

Differential Equations: A Toolbox for Modeling the World

Draft Version 0.62

Kurt Bryan

Copyright © 2020 Kurt Bryan

PUBLISHED BY SIMIODE

SIMIODE.ORG

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the “License”). You may not use this file except in compliance with the License. You may obtain a copy of the License at <http://creativecommons.org/licenses/by-nc/3.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

First printing, March 2020

Preface

This is a Work in Progress Preliminary Draft of the SIMIODE Online Text which we hope to offer for use in Spring Semester 2021 for in class use. Collegial feedback is sought to improve the work for first real release for commercial use in Fall 2021 semester.

Contents

1	Why Differential Equations?	11
1.1	The 2008 Olympic 100 Meter Dash	11
1.1.1	Usain Bolt's Olympic Victory	11
1.1.2	Modeling a Sprint	12
1.1.3	The Hill-Keller Differential Equation	13
1.2	Intracochlear Drug Delivery	15
1.2.1	The Challenge of Hearing Loss	15
1.2.2	A Compartmental Model for the Cochlea	16
1.2.3	The Differential Equation	17
1.3	Population Growth and Fishery Management	18
1.3.1	The Need to Manage Fish Harvesting	18
1.3.2	Modeling Fish Population	18
1.3.3	Modeling Harvesting	20
1.3.4	Parameter Estimation and Harvesting	21
1.4	Where Do We Go from Here?	22
1.4.1	A Toolbox for Describing the World	22
1.4.2	Some Terminology	22
1.4.3	You Already Know How to Solve Some Differential Equations	23
1.4.4	Exercises	25
1.5	The Blessing of Dimensionality	26
1.5.1	Definition of Dimension	26
1.5.2	The Algebra of Dimension	27
1.5.3	Derivatives, Integrals, Elementary Functions	28
1.5.4	Unit-Free Equations and Bending the Rules	29
1.5.5	Using Dimension to Find Plausible Models	29
1.5.6	Other Dimensions	30
1.5.7	Exercises	30

1.6 Modeling Projects	31
1.6.1 Project: Hang Time	31
1.6.2 Project: Money Matters	32
1.6.3 Project: Ant Tunneling	33
2 First Order Equations	37
2.1 First-Order Linear Equations	37
2.1.1 Example: Solving the Hill-Keller Equation as a Linear ODE	38
2.1.2 A General Procedure for Solving Linear ODE's	40
2.1.3 Some Common First-Order Linear Models	41
2.1.4 Exercises	45
2.2 Separable Equations	49
2.2.1 Application: Falling Objects	49
2.2.2 Separation of Variables: A First Example	50
2.2.3 The General Procedure for Separation of Variables	52
2.2.4 Example: Solving the Falling Object ODE	52
2.2.5 Example: Solving the Logistic Equation	54
2.2.6 Exercises	55
2.3 Qualitative and Graphical Insights	59
2.3.1 Direction Fields	59
2.3.2 Autonomous Equations	61
2.3.3 Phase Portraits	62
2.3.4 Fixed Points and Stability	64
2.3.5 Determining the Stability of Fixed Points	65
2.3.6 Bifurcations	67
2.3.7 Exercises	69
2.4 The Existence and Uniqueness of Solutions	72
2.4.1 Some Inspiration from Calculus 1	72
2.4.2 What Are Solutions to ODE's?	73
2.4.3 The Existence-Uniqueness Theorem for ODE's	74
2.4.4 Exercises	77
2.5 Modeling Projects	78
2.5.1 Project: Money Matters 2	78
2.5.2 Project: Chemical Kinetics	80
2.5.3 Project: A Shot in the Water	84
3 Numerical Methods for ODE's	87
3.1 The Need for Numerics	87
3.2 Euler's Method	88
3.2.1 Evaluate, Extrapolate, Repeat as Necessary	89
3.2.2 The Accuracy of Euler's Method	92
3.2.3 Exercises	94
3.3 Improvements to Euler's Method	96
3.3.1 Improving Euler's Method	97
3.3.2 The Improved Euler Method	99
3.3.3 Exercises	100

3.4 Modern Numerical Methods	102
3.4.1 The RK4 Algorithm	102
3.4.2 Adaptive Step Sizing and Error Control	104
3.4.3 Exercises	109
3.5 Parameter Estimation	111
3.5.1 Hill-Keller Revisited	111
3.5.2 Least-Squares Estimation	112
3.5.3 Hill-Keller Again	115
3.5.4 Least Squares For ODE Parameter Estimation	119
3.5.5 A Cautionary Example	120
3.5.6 Exercises	122
3.6 Modeling Projects	127
3.6.1 Project: Sublimation of Carbon Dioxide	127
3.6.2 Project: Fish Harvesting Revisited	129
3.6.3 Project: The Mathematics of Marriage	131
3.6.4 Project: Shuttlecocks and Model Selection	135
4 Second Order Equations	141
4.1 Vibration and the Harmonic Oscillator	141
4.1.1 The 2010 Chilean Earthquake	141
4.1.2 The Harmonic Oscillator	142
4.1.3 Initial Conditions	144
4.1.4 More Applications of Spring-Mass Models	144
4.1.5 Exercises	148
4.2 The Harmonic Oscillator	151
4.2.1 Solving the Harmonic Oscillator ODE: Examples	151
4.2.2 Solving Second Order Linear ODE's: The General Case	154
4.2.3 The Underdamped and Undamped Cases	157
4.2.4 The General Underdamped Case	160
4.2.5 The Critically Damped Case	162
4.2.6 The Existence and Uniqueness of Solutions	164
4.2.7 Summary and a Physical Perspective	165
4.2.8 Exercises	165
4.3 The Forced Harmonic Oscillator	171
4.3.1 Solving the Forced Harmonic Oscillator Equation	172
4.3.2 Finding a Particular Solution: Undetermined Coefficients	174
4.3.3 When the Guess Fails	180
4.3.4 Exercises	183
4.4 Resonance	186
4.4.1 An Example of Resonance	186
4.4.2 Periodic Forcing	187
4.4.3 Exercises	197
4.5 Scaling and Nondimensionalization for ODE's	200
4.5.1 Motivation: Nonlinear Springs	200
4.5.2 Characteristic Variable Scales	202
4.5.3 Rescaling Variables and Nondimensionalizing ODE's: Examples	204
4.5.4 The General Outline for Nondimensional Rescaling	208
4.5.5 Back to the Hard Spring	209
4.5.6 Exercises	212

4.6 Modeling Projects	216
4.6.1 Project: Earthquake Modeling	216
4.6.2 Project: Stayed Tuned—RLC Circuits and Radio Tuning	217
4.6.3 Project: Parameter Estimation with Second Order ODE's	219
4.6.4 Project: Bike Shock Absorber	221
4.6.5 Project: The Pendulum	222
4.6.6 Project: The Pendulum 2	224
5 The Laplace Transform	227
5.1 Discontinuous Forcing Functions	227
5.1.1 Motivation: Pharmacokinetics	227
5.1.2 Complication: Discontinuous Forcing	228
5.1.3 Complication: Impulsive Forcing	229
5.1.4 Discontinuous Forcing and Transform Methods	230
5.1.5 Exercises	230
5.2 The Laplace Transform	232
5.2.1 Definition of the Laplace Transform	232
5.2.2 What Kinds of Functions Can Be Transformed?	234
5.2.3 Laplace Transforms of Elementary Functions	235
5.2.4 Solving Differential Equations Using Laplace Transforms	238
5.2.5 The First Shifting Theorem	242
5.2.6 The Inverse Laplace Transform	243
5.2.7 The Initial and Final Value Theorems	246
5.2.8 Section Summary and Remarks	248
5.2.9 Exercises	248
5.3 Nonhomogeneous Problems and Discontinuous Forcing Functions	252
5.3.1 Some Nonhomogeneous Examples	252
5.3.2 Discontinuous Forcing	253
5.3.3 Laplace Transforming $H(t-c)$	255
5.3.4 The Second Shifting Theorem	255
5.3.5 Some More Models and Examples	259
5.3.6 Summary and Remarks	261
5.3.7 Exercises	262
5.4 The Dirac Delta Function	265
5.4.1 Motivational Examples	265
5.4.2 Definition of the Dirac Delta Function	268
5.4.3 Three Models	272
5.4.4 The Laplace Transform of the Dirac Delta Function	273
5.4.5 Solving ODE's with Dirac Delta Functions	274
5.4.6 Summary and a Few Remarks	276
5.4.7 Laplace Transform Table	276
5.4.8 Exercises	276
5.5 Input-Output, Transfer Functions, and Convolution	279
5.5.1 A System Identification Problem	279
5.5.2 Input-Output Systems	280
5.5.3 Convolution	282
5.5.4 The Impulse Response of a System	285
5.5.5 Using Transfer Functions and Impulse Responses	286
5.5.6 System Identification with Impulsive Input	287

5.5.7	Exercises	289
5.6	A Taste of Control Theory	292
5.6.1	The Need for Control	292
5.6.2	Modeling an Incubator	293
5.6.3	Open Loop Control	294
5.6.4	Closed-Loop Control	297
5.6.5	Proportional-Integral Control	301
5.6.6	Proportional-Integral-Derivative Control	303
5.6.7	Disturbances	304
5.6.8	Summary and Comments	307
5.6.9	Exercises	307
5.7	Modeling Projects	309
5.7.1	Project: Drug Dosage	309
5.7.2	Project: Machine Replacement	310
5.7.3	Project: Vibration Table Shakedown	313
5.7.4	Project: Segway Scooters and The Inverted Pendulum	315
6	Linear Systems of Differential Equations	319
6.1	Systems of Differential Equations	319
6.1.1	Motivation: More Pharmacokinetics	319
6.1.2	Existence and Uniqueness	324
6.1.3	Exercises	325
6.2	Linear Constant Coefficient Homogeneous Systems of Differential Equations	328
6.2.1	Matrix-Vector Formulation	328
6.2.2	Solving the Homogeneous Case	328
6.2.3	Complex Eigenvalues	332
6.2.4	Defective Matrices	335
6.2.5	Exercises	337
6.3	Linear Constant Coefficient Nonhomogeneous Systems of Differential Equations	339
6.3.1	The Nonhomogeneous Equation $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{f}$ via Laplace Transforms	339
6.3.2	Solution to $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{f}$ via Undetermined Coefficients	341
6.3.3	The Significance of Eigenvalues	344
6.3.4	Exercises	344
6.4	The Matrix Exponential	347
6.4.1	Inspiration	347
6.4.2	Definition of the Matrix Exponential	348
6.4.3	Properties of the Matrix Exponential	349
6.4.4	The Matrix $e^{t\mathbf{A}}$	350
6.4.5	Solving ODEs with the Matrix Exponential	351
6.4.6	Computing The Matrix Exponential: The Diagonal Case	352
6.4.7	Computing The Matrix Exponential: The Diagonalizable Case	353
6.4.8	Computing The Matrix Exponential: Putzer's Algorithm	355
6.4.9	Final Remarks	357
6.4.10	Exercises	358

6.5 Modeling Projects	359
6.5.1 Project: LSD Compartment Model	359
6.5.2 Project: Homelessness	361
6.5.3 Project: Tuned Mass Dampers	362
7 Nonlinear Systems of Differential Equations	367
7.1 Autonomous Nonlinear Systems and Direction Fields	368
7.1.1 Some Nonlinear ODE Models	368
7.1.2 Direction Fields	371
7.1.3 A Nonlinear Direction Field Example	374
7.1.4 Direction Fields in Higher Dimensions	375
7.1.5 Exercises	376
7.2 Direction Fields and Phase Portraits for Linear Systems	377
7.2.1 Direction Fields for Homogeneous Linear Systems	377
7.2.2 Application to the LSD Model	382
7.2.3 The Equation $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}$	385
7.2.4 Direction Fields for Larger Systems of ODEs	386
7.2.5 Exercises	386
7.3 Autonomous Nonlinear Systems and Phase Portraits	389
7.3.1 The Struggle for Existence Continues	389
7.3.2 Changing the Parameters	393
7.3.3 Sketching Phase Portraits with Unspecified Parameters	394
7.3.4 Linearizing Multivariable Functions	397
7.3.5 Linearizing ODEs at Equilibrium Points	399
7.3.6 Linearizing the Competing Species Model with General Parameters	402
7.3.7 Conclusions for Competing Species	404
7.3.8 Higher Dimensional Systems	405
7.3.9 Exercises	405
7.4 Modeling Projects	407
7.4.1 Project: Homelessness Revisited	407
7.4.2 Project: Predator-Prey Model	408
7.4.3 Project: Parameter Estimation for Competing Species	409
8 A Brief Introduction to Partial Differential Equations	413
A Appendix Complex Arithmetic	417
Appendices	
B Appendix Matrix Algebra Review	419
C Appendix Circuits	421
Bibliography	435
Index	443

1. Why Differential Equations?

To begin, we offer the reader mathematical models of three quite different physical situations. Remarkably, all can be described by similar mathematics. These three examples and many others are woven 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 Olympic Games’ 100 meter dash final in Beijing in 2008 [15]. The times belong to the gold medal winner Usain Bolt and represent a world record, 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 [9].

