 is complex-valued, as is λ_1 , even though \mathbf{A} has only real entries. ■

When the matrix \mathbf{A} has real entries the characteristic polynomial has real coefficients, but may have complex roots. In light of the discussion in Section A.4 this means that since the eigenvalues of \mathbf{A} are the roots of this polynomial, each eigenvalue is either real or part of a complex conjugate pair of eigenvalues. This is illustrated by Example B.8, in which the eigenvalues are $1 \pm 2i$. Further, the eigenvectors corresponding to conjugate eigenvalues are themselves conjugate, component by component.

Reading Exercise B.3.3 Compute an eigenvector for the eigenvalue $\lambda = 1 - 2i$ for the matrix \mathbf{A} in Example B.8.

The following Reading Exercises illustrate other possibilities for the eigenvalues and eigenvectors of a 2×2 matrix, and some properties of eigenvalues.

Reading Exercise B.3.4 Consider the 2×2 matrix $\mathbf{A} = 2\mathbf{I}$. Show that the characteristic polynomial for \mathbf{A} is $(\lambda - 2)^2$, and that every nonzero vector $\mathbf{v} \in \mathbb{R}^2$ is an eigenvector for \mathbf{A} .

Reading Exercise B.3.5 Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}.$$

Show that the characteristic polynomial for \mathbf{A} is $(\lambda - 2)^2$, and that both eigenvalues are equal to 2. Then show that all eigenvectors are scalar multiples of $\mathbf{v} = \langle 1, 0 \rangle$. So here there is only one eigenvector (up to scalar multiples) for both eigenvalues. In this case we say that the eigenvalue 2 is **defective**, something that may happen when the eigenvalue is a double root of the characteristic polynomial. Contrast the situation here to that of Reading Exercise B.3.4.

Reading Exercise B.3.6 Show that if $\lambda = 0$ is an eigenvalue for \mathbf{A} (here \mathbf{A} can be a square matrix of any size) then \mathbf{A} cannot be invertible. Hint: the eigenvector \mathbf{v} would satisfy $\mathbf{Av} = 0\mathbf{v} = \mathbf{0}$. Now recall Theorem B.2.1. Show that the converse is also true: if \mathbf{A} is not invertible then 0 is an eigenvalue for \mathbf{A} . Hint: again, consider Theorem B.2.1.

Reading Exercise B.3.7 Suppose \mathbf{A} is of one of the special forms

$$\mathbf{A} = \begin{bmatrix} a & b \\ 0 & d \end{bmatrix} \quad \text{or} \quad \mathbf{A} = \begin{bmatrix} a & 0 \\ c & d \end{bmatrix}.$$

These are 2×2 examples of an **upper triangular matrix** (all entries below the diagonal equal to zero) and a **lower triangular matrix** (all entries above the diagonal equal to zero), respectively. Show that in each case the characteristic polynomial of \mathbf{A} is $\lambda^2 - (a+d)\lambda + ad$ and that the eigenvalues are $\lambda = a$ and $\lambda = d$.

B.4 The Eigenvalues for a General Two by Two Matrix

There are some insights we can gain for the eigenvalues of a 2×2 matrix that don't require us to compute the eigenvalues explicitly. These insights are convenient for analyzing systems of ODEs

in the plane, especially systems that contain unspecified parameters. For a 2×2 matrix \mathbf{A} of the general form (B.5) this technique requires that we compute the determinant $\det(\mathbf{A}) = ad - bc$ and also the **trace** of \mathbf{A} , denoted and defined as $\text{tr}(\mathbf{A}) = a + d$. The trace is the sum of the diagonal elements of \mathbf{A} .

We can compute the eigenvalues of \mathbf{A} if we know the trace and the determinant of \mathbf{A} . To see this form the matrix

$$\mathbf{A} - \lambda \mathbf{I} = \begin{bmatrix} a - \lambda & b \\ c & d - \lambda \end{bmatrix}$$

and compute its characteristic polynomial $\det(\mathbf{A} - \lambda \mathbf{I}) = \lambda^2 - (a+d)\lambda + (ad - bc)$. For simplicity write $T = \text{tr}(\mathbf{A}) = a + d$ and $D = \det(\mathbf{A}) = ad - bc$, so the characteristic polynomial is $\lambda^2 - T\lambda + D$. The eigenvalues of \mathbf{A} are the roots of this polynomial, the solutions to

$$\lambda^2 - T\lambda + D = 0. \quad (\text{B.16})$$

These eigenvalues are easily obtained from the quadratic formula applied to (B.16) and are

$$\begin{aligned} \lambda_1 &= \frac{T + \sqrt{T^2 - 4D}}{2} \\ \lambda_2 &= \frac{T - \sqrt{T^2 - 4D}}{2}. \end{aligned} \quad (\text{B.17})$$

Thus knowledge of T and D completely determines the eigenvalues.

Eigenvalues Properties from the Trace and the Determinant

We can quickly and easily glean certain information about the eigenvalues of \mathbf{A} from T and D . Of particular interest is the sign of the real part of each eigenvalue (or just the sign if the eigenvalue is a real number). In what follows we assume that $D \neq 0$ and $D \neq T^2/4$; these are addressed in Reading Exercises B.4.1 and B.4.2.

Theorem B.4.1 — Trace-Determinant Analysis for Eigenvalues. Let \mathbf{A} be a 2×2 matrix with real entries, $D = \det(\mathbf{A})$ and $T = \text{trace}(\mathbf{A})$. Then the eigenvalues of \mathbf{A} fall into the following cases:

1. $D < 0$: In this case both eigenvalues in (B.17) are real numbers, with $\lambda_1 > 0$ and $\lambda_2 < 0$.
2. $0 < D < T^2/4$: In this case both eigenvalues are real and both are of the same sign as T .
3. $0 \leq T^2/4 < D$: In this case both eigenvalues are complex (and conjugate to one another), and both have real part with the same sign as T . If $T = 0$ both eigenvalues are purely imaginary.

These conditions are frequently easier to check than actually computing the eigenvalues explicitly, especially when the matrix contains unspecified parameters, and usually give us the information we need to determine the stability of equilibrium points for a pair of ODEs in the plane.

Proof of the Various Cases

To prove the assertions, first consider the case $D < 0$. Then $T^2 - 4D > T^2 \geq 0$, so $\sqrt{T^2 - 4D}$ is real and from (B.17) it follows that both eigenvalues are real numbers. Also, because $0 < -4D$ we have $T^2 < T^2 - 4D$, and so $|T| = \sqrt{T^2} < \sqrt{T^2 - 4D}$. This last inequality $|T| < \sqrt{T^2 - 4D}$ is entirely equivalent to

$$-\sqrt{T^2 - 4D} < T < \sqrt{T^2 - 4D}. \quad (\text{B.18})$$

The inequality (B.18) holds for any T when $D < 0$. The leftmost inequality in (B.18) is equivalent to $0 < T + \sqrt{T^2 - 4D}$, so from (B.17) it follows that $\lambda_1 > 0$. The rightmost inequality in (B.18) is equivalent to $T - \sqrt{T^2 - 4D} < 0$, so from (B.17) it follows that $\lambda_2 < 0$. This proves Case (1) above.

Let's now examine Case (2) in which $0 < D < T^2/4$. This is equivalent to $0 < 4D < T^2$ or $0 < T^2 - 4D$. Thus $\sqrt{T^2 - 4D}$ is real and so both eigenvalues in (B.17) are real. Since $D > 0$ we also see that $T^2 - 4D < T^2$, so all in all $0 < T^2 - 4D < T^2$. Taking square roots throughout and using $|T| = \sqrt{T^2}$ shows that $0 < \sqrt{T^2 - 4D} < |T|$. This last inequality is entirely equivalent to

$$0 < \sqrt{T^2 - 4D} < |T|. \quad (\text{B.19})$$

When $T > 0$ it's easy to see that $\lambda_1 > 0$, and in this case (B.19) becomes $\sqrt{T^2 - 4D} < T$ so that $T - \sqrt{T^2 - 4D} > 0$ and $\lambda_2 > 0$ also; both eigenvalues are thus positive. If $T < 0$ it's clear that $\lambda_2 < 0$, and in this case (since $|T| = -T$) the inequality (B.19) becomes $\sqrt{T^2 - 4D} < -T$ or equivalently, $T + \sqrt{T^2 - 4D} < 0$ and $\lambda_1 < 0$ also; thus both eigenvalues are negative. This proves Case (2).

Finally, consider Case (3) in which $0 < T^2/4 < D$. Then $T^2 - 4D < 0$ and $\sqrt{T^2 - 4D}$ is purely imaginary, and can be expressed as $i\sqrt{4D - T^2}$. From (B.17) the eigenvalues can then be written as

$$\frac{T}{2} \pm i \frac{\sqrt{4D - T^2}}{2}.$$

Both are complex, conjugate to one another, and have real part $T/2$, which has the same sign as T . If $T = 0$ the eigenvalues are purely imaginary.

Reading Exercise B.4.1 Argue that if $D = \det(\mathbf{A}) = 0$ then one eigenvalue equals 0 and the other equals T .

Reading Exercise B.4.2 Argue that if $D = T^2/4$ then both eigenvalues are real and equal to $T/2$.

Eigenvalues and Eigenvectors for Larger Matrices

The computation of the eigenvalues and eigenvectors for matrices of dimension 3×3 and larger is usually difficult to do by hand. The characteristic polynomial for an $n \times n$ matrix is of degree n , and so even the eigenvalues of a 3×3 matrix are the roots of a cubic polynomial. They would probably be easiest to compute numerically. The situation for 4×4 and larger matrices is even worse.

Here are a few general truths concerning eigenvalues, eigenvectors, and their computation. The interested reader should consult [19] for additional material on basic linear algebra, or [18] for the more algorithmic aspects of eigenvalue computation.

1. The characteristic polynomial for an $n \times n$ matrix \mathbf{A} is of degree n , and the eigenvalues of \mathbf{A} are the roots of this polynomial. Thus \mathbf{A} has n eigenvalues, if we count multiplicity. If λ is a root of multiplicity m of the characteristic polynomial then λ is said to have **algebraic multiplicity m** . In Reading Exercises B.3.4 and B.3.5, for example, $\lambda = 2$ was an eigenvalue of algebraic multiplicity 2 for the relevant matrices.
2. Each eigenvalue (which may have algebraic multiplicity of one or higher) has at least one eigenvector, but recall that scalar multiples of an eigenvector do not count as additional eigenvectors.
3. The computation of eigenvalues is a large area of research in numerical mathematics. These computations are not generally done through characteristic polynomials. The approach used depends very much on what is needed. Do we want just eigenvalues, or both eigenvalues and eigenvectors? Do we want all of the eigenvalues, or just some specified subset? Does the matrix \mathbf{A} have any special properties? For the light-duty needs of this text, e.g., 4×4 or smaller matrices, using any software package like Maple, Mathematica, Matlab, or Sage will work.

B.5 Diagonalization

This material is needed only for analyzing the matrix exponential in Section 6.4.

Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues, real or complex, of an $n \times n$ matrix \mathbf{A} . Let \mathbf{v}_k be an eigenvector with eigenvalue λ_k for $1 \leq k \leq n$. Suppose it is possible to choose these eigenvectors so that the $n \times n$ matrix

$$\mathbf{P} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_n] \quad (\text{B.20})$$

(with k th column equal to \mathbf{v}_k) is invertible. In this case we say that the matrix \mathbf{A} is **diagonalizable**, for reasons that will appear shortly. Also let \mathbf{D} be the diagonal matrix

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & \lambda_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix} \quad (\text{B.21})$$

formed from the eigenvalues strung down the diagonal, in the same order that the eigenvectors appear in \mathbf{P} .

Remark B.5.1 If there exist eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ such that \mathbf{P} in (B.20) is invertible then by Theorem B.2.1 the only solution to $\mathbf{P}\mathbf{c} = \mathbf{0}$ is to take $\mathbf{c} = \mathbf{0}$. In this case we say that the vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ are **linearly independent**. Conversely, if $\mathbf{c} = \mathbf{0}$ is the only solution to $\mathbf{P}\mathbf{c} = \mathbf{0}$ (that is, the vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ are linearly independent) then by Theorem B.2.1 \mathbf{P} is invertible.

■ **Example B.9** Let

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}.$$

The eigenvalues $\lambda_1 = 4$ and $\lambda_2 = -2$ and corresponding eigenvectors

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

were computed in Example B.5. In this case

$$\mathbf{P} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} 4 & 0 \\ 0 & -2 \end{bmatrix}.$$

The matrix \mathbf{P} is invertible since $\det(\mathbf{P}) = -2 \neq 0$. ■

Theorem B.5.1 — Diagonalizable Matrices. If \mathbf{A} is diagonalizable then

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1} \quad (\text{B.22})$$

where \mathbf{P} is as given in (B.20) and \mathbf{D} is as given in (B.21).

To prove this, rewrite (B.22) in the equivalent form

$$\mathbf{AP} = \mathbf{PD} \quad (\text{B.23})$$

obtained by multiplying both sides of (B.22) by \mathbf{P} on the right. From the definition of matrix multiplication and the definition (B.20) of \mathbf{P} the left side can be computed as

$$\begin{aligned} \mathbf{AP} &= \mathbf{A}[\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_n] \\ &= [\mathbf{Av}_1 \ \mathbf{Av}_2 \ \cdots \ \mathbf{Av}_n] \\ &= [\lambda_1\mathbf{v}_1 \ \lambda_2\mathbf{v}_2 \ \cdots \ \lambda_n\mathbf{v}_n] \end{aligned} \quad (\text{B.24})$$

where the last line follows since $\mathbf{A}\mathbf{v}_k = \lambda_k \mathbf{v}_k$; the \mathbf{v}_k are eigenvectors. In summary, multiplying \mathbf{P} by \mathbf{A} multiplies the k th column of \mathbf{P} by λ_k .

Now consider the right side of (B.23). The k th column of the product \mathbf{PD} can be computed as \mathbf{P} times the k th column of \mathbf{D} , which is

$$[\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_n] \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \lambda_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

This is the product of an $n \times n$ matrix times an n -dimensional vector. From the definition of matrix-vector multiplication (B.2) it is not hard to see that this product is precisely $\lambda_k \mathbf{v}_k$, which is the k th column of \mathbf{AP} as given by (B.24). This shows that (B.23), $\mathbf{AP} = \mathbf{PD}$, holds. If we multiply both sides of this last equation on the right by \mathbf{P}^{-1} (recall \mathbf{P} is invertible) we obtain (B.22).

■ **Example B.10** To illustrate the equation $\mathbf{AP} = \mathbf{PD}$, consider the matrix \mathbf{A} in Example B.9, with \mathbf{P} and \mathbf{D} as defined there. We can compute

$$\mathbf{AP} = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 4 & -2 \\ 4 & 2 \end{bmatrix}$$

and

$$\mathbf{PD} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} 4 & -2 \\ 4 & 2 \end{bmatrix}.$$

Since \mathbf{P} is invertible this also shows that $\mathbf{A} = \mathbf{PDP}^{-1}$. ■

The utility of diagonalizing matrices is illustrated in Section 6.4.

B.6 Additional Exercises

Exercise B.6.1 Compute the eigenvalues and eigenvectors for each matrix. Then diagonalize the matrix by writing out \mathbf{P} and \mathbf{D} as in (B.20) and (B.21). You have considerable flexibility in choosing \mathbf{P} . Verify that $\mathbf{A} = \mathbf{PDP}^{-1}$.

(a)

$$\mathbf{A} = \begin{bmatrix} -1 & -4 \\ -3 & -2 \end{bmatrix}$$

(b)

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 25 & 0 \end{bmatrix}$$

(c)

$$\mathbf{A} = \begin{bmatrix} -3 & 1 \\ 0 & 2 \end{bmatrix}$$

(d)

$$\mathbf{A} = \begin{bmatrix} 3 & 12 \\ 1 & -1 \end{bmatrix}$$

(e)

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

(f)

$$\mathbf{A} = \begin{bmatrix} 7 & 5 \\ -9 & 1 \end{bmatrix}$$

(g)

$$\mathbf{A} = \begin{bmatrix} -1 & 0 & 0 \\ 1 & 2 & 0 \\ 5 & -3 & 1 \end{bmatrix}$$

C. Circuits

C.1 Current, Voltage, and Resistance

This appendix provides additional detail on how ODEs can be used to model basic circuits. The goal here is not an exhaustive treatment of electrical circuits, but just enough information to understand how differential equations can be used to model simple circuits containing resistors, capacitors, inductors, and voltage sources. We'll also see how, in certain situations, one can analyze circuits without actually writing down any ODEs, by using the notion of impedance.

A First Example

Let's start with a very simple circuit containing a **voltage source** and **resistor**, as illustrated in Figure C.1. The main physical quantities of interest are the **current** through the circuit and the **voltage** differences between any two points in the circuit. The current through a wire at any position is the net rate at which **electric charge** is flowing past that position in some reference direction. The voltage between two points measures the change in **electric potential** between those points or how much work is done moving a unit charge from one of the points to the other. A simple and intuitive way to think about the situation is to consider the wire as a pipe and electric charge as water (but with no mass). Then current is analogous to the water flow rate, mass per time, past a point in the pipe. Voltage is like pressure. In a pipe it is differences in pressure that induce water to flow. In a circuit it is a potential or voltage difference that induces electric current to flow. Voltages are always measured between two points in a circuit; it doesn't make sense to talk about the "voltage at a point" in a circuit unless a second reference point is understood. Frequently such a point is chosen and deemed to be at 0 volts, and is called a **ground point** or just **ground**. Such a ground may be chosen at any point in the circuit, based purely on convenience. The voltage source may be time-dependent, though for this first example it won't make any difference.

What is the nature of this electric charge that's flowing through the wires? In reality it consists of conduction electrons, negatively charged and loosely bound to the atoms of the wire, that can flow by hopping from atom to atom in the wire. A voltage difference between two points in the circuit generates an electric field that pushes on these electrons and imparts a net flow of negative charge through the wire. However, the conventional current model of electrical conduction posits

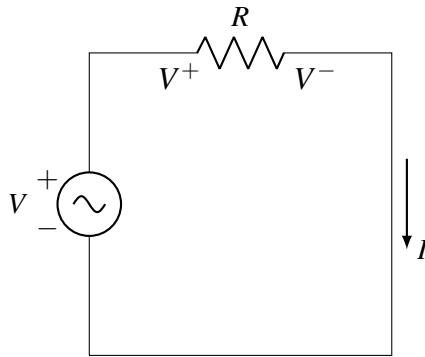


Figure C.1: Voltage source-resistor series circuit.

that current consists of positive charges that flow from higher potential to lower potential, just as water would flow from a region of higher pressure to a region of lower pressure. The flow of positive charge in this model corresponds to the flow of negative charge (electrons) moving in the opposite direction in the actual wire. In elementary circuit analysis this doesn't really change anything—we still obtain the correct answers for voltages, currents, etc. Current is measured in units of **amperes** or just **amps** and voltage or potential differences in **volts**. One ampere corresponds to one coulomb of charge per second flowing the wire, and one coulomb is approximately 6.24×10^{18} elementary charges, e.g., electrons.

Electric charge has its own physical dimension, independent of mass, length, or time, and we use Q to denote this physical dimension. In this case current has the dimension of QT^{-1} , charge per time. Voltage has the dimension of work per charge, which is $ML^2T^{-2}Q^{-1}$.

Kirchhoff's and Ohm's Laws

There are three essential tools to analyze the circuit of Figure C.1:

1. **Kirchhoff's voltage law:** The sum of all the voltage differences around a closed loop in a circuit must be zero.
2. **Kirchhoff's current law:** The current through a wire is the same at all points in the wire (at least at low frequencies) and more generally, at any junction where several wires meet, the net current into (or out of) that junction is zero. This is a consequence of the conservation of electric charge, and that conduction charges cannot pile up anywhere in the wire.
3. **Ohm's law:** In an ideal resistor the current through the resistor is proportional to the voltage drop across the resistor, with the current flowing from the higher potential side to the lower potential side. The constant of proportionality is called the **resistance** of the resistor and is measured in units known as **ohms**. Ohm's law is illustrated in Figure C.2, in which the resistor (the zigzag circuit element) is shown in isolation and we have the relation

$$V^+ - V^- = IR. \quad (\text{C.1})$$

A fluid analogue that illustrates Ohm's law is this: think of the resistor as a pipe of some diameter, $V^+ - V^-$ as the pressure difference across the resistor, and I as the rate at which water flows through the resistor. We can write (C.1) as $I = (V^+ - V^-)/R$; if R is large (the pipe has a small diameter) then for any pressure difference $V^+ - V^-$ the flow rate (current) will be relatively small. But if R is small (a pipe with large diameter) even a small pressure difference causes a lot of water to flow. In order for (C.1) to be dimensionally consistent, resistance must have dimension $[R] = ML^2T^{-1}Q^{-2}$; see Reading Exercise 2.1.10.

With these three principles we can determine the voltage and current at any point in the circuit of Figure C.1. We assume the wires themselves are perfect conductors, that is, they have no electrical

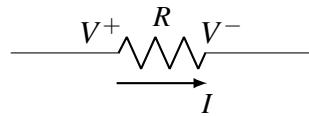


Figure C.2: Resistor schematic and Ohm's Law.

resistance. A consequence of this assumption is that the voltage is the same at all points in a perfect conductor.

Circuit Analysis

To analyze the circuit of Figure C.1, let us designate one side of the voltage source (arbitrarily) as the positive or plus side, the other as the negative or minus side, and in fact we shall designate the minus side of the voltage source as ground, at 0 volts. We take the clockwise direction around this circuit loop as the positive direction for current flow. We'll start on the minus side of the voltage source V and traverse the circuit in the clockwise direction. As we traverse each component in the circuit we will compute the change in voltage over that component. According to Kirchhoff's voltage law, when we return to the starting point the sum of the voltage changes must be zero.

The change in voltage as we step over the voltage source from the minus side at 0 volts to the plus side at V volts is V . This is also the voltage V^+ on the left side of the resistor. If V^- denotes the voltage on the right side of the resistor then from Ohm's law (C.1) we have $V^+ - V^- = IR$ where I is the current through the resistor from the left side to the right and R is the resistance of the resistor; $V^+ - V^-$ is the voltage change as we step over the resistor. As we move back to the minus side of the voltage source we find that $V^- = 0$ (remember, the wires have no resistance, so any two points connected by a wire have the same potential). From $V^+ = V$, $V_r = 0$, and equation (C.1) we find

$$V = IR. \quad (\text{C.2})$$

In summary, the voltage at all points in the wire connecting the plus side of the voltage source to the resistor is V (relative to the ground on the minus side of the voltage source), while the voltage at all points in the wire connecting the resistor to the ground side of the voltage source is 0. We can use (C.2) to determine the current at all points in the circuit as $I = V/R$, with clockwise being the positive direction.

If the voltage source is time dependent, so $V = V(t)$, this analysis still holds. In this case the current through the circuit is also time-dependent and (C.2) becomes $V(t) = I(t)R$.

C.2 Capacitors

Capacitors store electric charge, in the simplest case on two closely spaced conductive plates separated by a nonconductive space, e.g., air. One plate collects positive charge, the other collects an equal negative charge (negative charge here can also be viewed as a deficit of positive charge). This occurs when a voltage difference V is applied across the capacitor plates; the resulting potential difference pushes positive charge onto one plate and negative charge onto the other (equivalently, pulls positive charge from this plate). In an ideal capacitor the amount q of charge stored ($+q$ on one plate, $-q$ on the other) is

$$q = CV \quad (\text{C.3})$$

where C is the capacitance of the capacitor, measured in the SI unit **farads**, and $V = V^+ - V^-$ is the potential difference across the capacitor, as illustrated in Figure C.3. The higher potential V^+

side of the capacitor has the positive charge, the lower V^- side carries the negative charge. In order for (C.3) to be dimensionally consistent, capacitance must have dimension $[C] = M^{-1}L^{-2}T^2Q^2$; see Reading Exercise 2.1.10. The current $I(t)$ going into the left side of the capacitor equals the current $I(t)$ leaving the right side, thus the net charge on the capacitor is always zero.