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.
- (3) In a later chapter, extend the model to a greater range of running distances and use such a model to recommend the optimal strategy that a runner should adopt to cover any given distance as quickly as possible, given the runner’s physical abilities.

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 [15].

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 [56, 62]. 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 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 $x > 0$ 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, and 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 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 notated 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 Express the sprinter’s acceleration $a(t)$ in terms of velocity $v(t)$.

Reading Exercise 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 or "will," 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 3 gives some ideas on how we might interpret P physically.

Reading Exercise 3

- (a) Given that F_p has the physical dimension of force and m is a mass, what is the physical dimension for P ?
- (b) Suppose the runner is standing still, with no forces acting on the runner, and 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 [84]). In what follows this is the choice we'll use for P , at least for now. In doing so we are fixing SI units for our analysis, 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, this 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 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 flag 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 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 to show that the function $v(t)$ must satisfy

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

Note that m drops out! As previously noted, for the moment we'll use $P = 11$ meters per second squared, in SI units.

Equation (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 solve (1.3) for the unknown $v(t)$.

Reading Exercise 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 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 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) above 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, that you may have hit up on Reading Exercise 5.

Reading Exercise 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 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 exists 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 higher level 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 first, it will be helpful to develop some techniques for solving differential equations, and for determining the “best value” for k .

Reading Exercise 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 throughout $t > 0$? 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 was based on basic physics, specifically Newton's Second Law. The next section illustrates another common modeling technique.

Reading Exercise 9 Review the derivation of (1.3) and list every assumption we made in deriving this model.

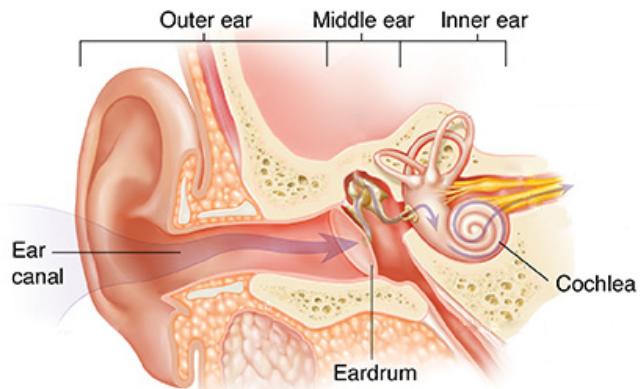


Figure 1.1: Human ear anatomy (adapted from [12]).

1.2 Intracochlear Drug Delivery

The material in this section is based on the SIMIODE project “Intracochlear Drug Delivery” [94].

1.2.1 The Challenge of Hearing Loss

Over 5% of the world’s population – or 466 million people – have disabling hearing loss [107]. The World Health Organization estimates that by 2050 over 900 million people, or about one in every ten people, will have a disabling 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 precisely dosing drugs to the cochlea.

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 [88]. This approach releases drugs directly to the inner ear, in order to establish 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. A prototype of such a system is shown in Figure 1.2. The implantable device is connected via 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, limitations 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 this concentration depends on the various physical parameters involved.

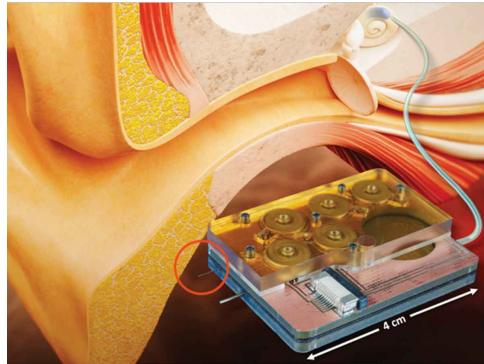


Figure 1.2: Delivery chip component of an intracochlear drug delivery device comprising microfluidic drug storage and flow control [83].

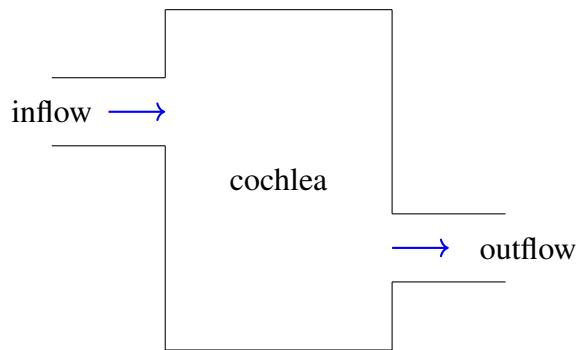


Figure 1.3: A compartmental model for cochlear drug delivery.

1.2.2 A Compartmental Model for the Cochlea

As a first approximation, consider a classic “compartmental” model in Figure 1.3. The input to the cochlea as depicted in Figure 1.2 is through a single “pipe” 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 from the cochlea.

However, we will model the situation as if the drug-containing fluid is introduced via one input (the “inflow” pipe in Figure 1.3) and withdrawn via 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 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 via one pipe; the drug then “instantaneously” diffuses 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 in for the drug is through the inflow pipe, the only way out is via the outflow pipe, and the drug is neither created nor destroyed in the cochlea. If during a given time interval Δt an amount A_1 micrograms (μg) enters via the inflow and an amount $A_2 \mu\text{g}$ exits via the outflow,

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 Δt time interval. The quantity $A_1/\Delta t$ is the average rate at which the drug enters via the inflow, and $-A_2/\Delta t$ is average rate the drug leaves via the outflow. The simple equation $(A_1 - A_2)/\Delta t = A_1/\Delta t - A_2/\Delta t$ merely states that

The average rate of change of the amount of drug contained 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

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

Equation (1.4) the basis of our mathematical model.

Reading Exercise 10 Let $u(t)$ denote the amount (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 quantity have, if the drug amount is in μg and time is in minutes?

Reading Exercise 10 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 via the inflow pipe, and the rate at which the drug is leaving. Suppose that fluid is entering the cochlea via the inflow pipe in Figure 1.3 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 11 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: it depends on r and c_1 .

Modeling Tip 2 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., “mass plus 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.

The rate at which the drug is leaving the cochlea is the trickiest part. Suppose that at a given time t there are $u(t)$ μg of the drug in the cochlea. Suppose the volume of the cochlea is V μL . 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 given 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 12 At what rate is the drug exiting the cochlea through the outflow pipe? Your answer should be in units of micrograms per minute ($\mu\text{g}/\text{min}$, physical dimensions of mass per unit time). Hint: it 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 10-12 we have

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

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, in this case that 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) equation (1.5) must be modified to account for this.

Note the similarity of equations (1.3) and (1.5). Despite the very different physical situations from which each equation arises, analysis techniques for one will likely be applicable to the other.

Reading Exercise 13 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 one μL 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 at a constant $200 \mu\text{L}$. What does equation (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 14 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 13, use this to determine how much of the drug (μg) will be in the cochlea after one week, and after two weeks. What is the concentration (μg per μL) of the drug in the cochlea at each of these times? Plot the amount and concentration of drug in the cochlea over time. What do you observe? Explain.

1.3 Population Growth and Fishery Management

The material in this section is based on the SIMIODE project "Fishery Harvesting" [40].

1.3.1 The Need to Manage Fish Harvesting

Fish are a valuable source of protein, and many people live to a large extent on fish and other seafood. However, over-fishing has driven many stocks of fish to near extinction [54, 55]. 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 [65]. 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 1.4. The 2012 assessments of Gulf of Maine and Georges Bank cod indicated that both stocks are seriously over fished 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 over fishing 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]. Now NOAA asks you to model the fish stock with harvesting in St. Georges Bank in order to fish sustainably.

1.3.2 Modeling Fish Population

Let us consider the cod population as existing or 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

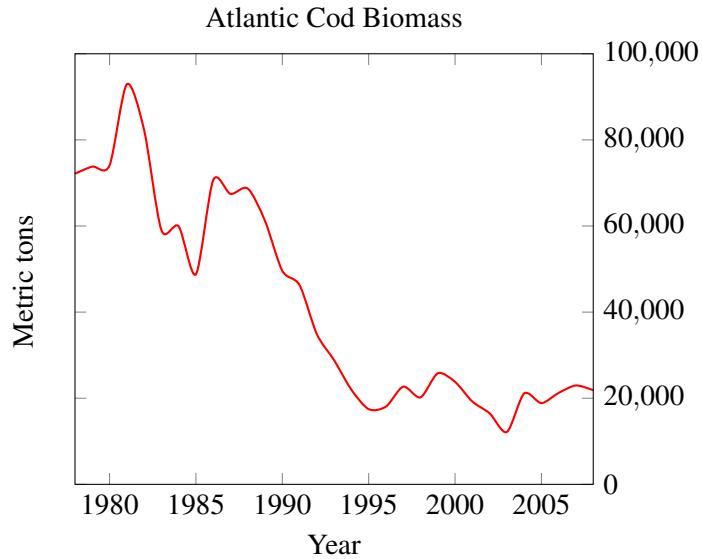


Figure 1.4: Atlantic Cod Biomass, 1978-2008.

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 embodied by the equation

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

This model is based on the assumption that at any given time the population produces new individuals at a rate proportional to the number of individuals present at that time. Here $r > 0$ is a constant of proportionality, and is called the *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) grows without limit, so the population tends to infinity! We need a better model.

Reading Exercise 15 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? The answer depends on r . How long for the population to quadruple? How long 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 $K > 0$.

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 16

- 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?
- 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?
- 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 r now carries the interpretation of 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 $u(t)$. Equation (1.10) is called the *logistic equation*.

Reading Exercise 17 What units on r are necessary for (1.10) (or (1.7)) to be dimensionally consistent, if u comes in units of “organisms” and t is measured in days? What units are necessary for K ?

Reading Exercise 18 As you will compute in the next chapter, the solution $u(t)$ to equation (1.10) with initial condition $u(0) = u_0$ is given by

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

Take $K = 10$, $r = 1$, and $u_0 = 2$ in (1.11). Plot the solution 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., 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 given by $hu(t)$. 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 them at rate $hu(t)$, the rate at which $u(t)$ is changing is just the difference between these quantities, $ru(t)(1 - u(t)/K) - hu(t)$. That is

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

Equation (1.12) is the *logistic equation with harvesting*. See [35] for more discussion of this model.

Reading Exercise 19 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?

Year	u_t	h_t	Year	u_t	h_t	Year	u_t	h_t
1978	72,148	0.18847	1988	68,702	0.231541	1998	20,196	0.189526
1979	73,793	0.149741	1989	61,191	0.208597	1999	25,776	0.170108
1980	74,082	0.219209	1990	49,599	0.335648	2000	23,796	0.156601
1981	92,912	0.176781	1991	46,266	0.295344	2001	19,240	0.281787
1982	82,323	0.282033	1992	34,877	0.331848	2002	16,495	0.252869
1983	59,073	0.34528	1993	28,827	0.350394	2003	12,167	0.255417
1984	59,920	0.206545	1994	21,980	0.282701	2004	21,104	0.081034
1985	48,789	0.338185	1995	17,463	0.199275	2005	18,871	0.0873972
1986	70,638	0.147236	1996	18,057	0.18781	2006	21,241	0.0819517
1987	67,462	0.19757	1997	22,681	0.193574	2007	22,962	0.105181
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 [108].

Reading Exercise 20 The solution to (1.12) (which we will deduce in the next chapter) is given by

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

Of course 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 K, r , and u_0 . Is it consistent with the modeling assumptions that were made?
- (b) Repeat part (a) but increase h to 0.5. What happens?
- (c) What is the limit $\lim_{t \rightarrow \infty} u(t)$ 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 [108] are given in Table 1.2 from 1978 to 2008. Note the fish population is estimated not in individuals, but in total mass. Nonetheless, 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 21 Although the harvest rate h in the table 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 $u(t)$ in (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 3.5 we'll look at more methodical and effective ways to find these parameters.

Reading Exercise 22 What is the long term behavior of the cod population with these parameters? What does harvesting do to the cod population? 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 [39, 60].

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 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 there was an additional piece of information, an initial condition of the form $u(t_0) = u_0$.

For the ODE's encountered so far we simply presented the reader with 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 ODE's like (1.14) and many others. In cases where an analytical solution cannot be found we will explore other techniques for gleaning information about solutions. The models developed in this chapter, their extensions, and additional 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 taken will depend greatly on the structure of the differential equation, so it's helpful to make a few definitions right up front.

Scalar ODE's, Systems, and PDE's

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,

$$\begin{aligned} u'(t) &= u(t) - u(t)v(t), \\ v'(t) &= -2v(t) + 3u(t)v(t) \end{aligned} \quad (1.15)$$

in which two (or more) unknown functions of a single independent variable t appear. The focus of the first five chapters will be scalar ODE's. In Chapters 6 and 7 techniques will be developed for analyzing system of ODE's.

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 provided by the *wave equation*

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0, \quad (1.16)$$

where $u(x, t)$ depends on two independent variables, x and t . One can even have systems of partial differential equations. Some elementary partial differential equations will be considered in Chapter 8.

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 in the ODE. Differential equations of second-order (involving $u''(t)$) are also common, and form much of the mathematical support for significant engineering applications. Occasionally higher-order equations make an appearance.

The distinction between “linear” versus “nonlinear” ODE’s 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

$$a_n(t)u^{(n)}(t) + a_{n-1}(t)u^{(n-1)}(t) + \cdots + a_1(t)u'(t) + a_0(t)u(t) = b(t) \quad (1.17)$$

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

$$F(t, u(t), u'(t), \dots, u^{(n)}(t)) = 0 \quad (1.18)$$

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 (1.17) then the differential equation is *linear*.

There are other important adjectives to describe ODE’s, such as “constant-coefficient,” “autonomous,” “separable,” “homogeneous,” and so on. These definitions will be presented later, in context when they naturally arise. Each type of equation demands different techniques for analysis. However, as noted in the next section, if you’ve taken a Calculus course, you already know how to solve some differential equations.

1.4.3 You Already Know How to Solve Some Differential Equations

The reader is 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 antiderivatives.

First-Order Equations

Consider a first-order differential equation of the form

$$u'(t) = g(t) \quad (1.19)$$

where $g(t)$ is a given function and $u(t)$ is the unknown to be found. Compare (1.19) to (1.14)—the right side in (1.19) does not involve the unknown function u . In this case one can simply antidifferentiate both sides of (1.19) with respect to t and find

$$u(t) = \int g(t) dt + C \quad (1.20)$$

where C is an arbitrary constant of integration; this is a *general solution* to the ODE, for all functions that satisfy (1.19) can be expressed in the form (1.20) for some choice of C . If an initial condition $u(t_0) = u_0$ is given then the constant C can be adjusted to obtain 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 $u(1) = 3$.

Antidifferentiating 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 thus $u(t) = t^3 + 2$.

■

Reading Exercise 23 Verify that $u(t) = t^3 + 2$ does 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 24 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

$$u''(t) = g(t). \quad (1.21)$$

Integrate both sides of (1.21) with respect to t to find

$$u'(t) = \int g(t) dt + C_1 \quad (1.22)$$

where C_1 is an arbitrary constant. Let $G(t) = \int g(t) dt$ be any antiderivative for $g(t)$, so (1.22) becomes $u'(t) = G(t) + C_1$. Integrating both sides of (1.22) with respect to t then yields

$$u(t) = \int G(t) dt + C_1 t + C_2 \quad (1.23)$$

where C_2 is a second arbitrary constant of integration. Equation (1.23) is a general solution to (1.21). To find the constants C_1 and C_2 one needs two additional pieces of information about the solution. This typically (but not always) takes 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 $u'(1) = 3$ forces $e + C_1 = 3$, so $C_1 = 3 - e$. Also, $u(1) = 2$ forces $e + (3 - e) + C_2 = 2$, so $C_2 = -1$. The solution with the required initial condition is

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

■

Reading Exercise 25 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 26 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 . Simply 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 2 The astute reader may note 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.

- (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 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 was the “salt.” The abstract situation is still accurately depicted by Figure 1.3.

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 any given time. If the tanks starts with no salt at time $t = 0$, use the reasoning that led to equation (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 is the limiting amount of salt in the tank? What is the limiting concentration? 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 and their application 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 ODE’s, and provide a sanity check for our work. This is the subject of *dimensional analysis*.

Mass, Length, and Time

To illustrate, 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 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 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 or negative. In many cases a, b , and c will be integers, but not all.

Dimension Versus Units

The notion of dimension isn’t quite the same as “units.” Thus “length” is a fundamental physical dimension, but it can be measured using many systems of units, e.g., meters, feet, inches, etc. If the author of this text ever gets sloppy and refers the dimension of a velocity as “meters per second” feel free to write him a stern email. However, specifying the units of a given quantity does allow us to determine its dimension.

Reading Exercise 27 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 this string equals the mass of this apple" or "this time interval is longer than that 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, divide 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

$$[xy] = [x][y] = M^{a_1+b_1} L^{a_2+b_2} T^{a_3+b_3} \quad \text{and} \quad [x/y] = [x]/[y] = M^{a_1-b_1} L^{a_2-b_2} T^{a_3-b_3}. \quad (1.24)$$

Reading Exercise 28 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 mathematical constants appear and these constants are *dimensionless*. 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 accord with (1.24) 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"? 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

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

Reading Exercise 29 Newton's Universal Law of Gravitation specifies that the force F of two point masses m_1 and m_2 separated by a distance r is given by

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

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

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 above 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 31 Water flows into a tank at a variable rate $r(t)$ 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 , and so on, all elementary transcendental functions 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 . 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 $P = 11$ 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 to the cosine function is dimensionless, in accord 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 a whatever units we choose to measure length. In principle this is what we’ll do, we just won’t remark on it. It should not cause any 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 29 you showed that $[G] = M^{-1}L^3T^{-2}$.

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

$$r = KG^\alpha c^\beta m^\gamma \quad (1.25)$$

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

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

If the right side above is to have the same dimension as r 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 (1.25), 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 (1.25). The dimensionless constant K can be anything.

In fact, with $K = 2$ the formula is correct! The radius r is called the *Schwarzschild radius* for 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 physics result 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 denote by Q . These are independent dimensions from mass, length, and time. In certain specific instances it can be helpful to temporarily assign other dimensions. For example, in an economic or “money” problem we could use V to denote the dimension “value,” that is, the worth of some quantity; it might be tempting to use “\$” but that is a unit, dollars. See the project “Money Matters” in Section 1.6 and [90]. 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 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 to the exponential function dimensionless? ■

Exercise 1.5.6 Recall the fish harvesting model of Section 1.3, and in particular, the ODE (1.10). Of course t in that equation is time, but u has no obvious dimension. Let us take $[u] = N$, where N denotes “fish population.” (Although we could consider u as dimensionless since it simply counts how many fish are present, but 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, 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 real 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 c in a gas as a function of its pressure P and density ρ , in the form $c = P^a \rho^b$. Note pressure is "force per area". ■

Exercise 1.5.11 Find a plausible formula the frequency f of a string's vibration in term of its linear density λ , the tension τ in the string (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 the 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" [100].

The phrase "hang time" is common in sports—an announcer in a football game might 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 phenomena. 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 $g \approx 9.81$ 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 referenced to this) and $y > 0$ 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 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 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 3 Put Modeling Exercises 1 and 2 together to find a second-order differential equation for $y(t)$. One initial condition from above 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 4 Find a general solution to the differential equation in Modeling Exercise 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 involve v_0 .

Modeling Exercise 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

$$v_0 t_1 - \frac{1}{2} g t_1^2 = 1, \quad (1.26)$$

in SI (metric) units, so the "1" on the right in (1.26) signifies 1 meter. Verify that all terms in (1.26) have the same dimension, length.

Modeling Exercise 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 (1.26) from Modeling Exercise 5 to find v_0 and t_1 . Work in SI units with $g = 9.81$ meters per second squared.

Modeling Exercise 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

$$p(t) = 50000e^{t/45} \quad (1.27)$$

dollars per year (so you make \$50,000 per year starting at age 22, and $p(45) \approx 135,914$ dollars the year you retire). Note that the argument $t/45$ to the exponential function in (1.27) 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 this 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 1 Explain why $S(t)$ obeys the ODE

$$S'(t) = 0.1p(t) \quad (1.28)$$

with $p(t)$ given by (1.27). 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 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 3 Find the particular solution to (1.28) that satisfies the initial condition from Modeling Exercise 2.

Modeling Exercise 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?

In later chapters we can add more realistic assumptions to this model (taxes, inflation, kids). But we’ll also show that there is hope for retiring, with compound interest working for you!

1.6.3 Project: Ant Tunneling

This project is based on the SIMIODE Modeling Scenario “Ant Tunnel Building” [98]; see also [96].

If you ever had an ant colony purchased for you by a well-meaning relative when you were in grade school, or even 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 1.5.¹ 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 .

Modeling Exercise 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 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 1.6, a depiction of the situation that is focused on the essential geometry. We’ll make some assumptions that reflect the geometry and physics of the situation, 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 $x + h$.

Modeling Exercise 2 On the right in (1.29), 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.

$$T(x+h) - T(x) = \underline{\hspace{10em}} \quad (1.29)$$

¹Ant drawing provided by Isaac H. All.

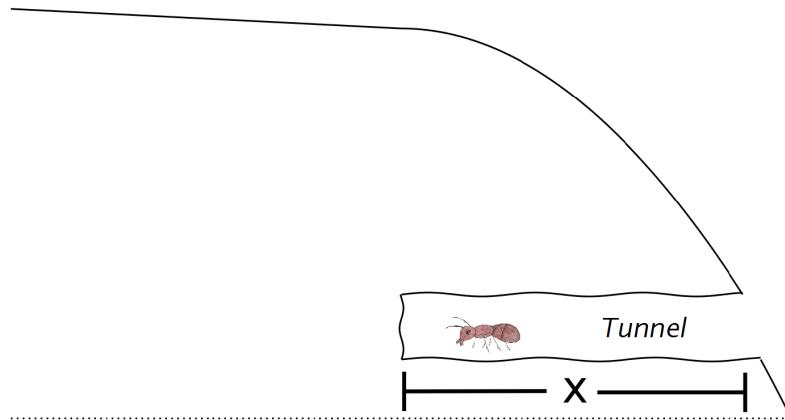


Figure 1.5: An ant building a tunnel. Here x is the current length of the tunnel and $T(x)$ is the time it has taken the ant to build the tunnel of length x .

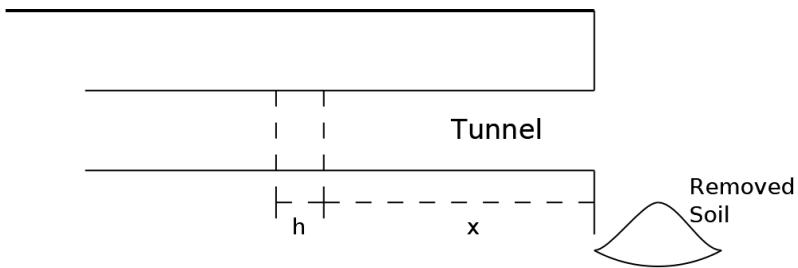


Figure 1.6: Useful diagram for discovering the time it takes to extend a small section of the ant tunnel from distance x to $x + h$.

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 (1.29). Defend or reject each and offer your reasons. Perhaps modify one or two to improve them. When trying to reject a model consider some trivial cases and see if it 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 3 Convert your model difference equation (1.29) to a related differential equation with appropriate initial conditions. It may be helpful to consider the expression $(T(x+h) - T(x))/h$ and what happens as h approaches 0.

Modeling Exercise 4 Solve the differential equation you create in Modeling Exercise 3 for $T(x)$. Hint: What initial condition $T(0)$ will you use?

Modeling Exercise 5 Use your solution from Modeling Exercise 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 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 7 Of course, the same principles can be applied to mode real tunnel building for engineers. If we were considering Modeling Exercise 6 as related to engineering 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 the tunnel.

Further SIMIODE Projects

Here would go a listing and brief descriptions of other suitable SIMIODE modeling projects for this chapter, perhaps simply a link to such a list on the SIMIODE website.

2. First Order Equations

In this chapter we'll look at techniques for solving or otherwise analyzing first order ODE's of the form (1.14), that is, $u'(t) = f(t, u(t))$. How we proceed depends on the function f , which determines the nature of the differential equation. We will examine solution techniques for two common classes of first-order ODE's: those that are *linear* and those that are *separable*; some ODE's 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 this case 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 ODE's 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 ODE's when an analytical solution cannot be found.

Remark 3 At times we will write ODE's without the independent variable explicitly attached to the unknown function. That is, we will write $u' = f(t, u)$ instead of $u'(t) = f(t, u(t))$. This sometimes makes it easier to see the structure of the equation.

2.1 First-Order Linear Equations

An ODE in the standard form $u' = f(t, u)$ is linear if $f(t, u) = g(t) + h(t)u$. That is, a linear first-order ODE is of the form

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

for some functions g and h . Linearity and several other common terms are part of the following definition.

Definition 2.1.1 — Linear First-Order ODE's. A first-order ODE that can be written in the form (2.1) is *linear*, and otherwise it is *nonlinear*. If $g(t)$ is the zero function then the ODE is *homogeneous*, otherwise the ODE is *nonhomogeneous*. If $h(t)$ is a constant function then (2.1)

is a *constant coefficient* ODE, otherwise (2.1) is a *variable coefficient* ODE.

■ **Example 2.1** Let us classify the ODE's (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 nonhomogeneous (if $P \neq 0$).
- 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 is a quadratic function of u .

In Section 1.4.3 certain ODE's 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: simply integrating both sides of (2.1) with respect to t leads to something like

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

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 (1.19) from the last chapter; direct integration succeeds there precisely because $h(t)$ was the zero function, so we only had to integrate $g(t)$.

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

Before showing the general technique for solving linear first-order ODE's, 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 of course 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} here is the *integrating factor* for this ODE. You can check using the product rule that

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

for any function $v(t)$, and the right side of (2.4) is precisely the left side of (2.3). The use of (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 . This gives

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

where C_1 and C_2 are arbitrary constants of integration. And as obvious as it sounds, by "constant" we mean that they do not depend on the independent variable t . Notice that all derivatives have disappeared—finding $v(t)$ is now an algebra problem.