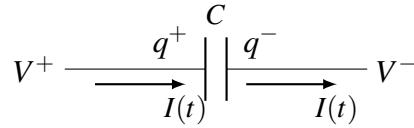


Figure C.3: Capacitor schematic.

A mechanical or fluid analogy that may be helpful is illustrated in Figure C.4. In this figure the uncharged capacitor is shown in the left panel. It consists of a tank with an inlet pipe (wire) on the left and an outlet pipe (wire) on the right separated by a thin massless divider (dark grey) that is attached to one of the walls by a spring, although the designation of which pipe is the inlet and which is the outlet is somewhat arbitrary. When the capacitor is uncharged the spring is at its equilibrium position; this occurs when there is no pressure differential between the inlet and outlet.

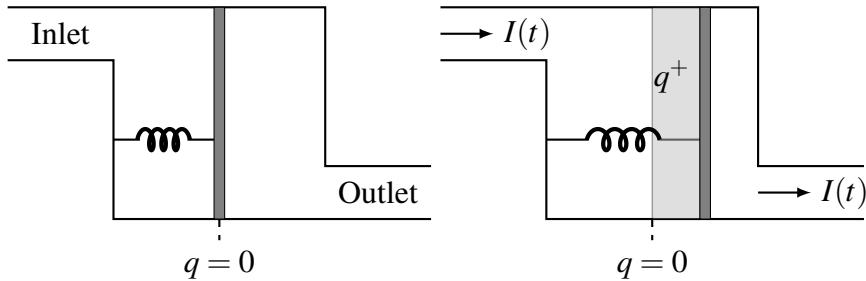


Figure C.4: Mechanical/fluid analogue to a capacitor, uncharged (left) and charged to $q = q^+$ (right).

The right panel in Figure C.4 shows the capacitor in the process of charging, when it carries a positive charge of q^+ (shaded) on the left side of the capacitor and a corresponding deficit of positive charge (effectively, a negative charge) on the right side. This situation occurs when a pressure (voltage) differential exists from the inlet to the outlet, causing water (positive charge) to flow into the left side of the capacitor and an equal amount to exit the right side of the capacitor. Since water is effectively incompressible, the amount that enters on the left always equals the amount that exits on the right, and the situation for electrical charge is the same. Also note that no water ever actually traverses the central divider, but merely displaces it. The divider will be displaced until the force exerted by the spring is sufficient to oppose the force exerted by the pressurized fluid entering on the left, at which point no more water/charge will flow into the capacitor. If a higher pressure differential is applied between the inlet and outlet then the divider is pushed farther to the right, and more water is pushed into the left side of the capacitor (and more exits the right).

The right panel of Figure C.4 also makes it easy to see that the rate at which water (positive charge) q is increasing in the left side of the capacitor (the gray shaded region) equals the rate at which water/charge is entering the left side of the capacitor. That is,

$$\frac{dq}{dt} = I. \quad (\text{C.4})$$

Finally, in our mechanical or fluid analogy for the capacitor, the inlet and outlet are effectively reversible. This is true for many types of electrical capacitors, but not all.

An RC Circuit

Consider the simple **RC circuit** in Figure C.5 with voltage source $V(t)$, where we'll explicitly assume that this source may depend on time t . Since $V(t)$ is changing in time we should expect all quantities associated with the circuit, e.g., voltages and currents, to also change with time.

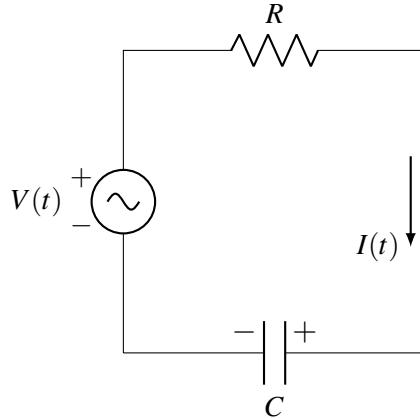


Figure C.5: Single loop RC series circuit.

To analyze this circuit we need one additional observation, specifically that if $q(t)$ denotes the amount of charge on the capacitor (positive charge on the side labeled with a plus sign, negative on the side labeled with a negative sign) and $I(t)$ denotes the current in the circuit (with $I > 0$ indicating clockwise flow of positive charge in Figure C.5) then (C.4) holds. We do have to be a bit careful with plus and minus signs in (C.4) and throughout. As a sanity check, suppose the amount of positive charge on the side of the capacitor labeled plus in Figure C.5 is increasing. This means that positive charge is flowing into that side, which is in accord with our choice of $I > 0$ as clockwise.

To analyze the behavior of the circuit, let's start at the minus or ground side of the source $V(t)$, designated as 0 V, and move clockwise around the circuit, back to our starting point. The net change in voltage over each component yields a net voltage change $V(t) - RI(t) - q(t)/C$ around the circuit (the voltage source, then the resistor, then the capacitor) which must be zero, so $V(t) - RI(t) - q(t)/C = 0$ or

$$RI(t) + \frac{q(t)}{C} = V(t).$$

If we now make use of (C.4) we arrive at

$$Rq'(t) + \frac{q(t)}{C} = V(t) \quad (\text{C.5})$$

a first-order differential equation for $q(t)$, the charge on the capacitor. We also need an initial condition, for example, $q(0) = 0$.

■ **Example C.1** Suppose $R = 8$ ohms and $C = 1.0 \times 10^{-3}$ F, with $q(0) = 0$. Let $V(t) = 5$ volts. Then

$$8q'(t) + 1000q(t) = 5$$

with $q(0) = 0$. The solution is $q(t) = (1 - e^{-125t})/200$. From (C.4) we can compute that $I(t) = 0.625e^{-125t}$, and from this one can use Ohm's law (C.2) to find the voltage across the resistor, or (C.3) to find the voltage across the capacitor at any time. ■

More generally, if $V(t)$ is constant in (C.5) and $q(0) = 0$ we find that the solution is $q(t) = VC(1 - e^{-t/(RC)})$ and the capacitor has charged to within one percent of its final value (VC) by time $t \approx 5RC$ (since $1 - e^{-5} \approx 0.993$). The product RC has the dimension of time and is referred to as the **RC time constant** for the circuit. The argument $-t/RC$ of the exponential function is then dimensionless.

■ **Example C.2** Consider an RC circuit with voltage source $V(t) = \cos(\omega t)$ for some frequency ω . The ODE (C.5) becomes

$$Rq'(t) + \frac{q(t)}{C} = \cos(\omega t).$$

For an initial condition $q(0) = q_0$ the solution is

$$q(t) = \underbrace{De^{-t/(RC)}}_{\text{transient}} + \underbrace{A \cos(\omega t) + B \sin(\omega t)}_{\text{periodic}} \quad (\text{C.6})$$

where

$$D = q_0 - \frac{C}{1 + C^2 R^2 \omega^2}, \quad A = \frac{C}{1 + C^2 R^2 \omega^2}, \quad B = \frac{C^2 R \omega}{1 + C^2 R^2 \omega^2}.$$

The first term in (C.6) is transient and dies out after about $t > 5RC$, regardless of q_0 . The remaining portion is periodic. If we look at the current through the circuit, after the transients have died out, we find that the periodic current $I_{per}(t)$ is

$$I_{per}(t) = -A\omega \sin(\omega t) + B\omega \cos(\omega t) = \frac{C^2 \omega^2 R}{1 + C^2 R^2 \omega^2} \cos(\omega t) - \frac{C\omega}{1 + C^2 R^2 \omega^2} \sin(\omega t).$$

■

In Example C.2 if ω is very large it's easy to see that the coefficient of the $\sin(\omega t)$ term is near zero, while the coefficient of the $\cos(\omega t)$ term is about $1/R$. That is, the current is $I_{per}(t) \approx V(t)/R$, which is exactly what we'd get if there was no capacitor present in the circuit. Contrast this to Example C.1 in which V was constant, the ultimate low frequency $\omega = 0$. In that example the current was asymptotically zero as t increases. That is, the periodic response was zero. Informally, capacitors permit little current to flow at low frequencies, but as frequency increases the capacitor acts more and more like a perfect conductor.

C.3 Inductors

The final circuit component of interest is the **inductor**. In a circuit inductors appear as illustrated in Figure C.6. For an inductor the relation between the potential difference $V = V^+ - V^-$ across the inductor and the current through the inductor is

$$V = L \frac{dI}{dt}, \quad (\text{C.7})$$

where L is called the **inductance** of the inductor. Inductors are, in their simplest form, just coils of wire. The inductance L depends on the size and geometry of the inductor, among other things. The voltage-current relation (C.7) shows that it takes very little voltage difference to induce current through an inductor, as long as the current is not changing rapidly. Conversely, a large voltage difference is needed to push a rapidly changing current through an inductor. In order for (C.7) to be dimensionally consistent, inductance must have the dimension $ML^2 Q^{-2}$. (Note we are using L for both inductance and the dimension of length, so we should be careful.)

A mechanical analogue for an inductor is shown in Figure C.7. The fluid version consists of a paddle wheel of some substantial mass (this mass is like the inductance L) enclosed in a housing



Figure C.6: Inductor schematic.

with an inlet and outlet as illustrated (the inlet/outlet are interchangeable). The wheel spins without friction. If the incoming fluid flow rate $I(t)$ is not changing then essentially no pressure differential between the inlet and outlet is needed to maintain the flow. If, however, we wish to increase $I(t)$ then we need to invest energy to speed up the rotation of the paddle wheel, and doing this work requires applying a pressure (voltage) differential across the inlet/outlet. Similar reasoning shows that to decrease $I(t)$ we need to slow down the wheel, and so apply a negative pressure differential (higher at the outlet, lower at the inlet) in order to accomplish this. The larger the mass (or the larger the inductor) the more work is required.

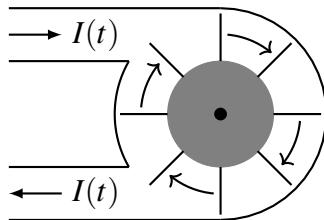


Figure C.7: Mechanical/fluid analogue to an inductor.

An RL Circuit

To understand the operation of an inductor in a circuit, let's consider a simple **RL circuit** consisting of a resistor R in series with an inductor, as illustrated in Figure C.8, analogous to the RC circuit of Figure C.5. However, the behavior of this circuit is quite different.

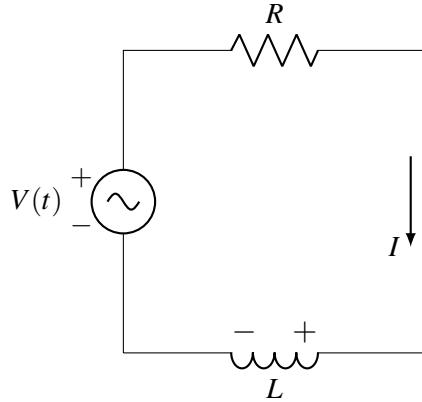


Figure C.8: Single loop RL series circuit.

We start at the minus side of the voltage source $V(t)$, designated as 0 V, and move clockwise around the circuit, back to our starting point. The net change in voltages over each component is $V(t) - RI(t) - LI'(t)$ (the voltage source, then the resistor, then the inductor, where we make use of (C.7)), and by Kirchhoff's voltage law this net change must be zero, so $V(t) - RI(t) - LI'(t) = 0$ or

$$LI'(t) + RI(t) = V(t), \quad (\text{C.8})$$

a first-order differential equation for $I(t)$, the current in the circuit. We also need an initial condition, say $I(0) = I_0$.