Reading Exercise 32 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, for any C equation (2.8) provides a solution to (1.3). That's why (2.8) is called the "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 $P/k + C = 0$. Solve for C as $C = -P/k$, and in (2.8) this yields the particular solution with the required initial condition $v(0) = 0$,

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

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

Reading Exercise 33 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 34 Let $k = 1$ (units reciprocal seconds) in the solution from Reading Exercise 33, 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 ODE's

The procedure used to solve the Hill-Keller ODE works more generally on 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 2.1.21.

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 in (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.

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 $u(1) = 2$, at least for $t > 0$. 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 $\ln|t|$; however, since we are interested in $t > 0$ we'll drop the absolute values. We'll discuss issues concerning the domain of the solution to an ODE later. The integrating factor is then $e^{-\ln(t)} = 1/t$. Multiply both sides of the ODE by $1/t$ and obtain

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

The left side above is just $\frac{d}{dt}(u(t)/t)$, which leads to

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

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

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

Multiply both sides by t to obtain general solution

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

To obtain $u(1) = 2$ substitute $t = 1$ 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 35 In step 2 of the integrating factor method above 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 36 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) that was obtained in equation (1.20) of Chapter 1.

Reading Exercise 37 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

$$u(t) = Ce^{H(t)} \tag{2.14}$$

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 ODE's. Each of the ODE's 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 corpse's body temperature. 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 [16].

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 the temperature is a function of position inside the object then 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 will be known as the *ambient* temperature.

- The object does not generate any internal heat, but changes temperature due simply to losing or gaining 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 $u(t)$ the temperature of the object at time t (again, u is constant throughout the object), so $[u] = \Theta$. Let A denote the ambient temperature in which the object exists, and 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 the 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 plain English. The language of ODE's makes it possible to state it much more concisely.

Reading Exercise 38

- What is a simple mathematical expression for the quantity "The rate at which the object's temperature changes"?
- What is a simple mathematical expression for the quantity "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.

In Reading Exercise 38 you should be led to the differential equation

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

This formulation clearly requires $k < 0$, for if $u(t) > A$ then $u'(t) < 0$ (the object is cooling), while $u(t) < A$ should force $u'(t) > 0$ (the object is warming). In view of Modeling Tip 1, require $k > 0$ and put an explicit minus sign on the right side of the ODE above to obtain Newton's Law of Cooling in the form

$$u'(t) = -k(u(t) - A). \quad (2.15)$$

Reading Exercise 39 Is the Newton Cooling DE (2.15) linear or nonlinear? If linear, is it constant or variable coefficient? Homogeneous or nonhomogeneous?

Reading Exercise 40 Given that $[u] = \Theta$, what is $[u']$? What must the dimension $[k]$ of k be?

■ **Example 2.4** 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 . Find the temperature of the object as a function of time t . At what time t is $u(t) = 80^\circ\text{F}$?

To answer this question first note that, as you will demonstrate in Exercise 2.1.13, the solution to the Newton Cooling ODE (2.15) with initial condition $u(0) = u_0$ is

$$u(t) = A + (u_0 - A)e^{-kt}. \quad (2.16)$$

With $u_0 = 98.6$ and $A = 72$ this becomes $u(t) = 72 + 26.6e^{-kt}$. Given the information that $u(3) = 94$, (2.16) yields $94 = 72 + 26.6e^{-3k}$. This can be solved for k to find $k \approx 0.0633$ (units are reciprocal hours). Then

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

approximately. The equation $u(t) = 80$ is then $72 + 26.6e^{-0.0633t} = 80$, which can be solved for $t \approx 18.98$ hours. ■

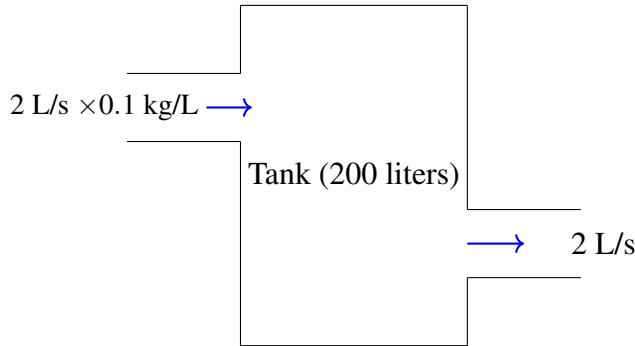


Figure 2.1: A typical “salt tank” compartmental model.

Salt Tank Problems

The next example is a generalization of the Cochlear Drug Delivery model that we considered in Section 1.2.2. In “salt tank” problems the quantity “salt” refers to any substance that is neither created nor destroyed in the course of the problem. These are another type of compartmental model.

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 via 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 $t > 0$, and how this amount of salt behaves in the limit that $t \rightarrow \infty$.

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 on a mass per time basis. 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 it 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 “rate in minus rate out,” so from the analysis above

$$\frac{dx}{dt} = 0.2 - \frac{x(t)}{100} \quad (2.17)$$

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 (2.17) is a linear first-order differential equation. In Exercise 2.1.14 at the end of this section you will show that the solution is

$$x(t) = 20 - 17e^{-t/100}. \quad (2.18)$$

From (2.18) it follows that $\lim_{t \rightarrow \infty} x(t) = 20$ kg, since the exponential term $e^{-t/100}$ decays to zero.

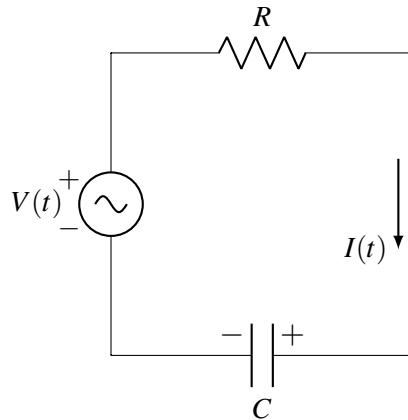


Figure 2.2: Single loop RC series circuit.

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 (in the sense of “conventional current”, the flow of positive charge), with $I > 0$ indicating clockwise current, and let $q(t)$ denote the charge on the capacitor (with the positive side $+q$ closest to the voltage source, $-q$ on the other side). Then $[q] = Q$ and since current is the flow rate of charge through the wire, $[I] = QT^{-1}$. If we start at the negative side of the voltage source and “step” around the RC loop the clockwise direction, we gain $V(t)$ volts over the source, then have a voltage drop across the resistor of $RI(t)$ 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

$$V(t) - q(t)/C - RI(t) = 0. \quad (2.19)$$

Since the charge entering the positive side of the capacitor flows in through wire from the voltage source it follows that $q'(t) = I(t)$, and then (2.19) can be written as

$$Rq'(t) + q(t)/C = V(t). \quad (2.20)$$

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

Reading Exercise 41 As shown in Appendix C, voltage has the dimensions of work per charge, which is $[V] = ML^2T^{-2}Q^{-1}$. Use this with (2.20) 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 time $t = 0$ when the circuit is completed 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 voltage V_R across the resistor.

With these values the ODE (2.20) becomes

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

with initial condition $q(0) = 0$. A general solution (the integrating factor approach is appropriate here) is $q(t) = 1/200000 + Ce^{-10000t}$ and $q(0) = 0$ yields

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

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

In Exercises 2.1.1 to 2.1.10, find a general solution to the linear ODE, and then find the specific solution with the given initial condition by using the integrating factor technique from Section 2.1.2.

Exercise 2.1.1 $u'(t) = u(t) + 3$, $u(0) = 3$

Exercise 2.1.2 $u'(t) = 2u(t) + 4$, $u(0) = 0$

Exercise 2.1.3 $u'(t) = -3u(t) + 3$, $u(0) = 5$

Exercise 2.1.4 $u'(t) = -3u(t) + 9t$, $u(0) = 5$

Exercise 2.1.5 $u'(t) = u(t) + 2\sin(t)$, $u(0) = 1$

Exercise 2.1.6 $u'(t) = -4u(t) + e^t$, $u(0) = 2$

Exercise 2.1.7 $u'(t) = tu(t) + t$, $u(0) = 2$

Exercise 2.1.8 $u'(t) = u(t)/t + 2$, $u(1) = 3$

Exercise 2.1.9 $u'(t) = \sin(t)u(t) + \sin(t)$, $u(0) = 4$

Exercise 2.1.10 $u'(t) = au(t) + b$, $u(0) = u_0$, a, b, u_0 constants.

Exercise 2.1.11 Solve equation (1.5) from the previous chapter with $u(0) = 0$, and so verify the formula (1.6). ■

Exercise 2.1.12 Many physical processes, e.g. radioactive decay, are governed by an ODE of the form

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

where $k > 0$ is some 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 limits to zero for any u_0 .
- (c) Show that the amount of time Δt necessary for the solution to decrease by half is given by $\Delta t = \ln(2)/k$, and that Δt has the dimension of time. 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 2.1.13 Solve the Newton Cooling ODE (2.15) with initial condition $u(0) = u_0$, by using the integrating factor approach, and so demonstrate that (2.16) is correct. ■

Exercise 2.1.14 Solve the compartmental salt tank ODE (2.17) with initial condition $x(0) = 3$, by using the integrating factor approach, and so demonstrate that (2.18) is correct. ■

Exercise 2.1.15 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 apothecary because her room mate 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 $u(t)$ equals 102.4. ■

Exercise 2.1.16 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 2.1.17 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!).

- (a) Find the volume $V(t)$ of liquid in the tank as a function of time.
- (b) 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.
- (c) Solve the ODE using the integrating factor approach and plot the solution. How much salt is in the tank at any given time?
- (d) When is the amount of salt in the tank maximized, and how much salt is in the tank at this

time?

Exercise 2.1.18 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 (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?
- How long after the rock hits the water will it hit the bottom of the river?
- What is the rock's velocity when it hits the bottom?

Exercise 2.1.19

- In the case that $V(t) = V_0$ is constant, use the integrating factor approach to show that the solution to (2.20) with $q(0) = 0$ is given by

$$q(t) = CV_0(1 - e^{-t/(RC)}).$$
- What is the limiting value of the charge on the capacitor as $t \rightarrow \infty$?
- Show that the product RC has the dimension time. Hint: Look back at Reading Exercise 41. The product RC is called the *RC time constant* for this circuit.
- How long does it take the charge to attain 99 percent of this limiting value from part (b)? Compare this to the rule of thumb that a capacitor takes $5RC$ time units to effectively reach full charge.

Exercise 2.1.20 Capacitors can be used in circuits to act as “filters” that allow sinusoidal input voltages in a given frequency range to “pass” while attenuating 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.

- Show that a general solution to (2.20) 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 $B = \frac{V_0}{10^6(\omega^2+1)}$.

- 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

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

- Recall that a function of the form $f(t) = A_1 \cos(\omega t) + A_2 \sin(\omega t)$ is sinusoidal with

frequency ω and amplitude $\sqrt{A_1^2 + A_2^2}$. 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*? ■

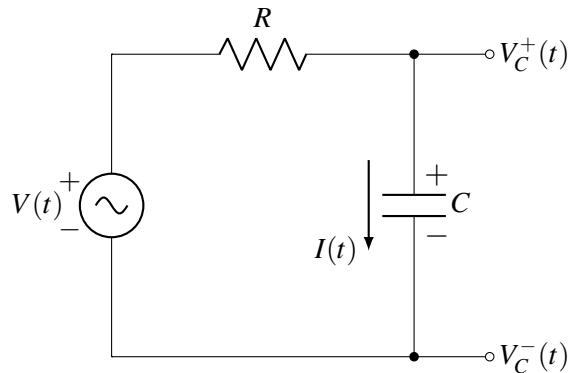


Figure 2.3: Single loop RC series circuit.

Exercise 2.1.21 In trying to solve a linear ODE written in the form

$$u'(t) - h(t)u(t) = g(t), \quad (2.21)$$

it's not at all obvious how anyone might hit upon the idea of multiplying both sides of (2.21) 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 (2.21) looks a little like the product rule for derivatives applied to the product $w(t)u(t)$ for some function $w(t)$, namely

$$(w(t)u(t))' = w(t)u'(t) + w'(t)u(t), \quad (2.22)$$

although the left side of (2.21) and right side of (2.22) aren't quite the same. But if we multiply the left side of (2.21) 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 (2.22) 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 $H'(t) = h(t)$. ■

2.2 Separable Equations

The other common class of first-order ODE's that can often be solved analytically are those that are *separable*. The precise definition is

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

$$u'(t) = g(t)h(u(t)) \quad (2.23)$$

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 this ODE 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 $f(t, u)$ isn't clearly separable, as in (c), it's probably not.

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. The gravitational force on the object is $F_g = mg$ with $g > 0$ as gravitational acceleration, if downward is the positive direction, which is convenient here. 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 42 gives some justification for this choice, however.

Reading Exercise 42 A spherical object with cross sectional area A moves at speed v 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, 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 42 to write

$$F_r = kv^2 \quad (2.24)$$

where $k = KA\rho$ is a constant with dimension $[k] = ML^{-1}$. More generally, we can posit that (2.24) holds whether the object is spherical or not; the constant k will certainly depend on the object's properties, e.g., size or cross-sectional area, as well as the density of the fluid.