■ **Example C.3** Suppose $R = 10$ ohms and $L = 1.0 \times 10^{-3}$ henries, with $I(0) = 0$. Let $V(t) = 5$ volts. Then

$$0.001I'(t) + 10I(t) = 5$$

with $I(0) = 0$. The solution is $I(t) = \frac{1}{2}(1 - e^{-10000t})$. ■

■ **Example C.4** Consider an RL circuit with voltage source $V(t) = \cos(\omega t)$ for some frequency ω , analogous to Example C.2, but with an inductor replacing the capacitor. The ODE (C.8) becomes

$$LI'(t) + RI(t) = \cos(\omega t).$$

For an initial condition $I(0) = I_0$ the solution is

$$I(t) = \underbrace{De^{-Rt/L}}_{\text{transient}} + \underbrace{A \cos(\omega t) + B \sin(\omega t)}_{\text{periodic}} \quad (\text{C.9})$$

where

$$D = I_0 - \frac{R}{L^2\omega^2 + R^2}, \quad A = \frac{R}{L^2\omega^2 + R^2}, \quad B = \frac{L\omega}{L^2\omega^2 + R^2}.$$

The first term in (C.9) is transient and dies out, regardless of the value of I_0 . The remaining portion is periodic and persists for as long as the voltage source is active. ■

In Example C.4 if $\omega = 0$ then the periodic portion of the current is $I(t) = 1/R$, precisely the same current that would be obtained with no inductor present. But as $\omega \rightarrow \infty$ the periodic portion of the current is $I(t) = 0$. At low frequencies the inductor allows current to pass (when $\omega = 0$, as if the inductor was not even present), but at high frequencies the inductor blocks the flow of current. This is in contrast to capacitors, which block low frequencies and allow high frequencies to pass.

C.4 RLC Circuits

Consider the circuit shown in Figure C.9, a single loop that contains a resistor, inductor, and capacitor in series, a so-called **RLC circuit**.

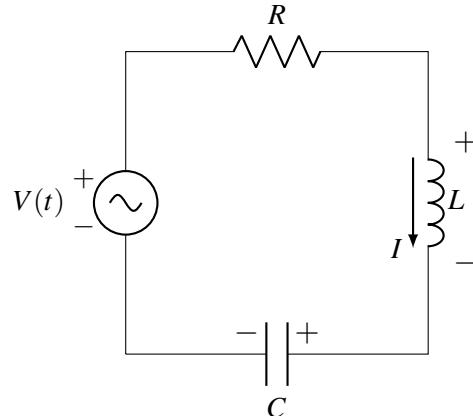


Figure C.9: Single loop RLC series circuit.

Let $q(t)$ denote the charge on the capacitor and $I(t)$ the current in the circuit as indicated. If we traverse the circuit clockwise starting at the ground side of the source $V(t)$ and add up the voltage changes while making use of Kirchhoff's voltage law and the voltage-current relations for each component we find that $V(t) - RI(t) - LI'(t) - q(t)/C = 0$ or

$$LI'(t) + RI(t) + q(t)/C = V(t). \quad (\text{C.10})$$

Now use $I(t) = q'(t)$ (and so $I'(t) = q''(t)$) in (C.10) to obtain

$$Lq''(t) + Rq'(t) + q(t)/C = V(t), \quad (\text{C.11})$$

a second order, linear, constant coefficient ODE that is identical to the mass-spring equations of Chapter 4. This second-order ODE needs two initial conditions, say of the form $q(0) = q_0$ and $I(0) = I_0$ (note that $I(0) = q'(0)$). We can solve (C.11) to find $q(t)$, from which we can compute the current $I(t) = q'(t)$ and then use (C.2), (C.3), and (C.7) to find the voltage across any of the components in the circuit at any time.

■ **Example C.5** Suppose $R = 2$ ohms, $C = 1.0 \times 10^{-6}$ F, and $L = 5 \times 10^{-5}$ h, with $V(t) = \cos(\omega t)$ for some ω . The ODE (C.11) is

$$(5 \times 10^{-5})q''(t) + 2q'(t) + 10^6q(t) = \cos(\omega t).$$

As with a damped spring-mass system, the solution will consist of a transient portion that satisfies the homogeneous ODE, plus a periodic portion. The characteristic equation for the homogeneous ODE is $(5 \times 10^{-5})r^2 + 2r + 10^6 = 0$ with roots $r \approx -20000 \pm (1.4 \times 10^5)i$. This tells us that the transient contains terms of the form $e^{-20000t}$ (times sines and cosines) and should substantially die out within about $t \approx 5/20000 = 1/4000$ th of a second. The system's natural oscillation frequency is determined by the imaginary part of the roots and is $\omega = 1.4 \times 10^5$ radians per second, about $(1.4 \times 10^5)/(2\pi) \approx 22,281$ hz.

Suppose we're interested in the long-term periodic portion of the solution, after transients have died out. We can find this using undetermined coefficients. If

$$q_{per}(t) = A \cos(\omega t) + B \sin(\omega t)$$

we find that in this case that

$$A = \frac{2 \times 10^{10} - \omega^2}{\omega^4 - 3.84 \times 10^{10}\omega^2 + 4 \times 10^{20}}, \quad B = \frac{8 \times 10^8}{\omega^4 - 3.84 \times 10^{10}\omega^2 + 4 \times 10^{20}}.$$

The corresponding periodic current is just dq_{per}/dt or

$$I_{per}(t) = -A\omega \sin(\omega t) + B\omega \cos(\omega t).$$

The amplitude $F(\omega)$ of the periodic current response as a function of ω is

$$F(\omega) = \omega \sqrt{A^2 + B^2} = \frac{20000\omega}{\sqrt{\omega^4 - 3.84 \times 10^{10}\omega^2 + 4 \times 10^{20}}}.$$

with the graph shown in Figure C.10. The circuit responds more favorably (a larger current flows) when the voltage source drives the circuit at certain frequencies, just like a driven, underdamped spring-mass system. ■

Reading Exercise C.4.1 A series RLC circuit in the configuration of Figure C.9 has $L = 1.0 \times 10^{-3}$ henry, $R = 4$ ohms, $C = 5.0 \times 10^{-6}$ farad, and a 5 volt source, so $V(t) = 5$. At time $t = 0$ the capacitor is uncharged and no current flows in the circuit. Set up and solve the relevant second-order ODE (C.11) for $q(t)$, then compute the current $I(t)$ through the circuit, and plot $I(t)$ from $t = 0$ to $t = 0.005$. Is this system overdamped or underdamped?

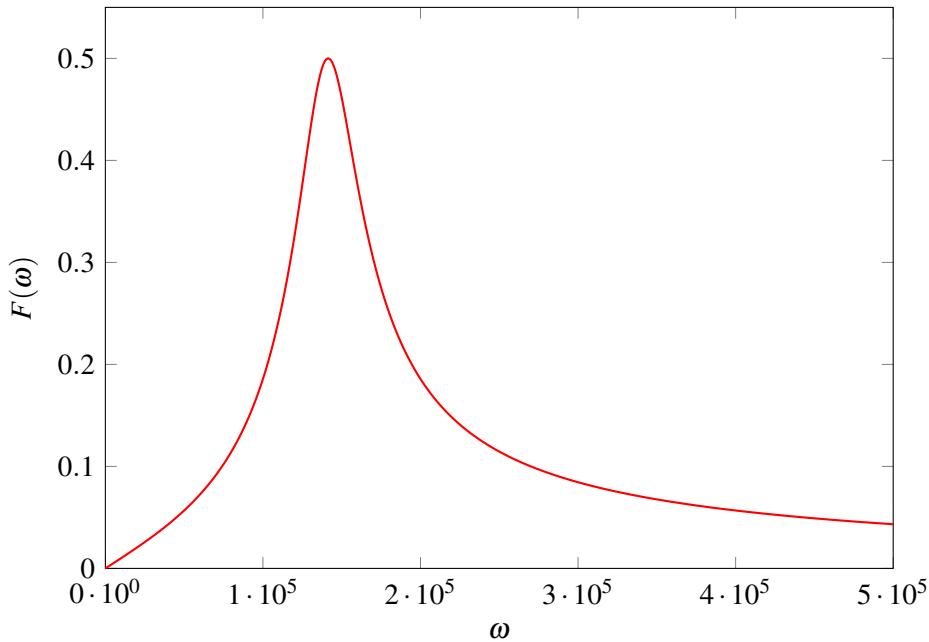


Figure C.10: Magnitude of periodic current through RLC circuit as a function of driving frequency.

Reading Exercise C.4.2 Repeat Reading Exercise C.4.1 if $V(t) = \cos(2000t)$ (starting at time $t = 0$), and plot from $t = 0$ to $t = 0.01$. Identify the transient and periodic portions of $I(t)$, both visually and in the formula for the solution.

Reading Exercise C.4.3 Repeat Reading Exercise C.4.1, but suppose now that the voltage source is

$$V(t) = \begin{cases} 0, & t < 2 \\ 5, & t \geq 2 \end{cases}$$

This might correspond to a switch being closed at time $t = 2$ and completing the circuit. Write down an ODE for $q(t)$, the charge on the capacitor; it should involve one or more Heaviside functions, and of course the Laplace transform provides a convenient solution method. Solve the ODE and plot the current $I(t)$ from $t = 1.999$ to $t = 2.005$ (just before the switch is closed to 0.005 seconds after the circuit is closed).

C.5 Complex-Valued Solutions and Periodic Forcing

Up this point when discussing the periodic forcing of systems we've used $\cos(\omega t)$, or $\sin(\omega t)$, or some linear combination thereof. The resulting periodic solutions are of the same form. Trigonometric functions have an intuitive appeal, but the resulting computations are unnecessarily messy. Using complex exponentials makes things easier, especially for circuit analysis, once you get used to it.

Suppose we have a linear differential equation, say of the form

$$ax''(t) + bx'(t) + cx(t) = f(t) \tag{C.12}$$

for some real numbers a, b, c , although what we're about to do works for linear constant coefficient ODEs of any order. Suppose that f is a complex-valued function of t , say $f(t) = f_r(t) + if_i(t)$ for some real-valued functions f_r and f_i . What does it mean to say that a function $x(t)$ is a solution to (C.12) when $f(t)$ is complex-valued?

First, the solution $x(t)$ will itself have a real and an imaginary part, say $x(t) = x_r(t) + ix_i(t)$, where $x_r(t)$ and $x_i(t)$ are real-valued functions of t . A straightforward computation shows that

$$\begin{aligned} ax''(t) + bx'(t) + cx(t) &= a(x_r''(t) + ix_i''(t)) + b(x_r'(t) + ix_i'(t)) + c(x_r(t) + ix_i(t)) \\ &= ax_r''(t) + bx_r'(t) + cx_r(t) + i(ax_i''(t) + bx_i'(t) + cx_i(t)) \\ &= f_r(t) + if_i(t). \end{aligned}$$

This makes it clear that the complex-valued $x(t)$ is a solution to the ODE if and only if each of

$$\begin{aligned} ax_r''(t) + bx_r'(t) + cx_r(t) &= f_r(t) \\ ax_i''(t) + bx_i'(t) + cx_i(t) &= f_i(t) \end{aligned}$$

holds. These are just the original ODE (C.12) but applied to the real and imaginary parts of $x(t)$ separately.

■ **Example C.6** Consider the ODE

$$x'(t) + 2x(t) = f(t)$$

with $f(t) = t + ie^{-t}$ and initial condition $x(0) = 2 + 3i$. The real part $x_r(t)$ of $x(t) = x_r(t) + ix_i(t)$ must satisfy $x_r'(t) + 2x_r(t) = t$ with $x_r(0) = 2$ and the imaginary part $x_i(t)$ must satisfy $x_i'(t) + 2x_i(t) = e^{-t}$ with $x_i(0) = 3$; be careful here, $x_i(0) = 3$, not $3i$. You can use any standard solution technique to find that

$$x_r(t) = \frac{t}{2} - \frac{1}{4} + \frac{9}{4}e^{-2t} \quad \text{and} \quad x_i(t) = e^{-t} + 2e^{-2t}.$$

Again, be careful: the $x_i(t)$ is the imaginary part of the solution, but does not itself include the multiplication by i . Then

$$x(t) = \frac{t}{2} - \frac{1}{4} + \frac{9}{4}e^{-2t} + i(e^{-t} + 2e^{-2t}).$$

■

■ **Example C.7** Consider using a forcing function of the form $V(t) = V_0e^{i\omega t}$ in equation (C.11), say with $R = 2$ ohms, $C = 1.0 \times 10^{-6}$ F, and $L = 5 \times 10^{-5}$ h. We can take V_0 to be any constant, real or complex. The method of undetermined coefficients (which works fine in this setting) can be used to find the periodic response of the circuit. This response will be of the form $q(t) = Ae^{i\omega t}$, the same form (and frequency ω) as the forcing function f . Use $q'(t) = i\omega Ae^{i\omega t}$ and $q''(t) = -\omega^2 Ae^{i\omega t}$ and substitute this information into the ODE, divide by $e^{i\omega t}$, and find

$$(-\omega^2 L + i\omega R + 1/C)A = V_0$$

so that

$$A = \frac{V_0}{-\omega^2 L + i\omega R + 1/C} \tag{C.13}$$

for the periodic response. This really is just the method of undetermined coefficients, but with complex exponentials instead of sines and cosines, and with a single complex-valued undetermined coefficient A .