An ODE for the Object's Motion

Take a coordinate system in which downward is positive and assume the object is falling straight down with velocity $v > 0$. Technically v is a vector, but since we are interested only in vertical motion we treat v as a scalar, and then the speed of the object is $|v|$. In view of the discussion above the net force F on the object is $F = F_g + F_r = mg - k|v|^2$ or, since we can drop the absolute values,

$$F = mg - kv^2 \quad (2.25)$$

where $g > 0$, and also $k > 0$ so that $-kv^2$ is upward. Note that if the object was rising ($v < 0$) then the force of air resistance would be downward and this would require $F = mg + kv^2$.

For such a falling object, Newton's Second Law $F = ma$, $a = v'(t)$, and (2.25) yield $mv' = mg - kv^2$ or

$$v'(t) = g - \frac{k}{m}v^2(t). \quad (2.26)$$

If the object is dropped at time $t = 0$ with zero initial velocity then $v(0) = 0$. Equation (2.26) is an ODE for $v(t)$. The equation is nonlinear, so its solution cannot be obtained using the integrating factor approach. We'll instead use separation of variables, but rather than start with (2.26) (which has some unpleasant algebra that clouds the central issues) let's begin with a more straightforward example and return to (2.26) shortly. See also the project "A Shot in the Water" in Section 2.5.

Reading Exercise 43 Verify that (2.26) is separable.

2.2.2 Separation of Variables: A First Example

Consider the problem of finding a general solution to the ODE $u'(t) = tu^2(t)$ by using separation of variables, and then finding the solution with initial condition $u(1) = 4$. This ODE is separable, of 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 is to manipulate the ODE so that integration actually makes progress toward a solution.

Step 1 is to separate as

$$\frac{u'(t)}{u^2(t)} = t. \quad (2.27)$$

Step 2 is to integrate both sides of (2.27) with respect to t to obtain

$$\int \frac{u'(t)}{u^2(t)} dt + C_1 = \int t dt + C_2. \quad (2.28)$$

The right side of (2.28) is easy to integrate and is $t^2/2 + C_2$. The left side of (2.28) 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 $w = u(t)$, so $dw = u'(t)dt$ 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 (2.28) can be written as

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

with $C = C_2 - C_1$.

Step 3 is to solve for $u(t)$. Multiply (2.29) through by -1 and reciprocate to find

$$u(t) = -\frac{1}{t^2/2 + C}. \quad (2.30)$$

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

Step 4 is to adjust C to obtain $u(1) = 4$. Substituting $t = 1$ and $u(1) = 4$ into (2.30) yields $4 = -\frac{1}{1/2+C}$, which can be solved to find $C = -3/4$. The solution 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 $u'(t) = tu^2(t)$ of Section 2.2.2 with this notational approach, and obtain 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 to 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

$$\frac{du}{u^2} = t dt. \quad (2.31)$$

Compare this to (2.27) above. 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 (2.31), treating the left side as an integral with respect to u and the right side as an integral with respect to t . This yields

$$\int \frac{du}{u^2} + C_1 = \int t dt + C_2 \quad (2.32)$$

where the constants of integration have already been incorporated on each side. Compare this to (2.28). Working both integrals in (2.32) yields

$$-\frac{1}{u} = t^2/2 + C \quad (2.33)$$

with $C = C_2 - C_1$. Compare this to (2.29); a substitution has been done to evaluate the integral on the left in (2.32) 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 (2.33) for u to find general solution

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

Then substitute in $t = 1$, $u = 4$ into the 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 (2.27). We'll use the Leibniz notational approach.

1. **Separate:** Write the ODE (2.23) in the form

$$\frac{du}{h(u)} = g(t) dt. \quad (2.34)$$

2. **Integrate:** Integrate both sides of (2.34) with respect to t :

$$\int \frac{du}{h(u)} + C_1 = \int g(t) dt + C_2 \quad (2.35)$$

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 (2.35) will become $G(t) + C_2$. The left side of (2.35) involves finding an antiderivative for $1/h(u)$ with respect to u . Define

$$Q(u) = \int \frac{du}{h(u)} \quad (2.36)$$

as the required antiderivative, so the left side of (2.35) is $Q(u) + C_1$. This means that (2.35) can be expressed as $Q(u) + C_1 = G(t) + C_2$, or

$$Q(u) = G(t) + C \quad (2.37)$$

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 (2.37) 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 (2.23), which can be written (conceptually, anyway) as

$$u(t) = Q^{-1}(G(t) + C). \quad (2.38)$$

4. **Obtain The Initial Condition:** The last step is to determine C in the general solution (2.38) to satisfy the initial condition $u(t_0) = u_0$, if an initial condition is given. This means solving $u_0 = Q^{-1}(G(t_0) + C)$ for C . If we back up to (2.37) this yields $C = Q(u_0) - G(t_0)$.

Reading Exercise 44 Solve $u'(t) = u^2(t) + 1$ with $u(0) = 0$ using separation of variables. Hint: write the right side of the ODE as $f(t, u) = 1(u^2 + 1)$.

2.2.4 Example: Solving the Falling Object ODE

Let's now find a general solution to the ODE (2.26) 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

$$\frac{dv}{dt} = -\frac{k}{m}v^2 + g. \quad (2.39)$$

Separate (2.39) by dividing both sides by $kv^2/m - g$ to obtain

$$\frac{dv}{kv^2/m - g} = -1. \quad (2.40)$$

Integrate both sides of (2.40) with respect to the appropriate variable to obtain

$$\int \frac{dv}{kv^2/m - g} = - \int 1 dt. \quad (2.41)$$

The integral on the right in (2.41) 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 2.2.14. An antiderivative is given by

$$\int \frac{dv}{kv^2/m - g} = \frac{1}{2} \sqrt{\frac{m}{gk}} \ln \left| \frac{v - \sqrt{mg/k}}{v + \sqrt{mg/k}} \right| + C_2. \quad (2.42)$$

The algebra is a bit simpler with the definition $\alpha = \sqrt{mg/k}$, in which case (2.42) becomes

$$\int \frac{dv}{kv^2/m - g} = \frac{\alpha}{2g} \ln \left| \frac{v - \alpha}{v + \alpha} \right| + C_2. \quad (2.43)$$

From (2.41) and (2.43) we obtain

$$\frac{\alpha}{2g} \ln \left| \frac{v - \alpha}{v + \alpha} \right| = -t + C \quad (2.44)$$

where $C = C_1 - C_2$ is an arbitrary constant.

The next step is to solve for v , which is merely algebra. Multiply both sides of (2.44) 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 simply 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

$$\ln \left| \frac{v - \alpha}{v + \alpha} \right| = -\frac{2g}{\alpha} t + C \quad (2.45)$$

where $C = C_1 - C_2$ is an arbitrary constant. Exponentiate both sides of (2.45) to obtain

$$\left| \frac{v - \alpha}{v + \alpha} \right| = e^C e^{-2gt/\alpha}. \quad (2.46)$$

Since C is an arbitrary (real) constant, isn't e^C also arbitrary? Not quite! 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 (2.46) becomes

$$\left| \frac{v - \alpha}{v + \alpha} \right| = C^+ e^{-2gt/\alpha}. \quad (2.47)$$

The absolute values in (2.47) 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 (2.47) can be written as

$$\frac{v - \alpha}{v + \alpha} = \pm C^+ e^{-2gt/\alpha} \quad (2.48)$$

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

$$\frac{v - \alpha}{v + \alpha} = C e^{-2gt/\alpha}. \quad (2.49)$$

Finally, routine algebra shows and the fact that $\alpha = \sqrt{mgk}$ shows that a general solution is given by

$$v(t) = \sqrt{\frac{mg}{k}} \left(\frac{1 + Ce^{-2t\sqrt{kg/m}}}{1 - Ce^{-2t\sqrt{kg/m}}} \right). \quad (2.50)$$

The constant C appears in two places in (2.50), which is fine. This solution could be written with C appearing only one, by why bother? (Recall Remark 2). Also, the restriction that $C \neq 0$ made above can be removed, since taking $C = 0$ in (2.50) 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 initial condition $v(0) = 0$ we need

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

which gives $C = -1$. From (2.50) the solution with this initial data is thus

$$v(t) = \sqrt{\frac{mg}{k}} \left(\frac{1 - e^{-2t\sqrt{kg/m}}}{1 + e^{-2t\sqrt{kg/m}}} \right). \quad (2.51)$$

This solution may also be expressed as

$$v(t) = \sqrt{\frac{mg}{k}} \tanh(t\sqrt{kg/m}) \quad (2.52)$$

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

Reading Exercise 45 Use (2.51) 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 down is the positive direction)?

Reading Exercise 46 Use (2.51) 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 as

$$\frac{du}{u(1 - u/K)} = r dt. \quad (2.53)$$

This isn’t the only way to separate—the constant r could be taken to the denominator on the left side of (2.53) 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 (2.53) 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 on the right in the expression above must equal 1 for all u , which forces $B - A/K = 0$ and $A = 1$, two equations in unknowns A, B . Solve to find $A = 1, B = 1/K$. Then

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

Integrating both sides in u shows that

$$\int \frac{du}{u(1-u/K)} = \ln|u| - \ln|K-u| + C_2. \quad (2.54)$$

Use this in (2.53) with the antiderivative $rt + C$ for the right side of (2.53) to find

$$\ln|u| - \ln|K-u| = rt + C \quad (2.55)$$

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 (2.55) (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,

$$\left| \frac{u}{K-u} \right| = C e^{rt} \quad (2.56)$$

for some $C > 0$. Equation (2.56) is equivalent to $u/(K-u) = \pm C e^{rt}$ or

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

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

$$u(t) = \frac{K}{1 + e^{-rt}/C}. \quad (2.57)$$

Redefine C as $1/C$ and write

$$u(t) = \frac{K}{1 + Ce^{-rt}}. \quad (2.58)$$

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 (2.58) can be written as

$$u(t) = \frac{Ku_0}{u_0 + e^{-rt}(K - u_0)}. \quad (2.59)$$

See Exercise 2.2.16 for an opportunity to compare the formula (2.59) to some real data.

2.2.6 Exercises

In Exercises 2.2.1 to 2.2.9, find the general solution to the separable ODE, and then find the specific solution with the given initial condition.

Exercise 2.2.1 $u'(t) = u(t) + 3$, $u(0) = 3$

Exercise 2.2.2 $u'(t) = 2u(t) + 4$, $u(0) = 0$

Exercise 2.2.3 $u'(t) = -3u(t) + 3$, $u(0) = 5$

Exercise 2.2.4 $u'(t) = tu(t) + t$, $u(0) = 2$

Exercise 2.2.5 $u'(t) = \sin(t)u(t) + \sin(t)$, $u(0) = 4$

Exercise 2.2.6 $u'(t) = au(t) + b$, $u(0) = u_0$, with a, b, u_0 constants

Exercise 2.2.7 $u'(t) = \sin(t)u(t)$, $u(0) = 1$

Exercise 2.2.8 $u'(t) = t^2u(t)$, $u(1) = 2$

Exercise 2.2.9 $u'(t) = e^t u(t)$, $u(0) = 3$

Exercise 2.2.10 Solve the Newton Cooling DE (2.15) with initial condition $u(0) = u_0$, by using separation of variables.

Exercise 2.2.11 Solve the Hill-Keller ODE (1.3) with $v(0) = 0$ using separation of variables. ■

Exercise 2.2.12 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 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 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

$$v'(t) = g - \frac{kr\mu}{m}v(t). \quad (2.60)$$

- Find a general solution to (2.60). What is the terminal velocity of the object, in terms of k, r, μ , and m ? ■

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 2.2.13 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 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 2.2.14 Evaluate the integral on the left in (2.41) 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 2.2.15 Solve the compartmental salt tank DE (2.17) with initial condition $x(0) = 3$, by using separation of variables, and so demonstrate that (2.18) is correct. ■

Exercise 2.2.16 Table 2.1 contains population data concerning the growth of a species of yeast in a closed vessel (from a classic study [33].) This data may also be found at the book website [6].

Use this data and the solution (2.59) 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? ■

Exercise 2.2.17 This problem is based on the SIMIOODE modeling project [101]. 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 [82]. The goal is to determine whether the ODE (2.26), which was $v'(t) = g - \frac{k}{m}v^2(t)$, provides a good model for an object falling in the presence of quadratic air

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
<hr/>									
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: Time (seconds) and distance (meters) for shuttlecock fall.

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 (2.26), 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 becomes

$$v'(t) = g - \tilde{k}v^2(t). \quad (2.61)$$

It is the variable \tilde{k} we will estimate. Verify that the solution to (2.61) with $v(0) = 0$ is

$$v(t) = \sqrt{\frac{g}{\tilde{k}}} \left(\frac{1 - e^{-2t\sqrt{g\tilde{k}}}}{1 + e^{-2t\sqrt{g\tilde{k}}}} \right) \quad (2.62)$$

or equivalently,

$$v(t) = \sqrt{\frac{g}{\tilde{k}}} \tanh(t\sqrt{g\tilde{k}}). \quad (2.63)$$

- (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 (2.62) or (2.63). 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) Comment: does this model with quadratic air resistance seem reasonable?
(e) A linear model for air resistance is easily obtained in the same manner as (2.26) and leads to $v'(t) = g - \frac{k}{m}v(t)$ or

$$v'(t) = g - \tilde{k}v(t) \quad (2.64)$$

where $\tilde{k} = k/m$ again. Justify this model with reasoning similar to that which led to (2.26).