If we take, for example, $V_0 = 1$ and $\omega = 10000$, then according to (C.13) the periodic response is

$$q(t) = Ae^{10000it} = A(\cos(10000t) + i\sin(10000t))$$

with

$$A = \frac{1}{(5 \times 10^{-5})(10000)^2 + 2 \times 10000i + 10^6} = \frac{1}{1005000 + 20000i} \approx 9.94 \times 10^{-7} - (1.98 \times 10^{-8})i.$$

If we want to know the system response with a forcing function of $\cos(10000t)$ (the real part of $V(t)$) we can just compute the real part of $Ae^{10000it}$, which turns out to be

$$q_r(t) \approx 1.0 \times 10^{-6} \cos(10000t) + 2.02 \times 10^{-8} \sin(10000t).$$

More generally, if we want the response with forcing $\sin(\omega t)$ we could take the imaginary part of $Ae^{i\omega t}$. ■

C.6 Impedance in Electrical Circuits

The notion of **impedance** generalizes that of resistance to circuits that contain capacitors, inductors, and possibly other circuit elements. It quantifies more generally the relation between voltage and current in circuits, especially when these quantities are periodic. Let's redo Example C.7, but in a slightly more general fashion.

Consider an RLC circuit governed by (C.11), with voltage source $V(t) = V_0 e^{i\omega t}$. We have

$$Lq''(t) + Rq'(t) + q(t)/C = V_0 e^{i\omega t}. \quad (\text{C.14})$$

Differentiate both sides of (C.14) with respect to t and use the fact that $q'(t) = I(t)$, so that $q''(t) = I'(t)$ and $q'''(t) = I''(t)$. We find

$$LI''(t) + RI'(t) + I(t)/C = i\omega V_0 e^{i\omega t}. \quad (\text{C.15})$$

The periodic response for the current, after transients have died out, will be of the form $I(t) = I_0 e^{i\omega t}$ for some constant I_0 , where I_0 is probably complex. Substitute this ansatz this into the ODE (C.15) (along with $I'(t) = i\omega e^{i\omega t}$ and $I''(t) = -\omega^2 e^{i\omega t}$), divide by $e^{i\omega t}$ and find

$$(-\omega^2 L + i\omega R + 1/C)I_0 = i\omega V_0.$$

Finally, divide both sides by $i\omega$ to obtain

$$\left(i\omega L + R + \frac{1}{i\omega C} \right) I_0 = V_0. \quad (\text{C.16})$$

This is a complex-valued version of Ohm's law appropriate to RLC circuits and periodic forcing. The parenthesized quantity on the left is called the **impedance** of the circuit. It's a generalization of the notion of resistance.

For example, consider a circuit with only a resistor (so the terms involving L and C in (C.16) are not present), with $\omega = 0$ and V_0 and I_0 as real numbers. Then (C.16) becomes the familiar Ohm's law $V = IR$ —no ODEs required to find the current in the circuit. But even if the capacitor and inductor are present in the circuit, (C.16) and the analysis that leads up to it makes it much easier to analyze the behavior of RLC circuits, also without every writing down any ODEs.

The quantity R in (C.16) is, of course, the resistance of the resistor. It is also the impedance of the resistor. The quantity L is the inductance of the inductor and $i\omega L$ is the impedance of the inductor at frequency ω radians per second, or just the impedance of the inductor. It's purely imaginary and its magnitude increases as ω increases; this reflects the fact that inductors oppose high frequencies, but let lower frequencies pass more easily. The quantity C is the capacitance of the capacitor, and $\frac{1}{i\omega C}$ is the impedance of the capacitor at frequency ω . This impedance decreases as ω increases and increases at lower frequencies. Capacitors oppose lower frequencies.

We often write Z to denote impedance, or Z_R , Z_L , and Z_C for the impedance of a resistor, inductor, and capacitor, respectively. The sum on the left in (C.16) is the impedance Z of the entire RLC series circuit—the impedances add, just like resistances in series, $Z = Z_L + Z_R + Z_C$. Ohm's law becomes $V = IZ$.

■ **Example C.8** Let's redo the circuit of Example C.7, an RLC circuit with $R = 2$ ohms, $C = 1.0 \times 10^{-6}$ F, and $L = 5 \times 10^{-5}$ h. We'll fix $\omega = 20000$ radians per second and take $V_0 = 5$. That is, we drive the circuit with $V(t) = 5e^{20000it}$. The periodic portion of the current is $I(t) = I_0 e^{20000it}$ where, from (C.16), $((20000)(5 \times 10^{-5})i + 2 + 1/((20000)(10^{-6})/i))I_0 = 5$ or, after simplifying,

$$(2 - 49i)I_0 = 5.$$

So the impedance of this RLC circuit at frequency $\omega = 20000$ is $Z = 2 - 49i$. Of course then $I_0 = 5/(2 - 49i) = 2/481 + 49i/481$. If we want the actual (real-valued) current $I(t)$ when $V(t) = 5 \cos(20000t)$ (the real part of $5e^{20000it}$) we just take the real part of $I_0 e^{20000it}$, which yields

$$I(t) = \frac{2}{481} \cos(20000t) - \frac{49}{481} \sin(20000t).$$

■

■ **Example C.9** Impedances are often more useful when expressed in polar form. For example, the impedance $Z = 2 - 49i$ of the circuit above at frequency $\omega = 20000$ can be written as

$$Z \approx 49.04e^{-1.53i}$$

since $|2 - 49i| \approx 49.04$ and $\arg(2 - 49i) \approx 1.53$. Express V_0 in polar form as $V_0 = 5e^{0i}$ and then from $V_0 = ZI_0$ we find

$$5e^{0i} = 49.04e^{-1.53i}|I_0|e^{i\phi}$$

where $|I_0|$ denotes the magnitude of I_0 and ϕ the argument or phase of I_0 . Then we find $5 = 49.04|I_0|$ and $0 = -1.53 + \phi$ so that

$$|I_0| = 5/49.04 \approx 0.102 \text{ and } \phi \approx 1.53.$$

That is, $I_0 \approx 0.102e^{1.53i}$, and so

$$I(t) = I_0 e^{i\omega t} = 0.102e^{i(\omega t + 1.53)}.$$

The conclusion: if we drive this circuit with a 5 volt signal at $\omega = 20000$ then the resulting current will have magnitude about 0.102 amps and will lead the voltage by 1.53 radians (that is, the current's graph is shifted 1.53 radians to the left, compared to the voltage). ■

Final Remarks

For the analysis of circuits driven with periodic sources, it's often more convenient to think of the driving voltage as complex-valued, in the form $V(t) = V_0 e^{i\omega t}$ for some constant V_0 . One can then determine the various periodic quantities of interest, e.g., currents, voltage drops, etc., after the transients have died out, without actually solving or even writing down ODEs. We instead use simple algebraic techniques involving the notion of impedance (a sort of generalized notion of resistance) that allows us to do an end run around the ODEs. This is not unlike the machinery of the Laplace transform, that allows us to reduce linear, constant coefficient ODEs to simpler algebra problems.

Bibliography

- [1] Project MKULTRA. 2012 (accessed 01 May 2021). http://en.wikipedia.org/wiki/Project_MKULTRA.
- [2] Risky decisions: How denial and delay brought disaster to New England's historic fishing grounds: A brief from The PEW Charitable Trusts, 2014 (accessed 01 May 2021). <https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2014/09/risky-decisions>.
- [3] <https://www.youtube.com/watch?v=o-urnlaJp0A>, (accessed 01 May 2021).
- [4] <https://www.youtube.com/watch?v=0ubvTOHWTms>, (accessed 01 May 2021).
- [5] <https://www.youtube.com/watch?v=kzVvd4Dk6sw>, (accessed 01 May 2021).
- [6] <https://www.youtube.com/watch?v=V8W4Djz6jnY>, (accessed 01 May 2021).
- [7] Parke systems, active vibration isolation table. <https://parksystems.com/park-afm-options/active-vibration-isolation-table>, (accessed 01 May 2021).
- [8] SIMIODE Textbook Web Site, (accessed 01 May 2021). <https://qubeshub.org/community/groups/simiode/textbook>.
- [9] Systemic initiative for modeling investigations and opportunities with differential equations (SIMIODE), (accessed 01 May 2021). <https://qubeshub.org/community/groups/simiode>.
- [10] Usain Bolt 100m 10 meter splits and speed endurance. <http://speedendurance.com/2008/08/22/usain-bolt-100m-10-meter-splits-and-speed-endurance>, (accessed 01 May 2021).
- [11] Wikipedia: hydrogen peroxide, (accessed 01 May 2021). http://en.wikipedia.org/wiki/Hydrogen_peroxide.

- [12] Christy refractories insulting fire brick datasheet. <http://christyrefractories.com/refractory-products/other-products/insulating-fire-brick/>, (accessed 10 March 2019).
- [13] Podcast: Determining time of death, air date 15 February 2015 (accessed 01 May 2021). <https://coronertalk.com/28>.
- [14] George K. Aghajanian and Oscar H. L. Bing. Persistence of lysergic acid diethylamide in the plasma of human subjects. *Clinical Pharmacology & Therapeutics*, 5(5):611–614.
- [15] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.
- [16] Christopher Arnold. *Designing For Earthquakes: A Manual for Architects, FEMA 454*. 2006 (accessed 01 May 2021). https://archexamacademy.com/download/Structural%20Systems/structures%20university/fema454_complete.pdf.
- [17] Stealth Fighter Association. About the f117. <https://www.f117sfa.org/about-the-f117>, (accessed 01 May 2021).
- [18] Kendall Atkinson. *An Introduction to Numerical Analysis*. John Wiley and Sons, New York, second edition, 1989.
- [19] Sheldon Axler. *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Springer, Princeton, third edition, 2015.
- [20] Grigory Barenblatt. *Scaling*, volume 34. Cambridge Texts in Applied Mathematics, New York, second edition, 2003.
- [21] Athanassios Bissas, Josh Walker, Catherine Tucker, and Gorgios Paradisis. Biomechanical report for the 100m women's IAAF world championships, London 2017. <http://centrostudiolombardia.com/wp-content/uploads/2018/10/1-100-donne.pdf>, (accessed 01 May 2021).
- [22] Åke Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, 1996.
- [23] Karen Bliss. 1-011A-S-Kinetics, 2015. <https://qubeshub.org/community/groups/simiode/publications?id=3310&v=1>.
- [24] M. Braun. *Differential Equations and Their Applications: An Introduction to Applied Mathematics*. Springer-Verlag, New York, fourth edition, 1993.
- [25] E.O. Brigham. *The Fast Fourier Transform and Its Applications*. Prentice-Hall, New Jersey, 1988.
- [26] S.A. Broughton and K. Bryan. *Discrete Fourier Analysis and Wavelets: Applications to Signal and Image Processing*. Wiley, New York, second edition, 2018.
- [27] K. Bryan and D. Walter. Geolocation of multiple noncooperative emitters using received signal strength: Sparsity, resolution, and detectability. *IEEE Access*, 8:121999–122012, 2020.
- [28] Kurt Bryan. Elementary inversion of the Laplace transform. 1998 (accessed 01 May 2021). <https://www.rose-hulman.edu/~bryan/invlap.pdf>.
- [29] Kurt Bryan. A tale of two masses. *PRIMUS*, 21(2):149–162, 2011.