- (f) Solve (2.64) with initial condition $v(0) = 0$. Then repeat parts (b)-(d) with this model, and find the optimal value for \tilde{k} . Compare to the fit obtained with the quadratic model. Is one convincingly superior?

See the project “Shuttlecocks and Model Selection” in Section 3.6.4 for more on this problem and the issue of which model is “best.”

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 DE's. 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. Newton's Law of Cooling as embodied by the DE (2.15) was analyzed for an ambient temperature A that is constant. 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

$$u'(t) = -k(u(t) - A(t)) \quad (2.65)$$

for some positive constant k . The DE (2.65) is a linear, constant coefficient, nonhomogeneous equation, and can be solved using the integrating factor approach. Let's consider (2.65) 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 DE (2.65) shows that

$$u'(5) = -(0.2)(u(5) - A(5)) = -(0.2)(8 - (10 + 5 \sin(5/2))) \approx 0.998. \quad (2.66)$$

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 $u(t)$ requires only elementary arithmetic

The essential takeaway message here is that the solution to the DE (2.65) that passes through $t = 5, u = 8$ must be tangent to this vector, whose slope we can determine from the right side of ODE (2.65).

Reading Exercise 47

Emulate the computation in (2.66) 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 47 you should find $u'(10) \approx 0.041$ for the solution passing through $t = 10, u = 5$, and $u'(15) \approx 0.940$ 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. This is shown in Figure 2.4 in the right panel. If we pick enough points in the tu plane, perform this computation at each and draw the corresponding vector then the result is a *vector field* that shows the slope of the solution passing through each point.

Of course this operation is highly repetitive and must be done quickly and accurately if it is to be of any value. Happily, computers were invented for jobs just such as this! The left panel of

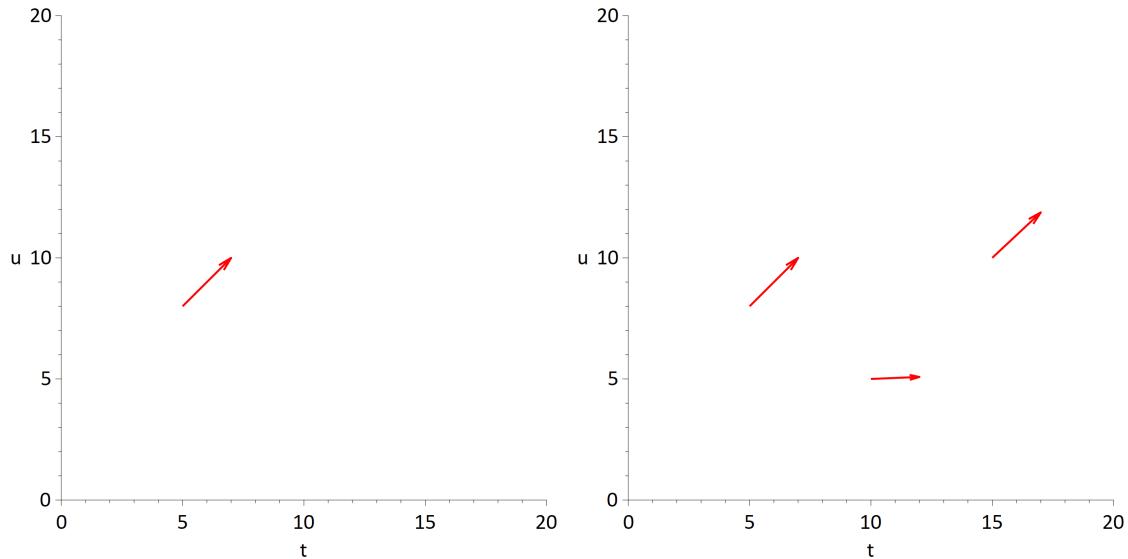


Figure 2.4: Vector to indicate slope of solution to (2.65) at $t = 5, u = 8$ (left) and several points (right).

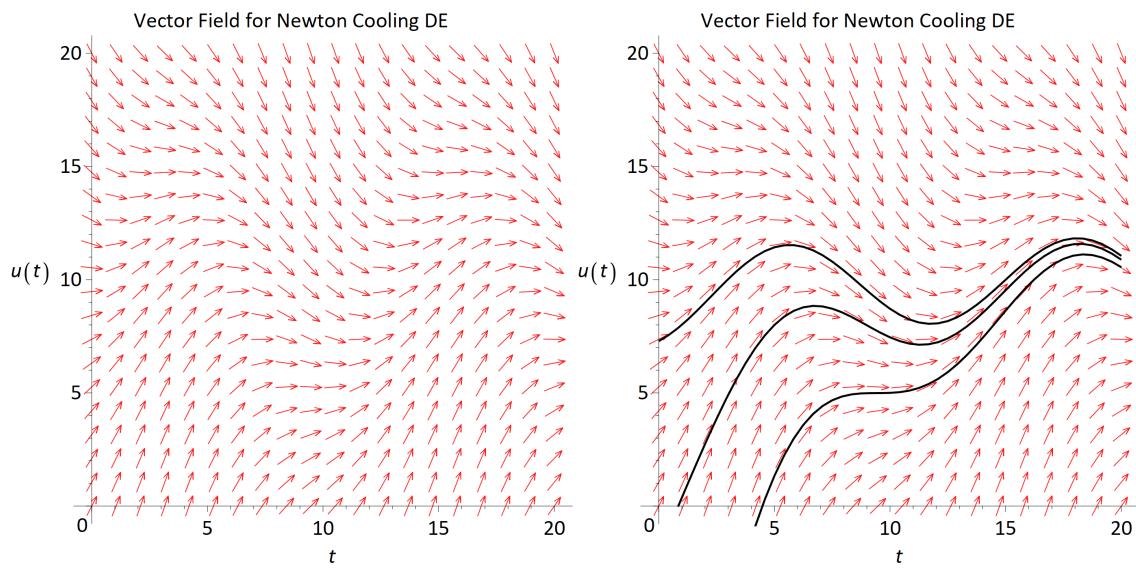


Figure 2.5: Vector field for (2.65) (left) and with superimposed solution curves (right).

Figure 2.5 shows what is obtained if this is 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 left panel in Figure 2.5 illustrates what slope a solution should have as it passes through each chosen point. The length of the vectors doesn't matter, only the slope; the length can be chosen by aesthetic considerations. Solution curves can now be drawn by simply “going with the flow” as illustrated in the right panel of Figure 2.5: through any given point $t = t_0, u = u_0$, sketch a curve that matches the slope of the direction field at all points through which the curve passes. This can be done forward in time ($t > t_0$) or backward in time ($t < t_0$); in the latter case go against the arrows.

Although this process doesn't provide a formula for the solution or prove anything quantitative in nature, it can be highly illuminating and build intuition. For example, Figure 2.65 makes a compelling case that all solutions asymptotically approach the same solution curve, itself some kind of periodic function. This can be proved by noting that a general solution to (2.65) with $k = 0.2$ and $A(t) = 10 + 5 \sin(t/2)$ can be found using the integrating factor approach and is given by

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

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

Reading Exercise 48 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 this point, 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 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. The general first-order ODE is 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 ODE's are often referred to as *time-invariant*, especially in engineering.

■ **Example 2.7** Consider the following ODE's:

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 (equation (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.

Each of the listed ODE's (1)-(3) are autonomous, and all are equations that have been considered in earlier sections. In fact most of the differential equations arising from the applications we've considered so far are autonomous. Only the equation $u' = -k(u - \cos(t))$ is not autonomous, due to the explicit t dependence of the right hand side. ■

Reading Exercise 49 Is the falling object ODE $v'(t) = g - kv^2(t)/m$ (equation (2.26)) autonomous?

Autonomous ODE's 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

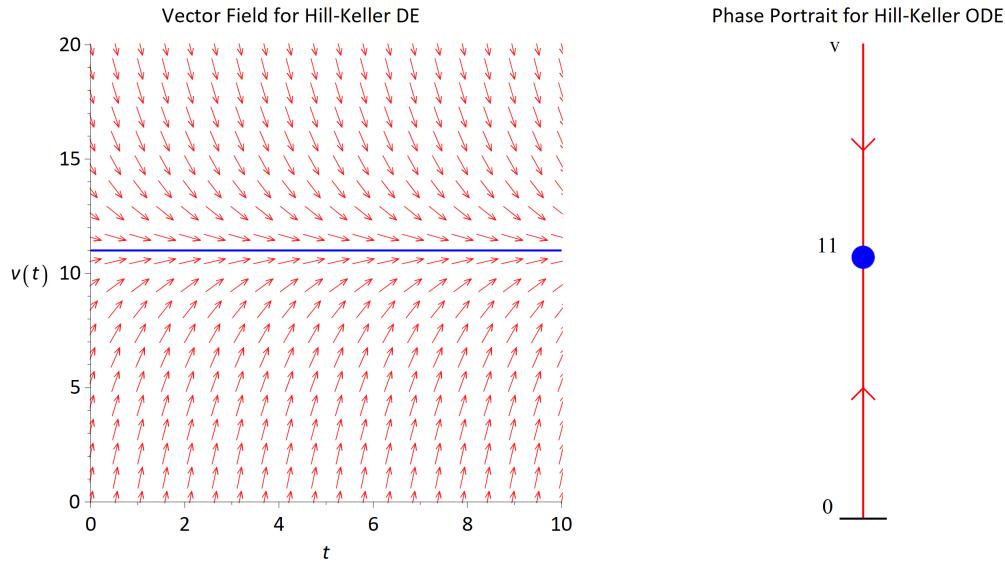


Figure 2.6: Vector field for Hill-Keller ODE $v' = 11 - v$ (left, $k = 1$) and phase portrait (right).

given by $f(u_0)$, 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 , instead $f(t, u)$ with both t and u varying. It also gives the direction field a characteristic appearance.

■ Example 2.8 Consider the ODE $v' = 11 - v$, 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, on the range $0 \leq t \leq 10, 0 \leq v \leq 20$. The slope of the vectors at any given point (t, v) does not depend on t , only v . As a consequence, the vectors on any given horizontal line (lines of constant v coordinate) all have the same slope. It's also obvious that there is a constant solution at $v = 11$, where all vectors have slope $v' = 11 - 11 = 0$. This is indicated by the blue line.

We could economize our efforts in sketching a direction field by eliminating the t axis. This is what has been done 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 blue dot in the panel on the right. The solutions to $v' = 11 - v$ that pass through points with $0 < v < 11$ are all increasing (as we can see in the direction field in the left panel), while those passing through points with $v > 11$ are decreasing. This is indicated with the arrows above and below $v = 11$ in the right panel in Figure 2.6.

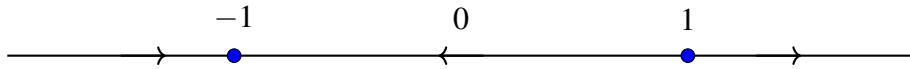
The figure in the right panel in Figure 2.6 is called a *phase portrait* for this ODE. It shows how solutions to this autonomous ODE behave: if $v(t) < 11$ at some time t then $v'(t) = 11 - v(t) > 0$ and the solution is increasing. If $v(t) > 11$ then $v'(t) = 11 - v(t) < 0$ and the solution is decreasing. 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 a DE. It can be done directly, and usually with very little effort. First, let's make a definition.

Definition 2.3.2 — Basic ODE Terminology. A constant solution $u(t) = u^*$ to an autonomous ODE $u' = f(u)$ is called a *fixed point* or *equilibrium solution* or *critical point* for the ODE. Thus if $u(t) = u^*$ is a fixed point for $u' = f(u)$ then $f(u^*) = 0$.

For example, based on Figure 2.6 it appears that $v(t) = 11$ is a fixed point for $v' = 11 - v$, as is easy to verify.

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

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.

1. Find the fixed points (equilibrium solutions) for the ODE, that is, solutions $u(t) = u^*$ that are constant. Such a solution satisfies $u'(t) \equiv 0$, so that $f(u^*) = 0$. Thus find all solutions to $f(u) = 0$. Plot these points as dots on a line, the “ u -axis.”
2. On any interval between a pair of fixed points either $u' > 0$ or $u' < 0$ must hold; the same goes for the regions above the largest fixed point (to ∞) and below the smallest (to $-\infty$) if applicable. In each such interval draw an arrow to indicate which direction solutions move, increasing or decreasing.

■ **Example 2.9** Let us sketch a phase portrait for the autonomous ODE $u'(t) = u^2(t) - 1$, then use this phase portrait to sketch solutions to the ODE with initial conditions $u(0) = -3/2$, $u(0) = 1/2$ and $u(0) = 2$. In this case the ODE is $u' = f(u)$ with $f(u) = u^2 - 1$ and the fixed points are constant solutions u , so $f(u) = 0$, that is, $u^2 - 1 = 0$. These solutions are $u = -1$ and $u = 1$, shown as the blue dots in the u -axis in Figure 2.7. If $u < -1$ then $f(u) > 0$, so solutions increase. If $-1 < u < 1$ then $f(u) < 0$, so solutions decrease. If $u > 1$ then $f(u) > 1$ and solutions increase. This behavior is summarized in the phase portrait of Figure 2.7; the arrows between the fixed points indicate whether solutions increase or decrease in that region.

Reading Exercise 50 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 Point 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. Based on Figure 2.7 if $u = -3/2$ at any time t then the solution $u(t)$ is increasing, and appears to increase toward $u = -1$. If $u = 1/2$ at any time t then the solution $u(t)$ is decreasing, and continues to decrease toward $u = -1$. If $u = 2$ at any time t then the solution $u(t)$ is increasing, and apparently grows without bound. This is illustrated in Figure 2.8. The solutions with 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 $t \approx 0.55$. Note also that the solutions can be projected backward in time t .

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 information about the scaling of the t -axis. ■

Reading Exercise 51 Based on the phase portrait you drew in Reading Exercise 50, 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 4 Example 2.9 illustrates 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

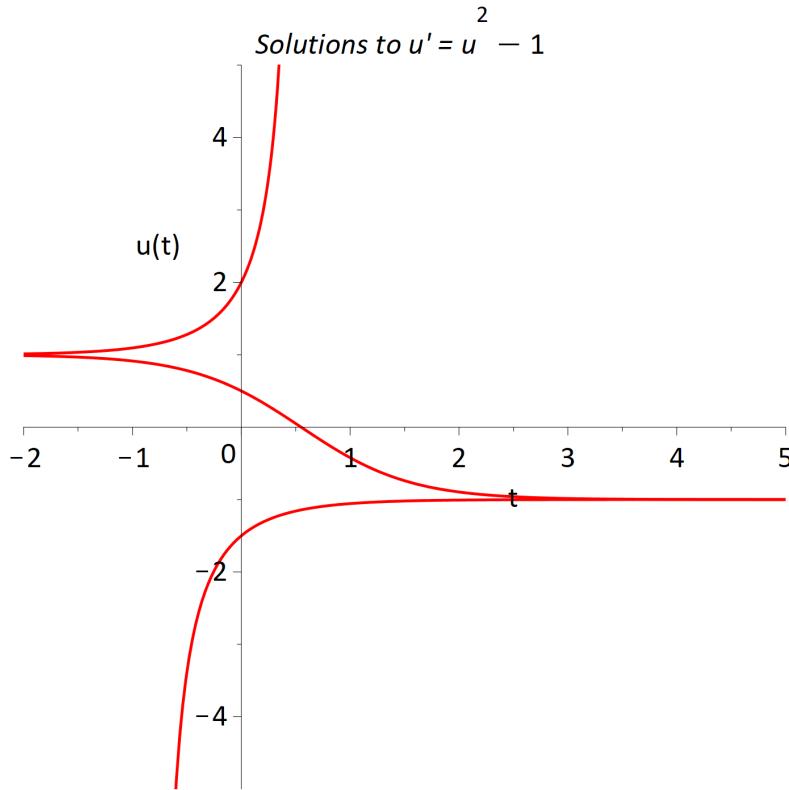


Figure 2.8: Three solutions to $u' = u^2 - 1$ plotted on $-2 \leq t \leq 5$ (right).

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 $t \approx -0.5$. Thus a solution $u(t)$ to an ODE $u' = f(t, u)$ with $u(t_0) = u_0$ is only defined on some interval $t_i < t_0 < t_f$, though frequently $t_i = -\infty$ and/or $t_f = \infty$.

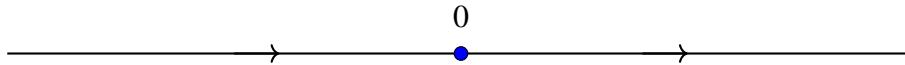
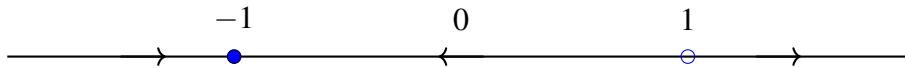
2.3.4 Fixed Points and Stability

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 $t \rightarrow \infty$. 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 said to be *stable*, and in particular, *asymptotically stable*. The fixed point at $u = 1$ is *unstable*.

- More generally, suppose that $u(t) = u^*$ is a fixed point 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 a bit technical.

Definition 2.3.3 — Basic ODE Terminology. Suppose $u(t) = u^*$ is a fixed point (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

Figure 2.9: Phase portrait for the ODE $u' = u^2$.Figure 2.10: Phase portrait version 2 for the ODE $u' = u^2 - 1$.

- The fixed point 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 is *asymptotically stable* if it is stable and for some $\delta > 0$, if $|u_0 - u^*| < \delta$ 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 2.9, in which the arrows point away from the fixed point on both sides, are called *sources*. Is anything else possible? Yes!

■ **Example 2.10** Consider the ODE $u' = u^2$. The only fixed point 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. ■

Reading Exercise 52 Based on your phase portrait from Reading Exercise 50, characterize each fixed point as asymptotically stable or unstable.

A Small Refinement in Phase Portrait Sketching

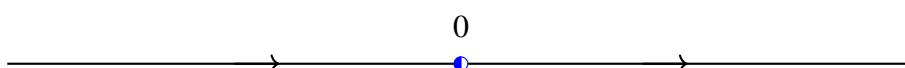
Some authors graphically delineate stable fixed points from unstable fixed points in a phase portrait by drawing a filled dot for a stable fixed point, and a circle for an unstable fixed point. In this case Figure 2.7 for the ODE $u' = u^2 - 1$ would be drawn as shown in Figure 2.10. A further refinement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

Reading Exercise 53 Improve your phase portrait from Reading Exercise 50 to graphically indicated 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 an autonomous ODE's $u' = f(u)$ are either stable or unstable. Suppose $u(t) = u^*$ is a fixed point for $u' = f(u)$, so $f(u^*) = 0$. Suppose also that u^* is “isolated”, in that there is an interval $I = (u^* - \delta, u^* + \delta)$ for some $\delta > 0$ that contains no other fixed points. Based on our work sketching phase portraits, it's easy to see that the stability of u^* can be determined by

Figure 2.11: Phase portrait version 2 for the ODE $u' = u^2$.

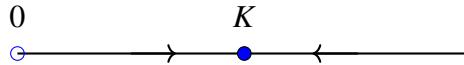


Figure 2.12: Phase portrait for the logistic equation $u' = ru(1 - u/K)$.

looking at the value 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^*) for some $\delta > 0$ 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 $u \in (u^* - \delta, u^* + \delta)$ (with $u \neq u^*$) for some $\delta > 0$ 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 $u^* < u < u^* + \delta$ for some $\delta > 0$ (to the right of u^*) then u^* is unstable.

And as remarked above, semi-stable fixed points are really 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. This makes it even easier to test the stability of most fixed points. In particular,

- **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/minima in Calculus 1. The direct approach of examining f on each side of u^* would be applicable, though.

Reading Exercise 54 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 2.11** 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. This 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 this 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 below, the fixed point at $u = 0$ will be unstable and $u = K$ will be unstable, so we'll just draw appropriate dots right now.

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, $u < 0$ 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 stable, in fact 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 this fixed point is unstable. But $f'(K) = r - 2r < 0$, so this fixed point is stable. ■

Take note: we just figured out how solutions to the logistic equation behave and all we did was some 8th grade algebra! Compare this to the work that was necessary to solve the logistic ODE in Section 2.2.5.

Reading Exercise 55 Based on the phase portrait in Figure 2.12, sketch the graph of solutions $u(t)$ to the logistic equation with $u(0) = K/2$, and with $u(0) = 2K$.

Reading Exercise 56 What's wrong with trying 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 57 Consider the ultimately simple 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 58 Sketch a phase portrait to show that all solutions to the salt tank problem quantified by the ODE (2.17) approach the fixed point $x(t) = x^* = 20$ kg (no need to solve the DE!) Of course you can confine your attention to $x \geq 0$, although it won't change the conclusion.

■ **Example 2.12** 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 above can be multiplied by any positive constant, or indeed, by any positive function of u , and the ODE still has the desired behavior. ■

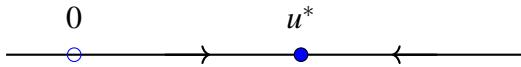
2.3.6 Bifurcations

In this section we'll take a brief look at the notion of *bifurcations*, a topic we'll explore again later in the text.

Bifurcations in the Harvested Logistic Equation

Very frequently the differential equation of interest contains unspecified parameters. For example, the Hill-Keller equation (1.3) contain parameters k and P , while the harvested logistic equation (1.12), reproduced here for convenience,

$$u'(t) = ru(t)(1 - u(t)/K) - hu(t) \tag{2.67}$$

Figure 2.13: Phase portrait for $u' = ru(1 - u/K) - h$, $h < r$.Figure 2.14: Phase portrait for $u' = ru(1 - u/K) - h$, $h > r$.

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 (2.67) 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 a phase portrait for (2.67). The fixed points are solutions to

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

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

$$u^* = K(1 - h/r). \quad (2.68)$$

Only the physically relevant region $u \geq 0$ is of interest. One thing is clear from (2.68): If $h < r$ then $u^* > 0$ and there is a physically meaningful solution in which the population has a constant positive value. That is, $u(t) = u^* > 0$ is a solution to (2.67). 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 of course solutions to the ODE may change behavior drastically. 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 (2.67) after simplifying.

- **$h < r$:** In this case there are two fixed points, $u = 0$. and from (2.68) $u = u^* > 0$. In the interval $0 < u < u^*$ solution directions can be determined by computing $f(u^*/2) = \frac{K(h-r)^2}{4r} > 0$ (after a bit of algebra). Solutions in this region increase. It's also easy to see that if $u > u^*$ then $f(u) < 0$, say by computing $f(2u^*) = -\frac{2K(h-r)^2}{r} < 0$. The resulting phase portrait is shown in Figure 2.13. 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 accord with the phase portrait.
- **$r < h$:** In this case $u^* < 0$ and this fixed point 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 phase portrait now looks like that shown in Figure 2.14. 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 $f'(0) = r - h < 0$ since $r < h$, confirming that the origin is stable. This is 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.
- **$h = r$:** See Reading Exercise 59.

Bifurcation Diagram for Harvested Logistic Equation

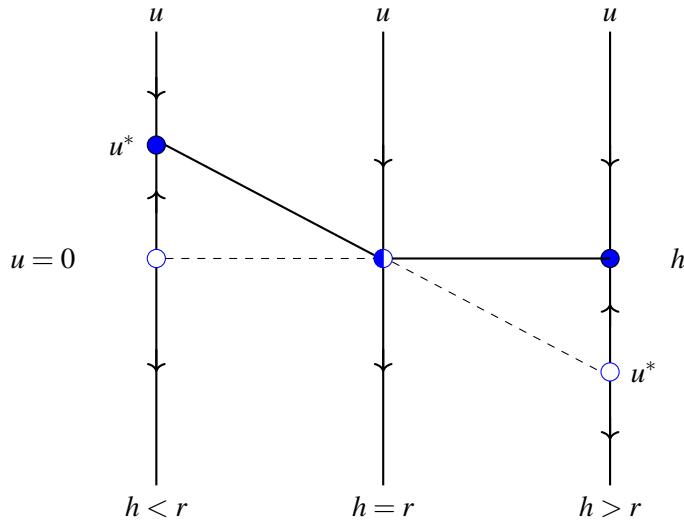


Figure 2.15: Bifurcation diagram for the harvested logistic equation (2.67).

Reading Exercise 59 Consider the harvested logistic equation (2.68) 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 let's us make a rather strong and physically relevant conclusion, without solving the harvested logistic equation (2.68): If $h < r$ (the harvesting rate h is less than the growth rate r) then there is a stable equilibrium point $u^* = K(1 - h/r)$ for the population, which all solutions asymptotically approach. If $h > r$ then there is no positive equilibrium population and 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 (2.67) 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, but 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 59. 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 $h > r$.

2.3.7 Exercises

In Exercises 2.3.1 to 2.3.4, for the given ODE and points $t = t_0, u = u_0$, the slope $u'(t_0)$ of the solution that passes through that point. Then plot appropriate vectors on a pair of t - u axes to form a

(crude) direction field; make the vectors fairly short, perhaps length $1/4$ or so.

Exercise 2.3.1 $u'(t) = u(t) - 2t$, for (t_0, u_0) pairs $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. ▀

Exercise 2.3.2 $u'(t) = u^2(t) + t + 1$, for (t_0, u_0) pairs $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. ▀

Exercise 2.3.3 $u'(t) = -u(t)$, for (t_0, u_0) pairs $(0, 1)$, $(0, 2)$, $(1, 1)$, and $(1, 3)$. ▀

Exercise 2.3.4 $u'(t) = -1/u(t)$, for (t_0, u_0) pairs $(0, 1)$, $(0, 2)$, $(1, 1)$, and $(1, 3)$. ▀

In Exercises 2.3.5 to 2.3.12, use whatever technology you have available to sketch a direction field for the given ODE on the given range. If the ODE is autonomous, visually identify the equilibrium solutions, if any.