- [30] Kurt Bryan. 5-010-S-MatrixExponential, 2015. <https://qubeshub.org/community/groups/simiode/publications?id=3348&v=1>.
- [31] Kurt Bryan. 3-095-S-ShotInWater, 2018. <https://qubeshub.org/community/groups/simiode/publications?id=3028&v=1>.
- [32] Kurt Bryan. 1-092-S-DashItAll, 2019. <https://qubeshub.org/community/groups/simiode/publications?id=3233&v=1>.
- [33] Kurt Bryan. 3-150-S-ItsABlastFurnace, 2019. <https://qubeshub.org/community/groups/simiode/publications?id=3042&v=1>.
- [34] J.C. Butcher. *Numerical Methods for Ordinary Differential Equations*. Wiley, New York, 2003.
- [35] Tor Carlson. Über geschwindigkeit und grösse der hefevermebrung in würze. *Biochem. Z*, 57:313–334, 1913.
- [36] Communicable Disease Surveillance Center. News and notes: Influenza in a boarding school. *British Medical Journal*, 1(6112):586–590, 1978 (accessed 01 May 2021). <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1603269/pdf/brmedj00115-0064.pdf>.
- [37] Colin W. Clark. *Mathematical Bioeconomics: The Optimal Management of Renewable Resources*. John Wiley and Sons, New York, second edition, 1990.
- [38] John Cloud. Was Timothy Leary Right? *Time Magazine*, 19 April 2007 (accessed 01 May 2021). <http://content.time.com/time/subscriber/article/0,33009,1612717,00.html>.
- [39] Lawrence Corwin and Robert Szczarba. *Calculus in Vector Spaces*. Marcel Dekker, New York, second edition, 1995.
- [40] Zina Deretsky. Hearing mechanics. 23 October 2006 (accessed 01 May 2021). https://commons.wikimedia.org/wiki/File:Hearing_mechanics.jpg.
- [41] Matthew Desmond. *Evicted: Poverty and Profit in the American City*. Broadway Books, New York, 2017.
- [42] Wandi Ding. 1-070-S-FisheryHarvest, 2015. https://qubeshub.org/community/groups/simiode/publications?id=3261&tab_active=about&v=1.
- [43] Wandi Ding and Suzanne Lenhart. Optimal harvesting of a spatially explicit fishery model. *Natural Resource Modeling*, 22(2):173–211, 2009.
- [44] A.N. Dmitriev, Y.A. Chesnokov, K. Chen, Ivanov O.Y., and M.O. Zolotykh. Monitoring the wear of the refractory lining in the blast furnace hearth. *Steel in Translation*, 43:732739, 2013.
- [45] Armand du Plessis. Taipei 101 tuned mass damper. 02 June 2010 (accessed 01 May 2021). https://commons.wikimedia.org/wiki/File:Taipei_101_Tuned_Mass_Damper_2010.jpg.
- [46] Phil Dyke. *An Introduction to Laplace Transforms and Fourier Series*. Springer London, London, second edition, 2014.

- [47] Practical Engineering. What is a tuned mass damper? <https://www.youtube.com/watch?v=f1U4SAgy60c>, (accessed 01 May 2021).
- [48] R.L. Finney and D.E. Ostberg. *Elementary Differential Equations With Linear Algebra*. Addison-Wesley, Reading MA, 1968.
- [49] Georgy F. Gause. Experimental studies on the struggle for existence. *Journal of Experimental Biology*, 9(4):389–402, 1932.
- [50] Georgy F. Gause. *The Struggle for Existence*. Dover Publications, 1971 (first published in 1934 by The Williams & Wilkins Company). Available at <http://www.ggause.com/Contgau.htm>.
- [51] Georgy F. Gause, O.K. Nastukova, and W.W. Alpatov. The influence of biologically conditioned media on the growth of a mixed population of paramecium cadatum and p. aureliax. *Journal of Animal Ecology*, 3(2):222–230, 1934.
- [52] Paul Goldberger. Architecture View; A Novel Design And Its Rescue From Near Disaster. *New York Times*, pages 45–53, 24 April 1988 (accessed 01 May 2021). <https://www.nytimes.com/1988/04/24/arts/architecture-view-a-novel-design-and-its-rescue-from-near-disaster.html>.
- [53] Kevin Gray. Starting salary projections positive for the class of 2021. January 21, 2021 (accessed 01 May 2021). <https://www.naceweb.org/job-market/compensation/starting-salary-projections-positive-for-the-class-of-2021>.
- [54] R. Grech, T. Cassar, J. Muscat, and et al. Review on solving the inverse problem in eeg source analysis. *NeuroEngineering Rehabil*, 5:349–372, 2008.
- [55] David J. Griffiths. *Introduction to Electrodynamics*. Cambridge University Press, New York, fourth edition, 2017.
- [56] Jack K. Hale. *Ordinary Differential Equations*. Dover, Mineola NY, 2009.
- [57] Philip Hartman. *Ordinary Differential Equations*. SIAM Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, second edition, 2002.
- [58] M. Hasanbulli, S.P. Rogovchenko, and Y.V. Rogovchenko. Dynamics of a single species in a fluctuating environment under periodic yield harvesting. *Journal of Applied Mathematics*, 2013 (accessed 01 May 2021). <https://doi.org/10.1155/2013/167671>.
- [59] Alan Hastings and Thomas Powell. Chaos in a three species food chain. *Ecology*, 72:896–903, 1991.
- [60] Kenneth L. Henold and F. Walmsley. *Chemicals: Principles, Properties, and Reactions*. Addison-Wesley, Reading MA, 1984.
- [61] Gudmund Hernes. The process of entry into first marriage. *American Sociological Review*, 37(2):173–182, 1972.
- [62] João P. Hespanha. *Linear Systems Theory*. Princeton University Press, Princeton, second edition, 2009.
- [63] Ray Hilborn. *Overfishing: What Everyone Needs to Know*. Oxford University Press, Oxford, England, 2012.

- [64] Ray Hilborn and Carl J. Walters. *Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty*. Chapman and Hall, Inc., London, 1992.
- [65] A.V. Hill. The physiological basis of athletic records. *The Scientific Monthly*, 21(4):409–428, 1925.
- [66] Paul Horowitz and Winfield Hill. *The Art of Electronics*. Cambridge University Press, Cambridge, England, third edition, 2015.
- [67] Motioneering Inc. Grand Canyon Skywalk and Custom TMDs. *At the Moment: Motion Control News and Views from Motioneering*, 5:1–2, 2006.
- [68] I. Jain, R.J. Singh, and D. Mazumdar. Measurements of some thermal properties of steel-refractory systems and heat losses from steelmaking furnaces. *Transactions of the Indian Institute of Metals*.
- [69] Hem Raj Joshi, Guillermo E. Herrera, Suzanne Lenhart, and Michael G. Neubert. Optimal dynamic harvest of a mobile renewable resource. *Natural Resource Modeling*, 22(2):322–343, 2009.
- [70] Erdi Kara. 1-126-S-MarriageMath-StudentVersion, 2020. <https://qubeshub.org/community/groups/simiode/publications?id=3013&v=1>.
- [71] J.B. Keller. A theory of competitive running. *Physics Today*, 26(9):43, 1973.
- [72] Steve Kemper. *Code Name Ginger: The Story Behind Segway and Dean Kamen's Quest to Invent a New World*. Harvard Business School Press, Boston MA, 2003.
- [73] Rose M. Kreider and Renee Ellis. Number, timing, and duration of marriages and divorces: 2009. *Current Population Reports, U.S. Census Bureau*, 2009 (accessed 01 May 2021). <https://www.census.gov/prod/2011pubs/p70-125.pdf>.
- [74] Tribikram Kundu. Acoustic source localization. *Ultrasonics*, 54(1):25–38, 2014.
- [75] Mark Kurlansky. *Cod: A Biography of the Fish that Changed the World*. Penguin Books, London, 1998.
- [76] J.D. Lambert. *Numerical Methods for Ordinary Differential Systems: The Initial Value Problem*. Wiley, 1991.
- [77] Keith Alan Landry and Brian Winkel. 5-040-S-TunedMassDampers-Part I, 2016. <https://qubeshub.org/community/groups/simiode/publications?id=2995&v=1>.
- [78] Keith Alan Landry and Brian Winkel. 5-040-S-TunedMassDampers-Part II, 2016. <https://qubeshub.org/community/groups/simiode/publications?id=3116&v=1>.
- [79] Glenn Ledder. *Mathematics for the Life Sciences. Calculus, Modeling, Probability and Dynamical Systems*. Springer, New York, 2010.
- [80] Oscar A. Linares and Annemarie L. Linares. Computational opioid prescribing: A novel application of clinical pharmacokinetics. *Journal of Pain and Palliative Care Pharmacotherapy*, 25:125–135, 2011.
- [81] Robert Loschke. Development of the f-117 flight control system. In *AIAA Guidance, Navigation, and Control Conference and Exhibit (Austin, Texas)*, pages 161–174, August 11-14, 2003 (accessed 01 May 2021). <https://doi.org/10.2514/6.2003-5762>.

- [82] Xu Luo and Jun Yang. A survey on pollution monitoring using sensor networks in environment protection. *Journal of Sensors*, 2019.
- [83] Rich Marchand and Timothy J. McDevitt. Learning differential equations by exploring earthquake induced structural vibrations: A case study. *Int. J. Engng Ed.*, 15(6):477–485, 1999.
- [84] Jörg Matthes, Lutz Gröll, and Hubert B. Keller. Source localization based on pointwise concentration measurements. *Sensors and Actuators: A. Physical*, 115:32–37, 2004.
- [85] Desmond Matthew, Ashley Gromis, Lavar Edmonds, James Hendrickson, Katie Krywokulski, Lillian Leung, and Adam Porton. Eviction lab national database: Version 1.0, (accessed 01 May 2021). <https://www.evictionlab.org>.
- [86] Mehdi Mazaheri, Jamal Mohammad Vali Samani, and Hossein Mohammad Vali Samani. Mathematical model for pollution source identification in rivers. *Environmental Forensics*, 16(4):310–321, 2015.
- [87] Matt McFarland. The Segway is Officially Over. *CNN Business*, 25 June 2020 (accessed 01 May 2021). <https://www.cnn.com/2020/06/23/tech/segway-pt-shut-down/index.html>.
- [88] Carl M. Metzler. A mathematical model for the pharmacokinetics of LSD effect. *Clin Pharmacol Ther.*, 10(5):737–740, 1969.
- [89] Sheila Miller. 6-001-S-Epidemic, 2015. <https://qubeshub.org/community/groups/simiode/publications?id=2956&v=1>.
- [90] Lai Ming-Lai. Tuned mass damper. www.google.com/patents/US5558191, 1996 (accessed 01 May 2021).
- [91] Joe Morgenstern. The Fifty-Nine Story Crisis. *The New Yorker*, pages 45–53, 29 May 1995.
- [92] James D. Murray. *Mathematical Biology*. Springer-Verlag, New York, second, corrected edition, 1993.
- [93] National Information Service for Earthquake Engineering (NISEE). The earthquake engineering online archive. <https://nisee.berkeley.edu>, (accessed 01 May 2021).
- [94] National Law Center on Homelessness and Poverty. Protect tenants, prevent homelessness, (accessed 01 May 2021). <https://www.nlchp.org>.
- [95] J. Norman. One-compartment kinetics. *British Journal of Anaesthesia*, 69:387–396, 1992.
- [96] "Royal Observatory of Belgium". <https://wwwbis.sidc.be/silso/datafiles>, (accessed 04 November 2021).
- [97] Mark Peastrel, Rosemary Lynch, and Jr. Angelo Armenti. Terminal velocity of a shuttlecock in vertical fall. *American Journal of Physics*, 48(7):511–513, 1980.
- [98] William G. Pritchard. Mathematical models of running. *SIAM Review*, 35(3):359–379, 1993.
- [99] E.J. Putzer. Avoiding the Jordan canonical form in the discussion of linear systems with constant coefficients. *The American Mathematical Monthly*, 73(1):2–7, 1966.

- [100] David A. Sánchez. *Ordinary Differential Equations and Stability Theory*. Dover Publications, Mineola, New York, 2019.
- [101] George F. Simmons. *Differential Equations with Applications and Historical Notes*. Chapman and Hall/CRC, New York, third edition, 2017.
- [102] K. Sorli and I. Skaar. Monitoring the wear-line of a melting furnace. In *Inverse Problems in Engineering: Theory and Practice, 3rd Int. Conference on Inverse Problems in Engineering (Port Ludlow WA)*, June 13-18, 1999.
- [103] Frank Parker Stockbridge. How far off is that german gun? how 63 german guns were located by sound waves alone in a single day. *Popular Science monthly (scanned by Google Books)*, page 39, December 1918.
- [104] John M. Stockie. The mathematics of atmospheric dispersion modeling. *SIAM Review*, 53:349–372, 2011.
- [105] Walter A. Strauss. *Partial Differential Equations: An Introduction*. Wiley, New York, 1992.
- [106] Steven Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Westview Press, Boulder CO, 1994.
- [107] Jennifer Switkes. A modified discrete SIR model. *The College Mathematics Journal*, 34(5):399–402, 2003.
- [108] V. Tandon, W.S. Kang, T.A. Robbins, A.J. Spencer, E.S. Kim, M.J. McKenna, S.G. Kujaawa, J. Fiering, E.E. Pararas, M.J. Mescher, and W.F. Sewell. Microfabricated reciprocating micropump for intracochlear drug delivery with integrated drugfluid storage and electronically controlled dosing. *Lab. Chip.*, 16(5):829–846, 2018. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4766044/>.
- [109] American Red Cross Multi-Disciplinary Team. *Report on the 2010 Chilean earthquake and tsunami response: U.S. Geological Survey Open-File Report 2011-1053*, v. 1.1. 2011 (accessed 01 May 2021). <https://pubs.usgs.gov/of/2011/1053/>.
- [110] Miguel Alvarez Texocotitla, M. David Alvarez-Hernández, and Shaní Eneida Alvarez-Hernández. Dimensional analysis in economics: A study of the neoclassical economic growth model. *Journal of Interdisciplinary Economics*, 32(2):123144, Jul 2019.
- [111] TMC Vibration Control. Optical table advances quiet vibrations in highly sensitive applications. <https://www.techmfg.com/learning/whitepapers/optical-table-advances>, (accessed 01 May 2021).
- [112] Mary Vanderschoot. 5-026-S-Evictions, 2018. <https://qubeshub.org/community/groups/simiode/publications?id=2986&v=1>.
- [113] J.G. Wagner, G.K. Aghajanian, and O.H.L Bing. Correlation of performance test scores with “tissue concentration” of lysergic acid diethylamide in human subjects. *Clinical Pharmacology and Therapeutics*, 9(5):635–638, 1968.
- [114] Paul Waltman. *A Second Course in Elementary Differential Equations*. Dover Publications, New York, 2004.
- [115] Jue Wang. 1-138-S-InnerEarDrugDelivery, 2018. <https://qubeshub.org/community/groups/simiode/publications?id=3201&v=1>.

- [116] Tracy Weyand. 1-136-S-MarriageAge, 2020. <https://qubeshub.org/community/groups/simiode/publications?id=3203&v=1>.
- [117] Richard E. Williamson. *Introduction to Differential Equations and Dynamical Systems*. McGraw-Hill, New York, second edition, 2001.
- [118] Brian Winkel. Ants, tunnels, and calculus: An exercise in mathematical modeling. *Mathematics Teacher*, 87(4):284–287, 1994.
- [119] Brian Winkel. 1-006-S-FinancingSavingsAndLoans, 2015. <https://qubeshub.org/community/groups/simiode/publications?id=2974&v=1>.
- [120] Brian Winkel. 1-007-S-AntTunnelBuilding, 2015. <https://qubeshub.org/community/groups/simiode/publications?id=2993&v=1>.
- [121] Brian Winkel. 1-012-S-SublimationCarbonDioxide, 2015. <https://qubeshub.org/community/groups/simiode/publications?id=2950&v=1>.
- [122] Brian Winkel. 3-008-S-HangTime, 2015. <https://qubeshub.org/community/groups/simiode/publications?id=3052&v=1>.
- [123] Brian Winkel. 3-019-S-ShuttlecockFall, 2015. <https://qubeshub.org/community/groups/simiode/publications?id=3176&v=1>.
- [124] Brian Winkel. 5-001-S-LSDAndProblemSolving, 2015. <https://qubeshub.org/community/groups/simiode/publications?id=3333&v=1>.
- [125] Brian Winkel. 7-008-S-MachineReplacement, 2015. <https://qubeshub.org/community/groups/simiode/publications?id=2999&v=1>.
- [126] Brian Winkel. 1-061-S-PotatoCooling, 2016. <https://qubeshub.org/community/groups/simiode/publications?id=3269&v=1>.
- [127] Brian Winkel. 6-040-S-StruggleForExistence, 2016. <https://qubeshub.org/community/groups/simiode/publications?id=2985&v=1>.
- [128] Brian Winkel. 3-002-S-ModelsMotivatingSecondOrder, 2017. <https://qubeshub.org/community/groups/simiode/publications?id=3186&v=1>.
- [129] World Health Organization. Deafness and hearing loss. <http://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>, 2018 (accessed 01 May 2021).
- [130] A.A. Yakubu, N. Li, J.M. Conrad, and M-L. Zeeman. Constant proportion harvest policies: dynamic implications in the pacific halibut and atlantic cod fisheries. *Mathematical Biosciences*, 232:66–77, 2011.
- [131] N. Yarlikina and H. Walrath. Determining the length of a one-dimensional bar from thermal measurements, 2004 (accessed 10 October 2021). <https://www.rose-hulman.edu/~bryan/reu2/natholpaper.pdf>.
- [132] Armen H. Zemanian. *Distribution Theory and Transform Analysis: An Introduction to Generalized Functions, with Applications*. Dover Publications, New York, 2010.
- [133] Steven S. Zumdahl. *Chemical Principles*. D. C. Heath and Company, Lexington MA, 1992.

Index

- L^1 -minimization, 112
- L^2 convergence, 470
- s -domain, 234
- t -domain, 234

- accelerometer, 306
- accuracy
 - first-order, 82
 - second-order, 90
- accuracy
 - fourth-order, 93
- adaptive step size, 93, 94, 98, 407
- adaptive step size method, 93
- advection equation, 502
- AIC figure of merit, 127
- Akaike information criterion, 106, 127
- ambient temperature, 32
- amplitude, 156
- ansatz, 141
- Archimedes' principle, 137
- asymptotically stable equilibrium solution, 54, 387
 - asymptotically stable fixed point, 54, 387
 - asymptotically stable improper node, 373
 - asymptotically stable node, 372
 - asymptotically stable spiral point, 374
- Atlantic cod, 8
- autonomous, 51
- auxiliary equation, 140

- backward Euler's method, 415
- backward heat equation, 517
- basin of attraction, 433
- basis, 143
 - basis functions, 143
- beat, 184
- bicycle shock absorber, 132, 167
- bifurcation, 58
 - pitchfork, 61
 - transcritical, 59
- bifurcation diagram, 59
- blow up, 64
- Bolt, Usain, 1, 100
- bolus, 217
- boundary condition
 - Robin, 463
- boundary conditions
 - Dirichlet, 457
 - insulating, 458
 - Neumann, 457
- bounded function, 466
- Bromwich integral, 233
- butadiene, 73

- capacitor, 571
- carrying capacity, 10
- causality, 513
- center, 374
- characteristic equation, 140, 561
- characteristic polynomial, 561

- characteristic variable scales, 191, 193
 characteristics, 503
 circuit
 RC, 34, 573
 RL, 575
 RLC, 135, 576
 closed-loop control, 288, 290
 closed-loop transfer function, 290
 cochlea, 5
 cohort, 120
 commutative diagram, 228
 compartmental model, 6, 33, 311
 LSD, 353
 competing species model, 362
 complete, 470, 474
 complex conjugate, 157
 complex number
 arithmetic, 542
 conjugate, 542, 546
 definition, 542
 exponentiation, 544
 imaginary part, 542
 modulus, 542
 real part, 542
 complex plane, 542
 compliance, 135
 conjugate, 145, 542, 546
 conjugate roots, 145
 conservation law, 8, 451
 conservative, 427
 consistent linear system, 553
 constant-coefficient, 27, 131
 constant-coefficient linear system, 314, 320
 constitutive relation, 456
 continuity equation, 453
 control
 closed-loop, 288, 290
 feedback, 288, 290
 open-loop, 287
 PI, 292
 PID, 294
 proportional, 288
 proportional-integral, 292
 proportional-integral-derivative, 294
 tuning, 293, 297
 control function, 284
 control theory, 222
 control volume, 452
 conventional current, 569
 convergence
 L^2 , 470
 mean square, 470
 pointwise, 475
 uniform, 476
 convolution, 272, 274
 convolution theorem, 275
 Coulomb damping, 136
 critical point, 52
 critically damped, 145, 150, 153
 current, 569
 D'Alembert solution, 512
 damped harmonic oscillator, 132
 damped pendulum, 215, 307
 damped wave equation, 529
 damper, 130
 damping
 Coulomb, 136
 viscous, 131
 dashpot, 130
 DCT, 534
 deconvolution, 280
 defective eigenvalue, 562
 defective matrix, 327
 deformation
 elastic, 130
 plastic, 130
 derivative gain, 294
 determinant, 557
 diagonal, 554
 diagonalizable matrix, 565
 diagonalization, 347, 564
 diagonalize, 347
 difference equation, 68
 differential equation
 linear, 13
 nonlinear, 13
 ordinary, 4
 partial, 453
 differential operator, 504
 diffusion, 490
 diffusion equation, 491
 source term, 494
 dimension, 16
 dimensional analysis, 16
 dimensionless, 17
 dimensionless variables, 193
 dinitrogen pentoxide, 74
 Dirac delta function, 258–260

- Dirac mass, 262
direct problem, 494, 521
direction field, 50, 367
Dirichlet boundary conditions, 457
discrete cosine transform, 534
disturbances, 288, 295
driven harmonic oscillator, 131, 159

eigenvalue, 320, 559
 defective, 562
eigenvector, 320, 559
elastic deformation, 130
electric charge, 570
envelope, 185
equation
 advection, 502
 continuity, 453
 transport, 502
 wave, 505
equilibrium position, 130
equilibrium solution, 51, 363
 asymptotically stable, 54, 387
 semi-stable, 54
 stable, 54, 387
 unstable, 54, 387
error control, 98
error, Euler's method, 83
Euler's formula, 140, 145, 544
Euler's method, 80, 405
 improved, 88, 406
Euler's method error theorem, 83
exchange of stability, 59
existence-uniqueness theorem, 65
existence-uniqueness theorem for first-order systems, 316
explicit ODE method, 415
exponential growth, 9
exponential order, 225

fast Fourier transform, 533
feedback control, 288, 290
FFT, 533
Fick's law, 491
filter
 low-pass, 37
final value theorem, 237
finite difference, 118
first integral, 427
first shifting theorem, 232
first-order accurate, 82

first-order ODE, 13
first-order ordinary differential equation
 standard form, 13, 313
first-order system, 313, 361
 linear, 314
fish, 8
fixed point, 52, 363
 asymptotically stable, 54, 387
 semi-stable, 54
 stable, 54, 387
 unstable, 54, 387
forced harmonic oscillator, 131, 159
forcing function, 163
forward problem, 494, 521
Fourier cosine expansion, 470, 471
Fourier cosine expansion, two-dimensional, 480
Fourier cosine series, 470
Fourier series
 complex exponential, 482
 cosine, 470, 471
 cosine, two-dimensional, 480
 sine, 473
 sine-cosine, 480
Fourier sine expansion, 473
Fourier sine series, 474
Fourier sine-cosine expansion, 480
Fourier's law, 455
fourth-order accurate, 93
frequency domain, 234
fundamental frequency, 510
fundamental matrix solution, 344
fundamental set of solutions, 143
fundamental theorem of algebra, 546
furnace, 449

gain, 181, 289
 derivative, 294
 integral, 292, 294
 proportional, 289, 292, 294
gain function, 181
Galilean transformation, 536
Gause, G.F., 362
Gedankenexperiment, 455
general solution, 4, 14, 141, 143, 153, 322
 nonhomogeneous, 162
 real-valued, 148, 150, 325
global error, 98
globally asymptotically stable, 433
globally attractive, 433
ground, 569

- growth rate
 intrinsic, 9

 half-life, 36, 217
 hang time, 22
 hard spring, 189
 harmonic, 510
 harmonic oscillator
 critically damped, 145, 150, 153
 damped, 132
 driven, 131, 159
 forced, 131, 159
 overdamped, 139, 144, 153
 pure, 131
 undamped, 131, 145, 153
 underdamped, 145, 153
 unforced, 131

 Hartman-Grobman theorem, 389, 390
 harvested logistic equation, 195
 harvesting, 11, 195
 heat equation
 source term, 494
 Heaviside function, 239, 243
 henry, 135
 Heun's method, 89
 Hill-Keller model, 2, 100
 homelessness, 354, 444
 homogeneous, 27, 131, 320
 homogeneous linear system, 320
 Hooke's law, 130, 189
 hydrogen peroxide, 72
 hyperbolic, 390
 hyperbolic equilibrium point, 390

 identity matrix, 555
 image compression, 472
 imaginary part, 542
 impedance, 580
 implicit Euler's method, 415
 implicit method, 415
 improved Euler's method, 88
 impulse, 256
 impulse response, 272, 276
 impulsive, 219
 inconsistent linear system, 553
 incubator, 283
 inductance, 135, 574
 inductor, 135, 574
 infusion pump, 217
 initial condition, 4

 initial conditions, 132
 initial value problem, 4
 initial value theorem, 236
 input-output system, 270
 insulating boundary conditions, 458
 integral equation, 303
 integral gain, 292, 294
 integrating factor, 28, 30
 integrodifferential equation, 309
 intermediate value theorem, 61
 intrinsic growth rate, 9
 inverse Laplace transform, 228, 233
 inverse matrix, 555, 556
 inverse problem, 494, 522
 inverted pendulum, 307
 invertible matrix, 556
 investing, 23

 Jacobian matrix, 390
 JPEG, 472
 jump discontinuity, 225

 Kamen, Dean, 306
 Kirchhoff's current law, 570
 Kirchhoff's voltage law, 135, 570
 KISS philosophy, 117

 Laplace *s*-domain, 234
 Laplace transform, 220, 332
 inverse, 228, 233
 linearity, 225
 LaSalle's invariance principle, 434
 law of mass action, 70
 least-squares, 102
 least-squares estimation, 102, 104
 Leibniz notation, 41
 Liapunov function, 431
 linear, 13
 linear first-order system, 314
 linear independence, 322
 linear ordinary differential equation, 27
 linear system
 consistent, 553
 constant-coefficient, 314, 320
 inconsistent, 553
 linear system of equations, 551
 linearization, 78, 387
 linearly independent, 144, 347
 linearly independent vectors, 565
 loans, 67

- local truncation error, 94
- logistic equation, 10, 193
 - harvested, 195
- logistic equation with harvesting, 11
- Lorentz factor, 538
- Lorentz transformation, 538
- Lotka-Volterra competing species model, 362
- Lotka-Volterra predator-prey equations, 440, 442, 445
- low-pass filter, 37
- lower triangular matrix, 562
- LSD, 311, 376
- LTE, 94
- Lyapunov function, 431
- Lyapunov's direct method, 431
- Lyapunov's first method, 391
- Lyapunov's second method, 431
- lysergic acid diethylamide, 311
- marching, 81
- matrix, 552
 - defective, 327
 - diagonalizable, 565
 - invertible, 556
 - lower triangular, 562
 - square, 553
 - upper triangular, 562
- matrix exponential, 343
- matrix inverse, 556
- matrix multiplication, 554
- maximum domain, 63
- mean square convergence, 470
- mean value theorem, 61
- method of undetermined coefficients, 163
- modified Euler's method, 89
- modulus, 542
- money
 - investing, 23
- morphine, 217
- morphine sulfate, 217
- mortgage, 67
- multiplication, matrix-matrix, 554
- multiplication, matrix-vector, 552
- multiplicity, 235, 546
- natural frequency, 150
- negative definite function, 429
- negative definite quadratic form, 440
- negative semidefinite function, 429
- Neumann boundary conditions, 457
- Newton's law of cooling, 31, 32
- Newton's second law, 2, 23
- Newton's universal law of gravitation, 138
- nitrous oxide, 71
- node
 - asymptotically stable, 372
 - asymptotically stable improper, 373
 - unstable, 372
 - unstable improper, 373
- nondimensionalization, 190, 193
- nondimensionalize, 193
- nonhomogeneous, 27, 131, 320
- nonhomogeneous general solution, 162
- nonhomogeneous linear system, 320, 332
- nonlinear, 13
- nonlinear ordinary differential equation, 27
- nonlinear pendulum, 364
- nullcline, 377
- Nyquist frequency, 534
- ODE, 4
 - order, 13
- ODEs
 - stiff, 408
- Ohm's law, 570
- one-step ODE method, 415
- one-way wave equation, 502
- open-loop control, 287
- optimization, 108
- order, second, 13
- order, first, 13
- ordinary differential equation, 4
 - autonomous, 51
 - constant-coefficient, 27
 - homogeneous, 27
 - linear, 27
 - nonhomogeneous, 27
 - nonlinear, 27
 - separable, 38
 - time-invariant, 51
 - variable-coefficient, 27
- orthogonal functions, 469
- overdamped, 139, 144, 153
- parameter estimation, 11, 494, 522
- parsec, 191
- partial differential equation, 453
- partial differential equations, 31
- partial fraction decomposition, 548
- pendulum, 211, 213

- damped, 215, 307
 inverted, 307
 nonlinear, 364
 periodic, 161, 166
 pharmacokinetics, 217, 311
 phase line portrait, 51
 phase plane, 365
 phase portrait, 51, 379, 385
 phase shift, 156
 phase space, 365
 PI control, 292
 PID control, 294
 piecewise continuous, 225
 pitchfork bifurcation, 61
 plant, 286
 plastic deformation, 130
 pointwise convergence, 475
 poles, 236, 547
 polynomial, 546
 polynomial roots, 541
 positive definite function, 429
 positive semidefinite function, 429
 Post inversion formula, 233
 predator-prey, 445
 principal, 68
 principle of causality, 513
 principle of superposition, 141
 process variable, 285
 proportional control, 288
 proportional gain, 289, 292, 294
 proportional-integral control, 292
 proportional-integral-derivative control, 294
 pure harmonic oscillator, 131
 pure resonance, 182
 Q-factor, 187
 quadratic form
 negative definite, 440
 quadratic formula, 541
 radian, 18
 rational function, 234, 547
 poles, 547
 zeros, 547
 RC circuit, 34, 573
 RC time constant, 37
 reaction
 first-order, 71
 second-order, 73
 zeroth-order, 71
 reaction order, 70
 reaction rates, 70
 real part, 542
 real-valued general solution, 148, 150, 325
 reduction of order, 151
 reference signal, 285
 refractory material, 449
 rescaling, 193, 194
 residual, 103, 106
 residual sum of squares, 103, 106
 resonance, 174, 181
 pure, 182
 resonant frequency, 182
 RK4, 92, 406
 RL circuit, 575
 RLC circuit, 135, 576
 Robin boundary conditions, 463
 root
 multiplicity, 235, 546
 roots
 polynomial, 541, 546
 Routh-Hurwitz Theorem, 401, 435
 Runge-Kutta fourth-order method, 92, 406
 Runge-Kutta method, 89, 92
 saddle point, 372
 salt tank, 332
 salt tank model, 16, 33
 sampling, 530
 sampling rate, 531
 scaling, 190, 193
 Schwarzschild radius, 20
 second shifting theorem, 245
 second-order accurate, 90
 second-order ODE, 13
 second-order reactions, 73
 Segway scooter, 306
 semi-stable, 55
 semi-stable equilibrium solution, 54
 semi-stable fixed point, 54
 separable, 458
 separable ODE, 38
 separation of variables, 40, 41
 setpoint, 285
 sgn function, 136
 shifting theorem
 first, 232
 second, 245
 shock absorber, 132, 167
 sifting property, 260

- sink, 55, 372
SIR model, 364, 441
slope field, 50
solar activity cycle, 535
solution
 general, 4, 141, 322
source, 55, 372
source localization, 491, 523
source term, 494
spacetime, 536
special relativity, 535
specific heat, 454
spiral sink, 374
spiral source, 374
spring constant, 130
square matrix, 553
stable equilibrium solution, 54, 387
stable fixed point, 54, 387
stable star point, 372
standard form, 13, 313
star point
 stable, 372
 unstable, 372
steady-state, 166
step response, 280
step size, 80
stiff, 410, 411, 413, 415
stiff ODEs, 408
strict Lyapunov function, 431
stuff, 451
 conservation, 451
sublimate, 116
sum of squares, 102, 107
 residual, 103, 106
superposition, 141
susceptible-infected-recovered model, 364
system
 stiff, 411, 413
system identification, 270, 278
tangent line approximation, 78
temperature, 32, 454
theorem
 convolution, 275
 Euler's method error, 83
 existence-uniqueness, 65
 existence-uniqueness for first-order sys-
 tems, 316
 final value, 237
 Hartman-Grobman, 390
initial value, 236
Routh-Hurwitz, 401, 435
thermal conductivity, 455
thought experiment, 455
three species food chain, 435
time domain, 234
time-invariant ODE, 51
TMD, 356
transcritical bifurcation, 59
transfer function, 271
 closed-loop, 290
transient, 161, 166
transport equation, 502
tuned mass damper, 356
tuning, 293, 297
two-compartment model, 312
undamped, 145, 153
undamped harmonic oscillator, 131
underdamped, 145, 147, 153
undetermined coefficients, 163, 334
 failed guess, 169
unforced harmonic oscillator, 131
uniform convergence, 476
unit impulse response, 272, 276
unit step function, 239, 243
unit-free, 19
unstable equilibrium solution, 54, 387
unstable fixed point, 54, 387
unstable improper node, 373
unstable node, 372
unstable spiral point, 374
unstable star point, 372
upper triangular matrix, 562
variable-coefficient, 27
vector field, 50
vibration isolation, 133
vibration isolation table, 133
viscosity, 46
viscous damping, 131
vodka, 362
voltage, 569
wave equation, 505
 damped, 529
wave speed, 505
zeros, 547



Catch the spirit of modeling first and throughout while learning important mathematics in context. *Differential Equations: A Toolbox for Modeling the World* puts applications and modeling front and center in an introduction to ordinary differential equations. This approach does not skimp on or skip over the mathematics, but uses applications to motivate both subject and technique. Differential equations are interwoven with modeling to drive forward both the mathematics and the reader's understanding of the application under study. This approach makes it clear that differential equations provide a powerful and indispensable toolbox for describing the world.

The book includes some important topics not usually offered in introductory texts: dimensional analysis, scaling and nondimensionalization for differential equations, parameter estimation, and a brief introduction to control theory via Laplace transforms. There is also more material on modern numerical methods than is typical in an introductory text. The incorporation of these topics is structured so that they may be taken advantage of, or omitted, as time and student interest permits, without disrupting the flow of later topics.

The text includes numerous activities for students, including:

- Over 200 inline Reading Exercises woven into the text itself, to immediately engage and reinforce the reader's mastery of the material.
- Over 230 traditional section-end exercises, ranging from routine computation and solution techniques to more involved modeling and theory.
- Three to six Modeling Projects at the end of each chapter, twenty-six in all. Many are based on those published by SIMIODE, many are entirely new.

Dr. Glenn Ledder, University of Nebraska, Lincoln NE USA, says in his review in *The UMAP Journal*, “This book is the only one this reviewer is aware of that presents differential equations in a modeling context rather than merely adding a bit of modeling to the standard presentation. If you want to study the mathematics of differential equations in a modeling context, you are in the right place.”

Kurt Bryan is Professor of Mathematics at the Rose-Hulman Institute of Technology. He has worked in industry, government, and academia as an educator, researcher, and consultant. He is the author of a textbook on signal and image processing and approximately 30 journal articles. His research interests include partial differential equations, inverse problems, and signal and image processing.

This work was supported in part by the National Science Foundation through NSF:DUE-IUSE Grant # 1940532.