Exercise 2.3.5 $u'(t) = u(t) - 2t$, $0 \leq t \leq 2$, $0 \leq u \leq 2$. ▀

Exercise 2.3.6 $u'(t) = u^2(t) + t + 1$, $-2 \leq t \leq 2$, $-2 \leq u \leq 2$. ▀

Exercise 2.3.7 $u'(t) = -u(t)$, $-2 \leq t \leq 2$, $-2 \leq u \leq 2$. ▀

Exercise 2.3.8 $u'(t) = -1/u(t)$, $-2 \leq t \leq 2$, $-2 \leq u \leq 2$. ▀

Exercise 2.3.9 $u'(t) = u(t)(u(t) - 3)$, $-2 \leq t \leq 5$, $-2 \leq u \leq 5$. ▀

Exercise 2.3.10 $u'(t) = (u(t) - 1)(u(t) + 1)$, $-2 \leq t \leq 5$, $-2 \leq u \leq 5$. ▀

Exercise 2.3.11 $u'(t) = t \sin(u) - t^2/4$, $-2 \leq t \leq 5$, $-2 \leq u \leq 5$. ▀

Exercise 2.3.12 $u'(t) = \cos(u+t)$, $-2 \leq t \leq 5$, $-2 \leq u \leq 5$. ▀

In Exercises 2.3.13 to 2.3.20 sketch, by hand, a phase portrait for the given autonomous ODE, following the procedure of Examples 2.9, 2.10, and 2.11, and classify each fixed point as asymptotically stable or unstable. Use this to sketch solutions for the given initial conditions on pair of tu axes with a reasonable range for u .

Exercise 2.3.13 $u'(t) = -u(t)$, sketch solutions with $u(0) = 2$ and $u(0) = -2$. ▀

Exercise 2.3.14 $v'(t) = 11 - 2v(t)$, sketch solutions with $v(0) = 0$ and $v(0) = 15$. ▀

Exercise 2.3.15 $v'(t) = 11 - kv(t)$ (k a positive constant), sketch solutions with $v(0) = 0$ and $v(0) = 15/k$. ▀

Exercise 2.3.16 $u'(t) = -(u(t) - 1)(u(t) - 3)$, sketch solutions with $u(0) = 1/2$, $u(0) = 2$, $u(0) = 4$. ■

Exercise 2.3.17 $u'(t) = u(t)(1 - u(t)) - u(t)/10$ (the harvested logistic equation (1.12) with $r = 1, K = 1, h = 1/10$), sketch solutions with $u(0) = 1/2$, $u(0) = 3/2$. Note only $u \geq 0$ makes physical sense here. What is the long-term fate of the species? ■

Exercise 2.3.18 $u'(t) = u(t)(1 - u(t)) - 2u(t)$ (the harvested logistic equation (1.12) with $r = 1, K = 1, h = 2$), sketch solutions with $u(0) = 1/2$, $u(0) = 3/2$. Note only $u \geq 0$ makes physical sense here. What is the long-term fate of the species? ■

Exercise 2.3.19 $u'(t) = rc_1 - ru(t)/V$ (the conservation law ODE (1.5), $r, c_1, V > 0$). Recall that this model is only appropriate for $u \geq 0$. Label the fixed point(s) in terms of r, c_1, V , and sketch solutions for which $u(0) = 0$ and $u(0) = 2c_1V$. ■

Exercise 2.3.20 $v'(t) = g - kv^2(t)/m$ (the falling body ODE (2.26), $m, g, k > 0$). Recall that this model is only appropriate for $v \geq 0$. Label the fixed point(s) in terms of m, g, k , and sketch solutions for which $v(0) = 0$ and $v(0) = 2\sqrt{mg/k}$. ■

In Exercises 2.3.21 to 2.3.24, make up an autonomous ODE $u' = f(u)$ (by finding a suitable function f) that has the desired fixed points with the indicated stability. Hint: review Example 2.12.

Exercise 2.3.21 Fixed points $u = 1$ (stable) and $u = 3$ (unstable). ■

Exercise 2.3.22 Fixed points $u = -3$ (stable), $u = 0$ (unstable), $u = 4$ (stable) and $u = 5$ (unstable). ■

Exercise 2.3.23 Fixed points $u = 1$ (semistable) and $u = 3$ (stable). Hint: include a term $(u - 1)^2$ in $f(u)$. ■

Exercise 2.3.24 Fix 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*. ■

In Exercises 2.3.25 to 2.3.27, an ODE that depends on a parameter h is given (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.

Exercise 2.3.25 $u' = hu - u^2$. ■

Exercise 2.3.26 $u' = hu - u^3$. The resulting bifurcation diagram illustrates a *pitchfork bifurcation*.. ■

Exercise 2.3.27 $u' = ru(1 - u/K) - h$, with $r = 1$ and $K = 1$. This is similar to the harvested logistic equation (2.67), but here the harvest rate is a constant h , instead of hu . Assume $h > 0$,

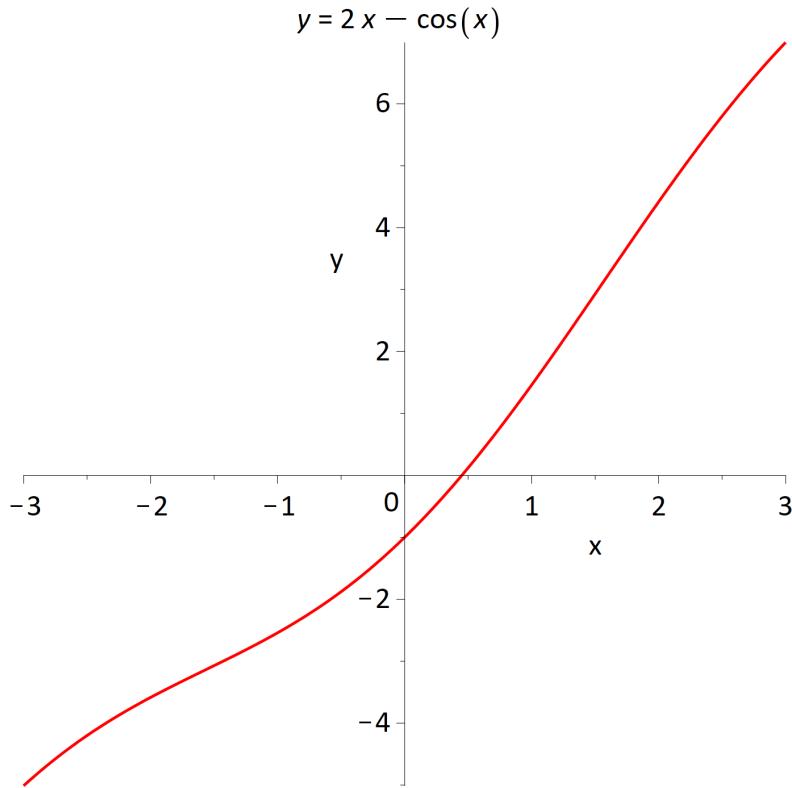


Figure 2.16: Graph of $f(x) = 2x - \cos(x)$.

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

Forget about differential equations for a moment. Instead, let's go back to Calculus 1, by considering the equation

$$2x - \cos(x) = 0. \quad (2.69)$$

The goal is to find a real number x that satisfies (2.69). If you try to solve (2.69) by using elementary algebraic operations, e.g., the basic four 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 (2.69) 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 $x = 1$, by the Intermediate Value Theorem $f(x)$ must be zero somewhere between $x = 0$ and $x = 1$. This is true even though we can't write down the solution. And as asserted above, this solution is unique: 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 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 it 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 60 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 ODE's?

Up to this point we've been talking about solutions to ODE's, yet 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 this works. But later in the text there will appear differential equations in which we wish to discuss 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 some region in the tu plane. A function $u(t)$ is a *solution* to $u'(t) = f(t, u(t))$ if

- $u(t)$ is defined and differentiable at every point in some interval $a < t < b$ (and hence continuous);
- The graph of $u(t)$, $a < t < b$, lies in the domain of f ;
- $u'(t) = f(t, u(t))$ for every point t in $a < t < b$.

As noted above, in every example so far it's been easy to see that solutions fit the above definition. But let's take a look at an example with a slight subtlety.

■ **Example 2.13** Consider the ODE

$$u'(t) = u^2(t) \tag{2.70}$$

with the initial condition $u(0) = 1$. Solving via separation of variables shows that

$$u(t) = \frac{1}{1-t}. \tag{2.71}$$

The function $u(t)$ is plotted in Figure 2.17, on the range $-3 < t < 3$. The graph of the solution has a vertical asymptote at $t = 1$, as the formula (2.71) also makes obvious. It might be tempting to conclude that (2.70) with $u(0) = 1$ has a solution for all $t \neq 1$, but this does not 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 only be extended forward in t up to, but not including, $t = 1$, but 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. ■

The issue raised in Example 2.13 also appeared in Example 2.9, as noted in Remark 4.

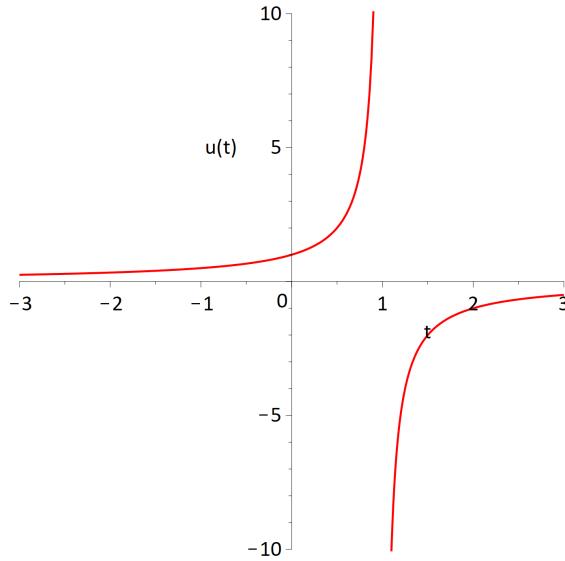


Figure 2.17: “Solution” to $u'(t) = u^2(t)$ with $u(0) = 1$.

Reading Exercise 61 Go back and look at Example 2.9, and in particular 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.

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 61 you should find 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.

In Chapter 5 we’ll encounter “solutions” to ODE’s 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 ODE’s

Under what circumstances can it 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

$$u'(t) = t \cos(u(t)) - \sin(t) \quad (2.72)$$

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 (2.72), shown in the left panel of Figure 2.18, ought to convince you that a solution should exist, for the simple reason that 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.

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,

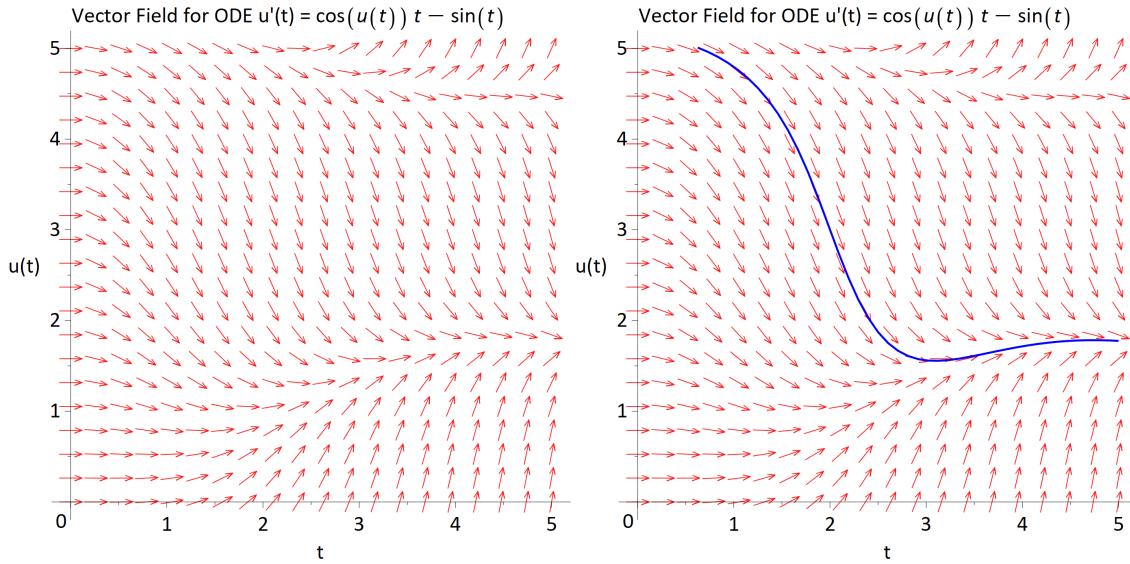


Figure 2.18: Direction field for (2.72) (left) and with solution curve through $t = 2, u = 3$ (right).

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. However, there are a couple of restrictions on the nature of the ODE. 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 (in more advanced texts weaker conditions are permitted). This is summarized in the following theorem.

Theorem 2.4.1 — Existence and Uniqueness of Solutions to ODE's. Let R denote a rectangle $a < t < b, c < u < d$ 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 $\delta_1, \delta_2 > 0$.

A fairly accessible proof of Theorem 2.4.1 can be found in [27].

■ **Example 2.14** To illustrate Theorem 2.4.1, consider the right panel of Figure 2.18, where a solution to (2.72) 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 $0 < t < 5, 0 < u < t$ 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.

In fact by using more advanced techniques it can be shown in this case 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 2.4.1, and so it will be possible to assert that the solutions 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.

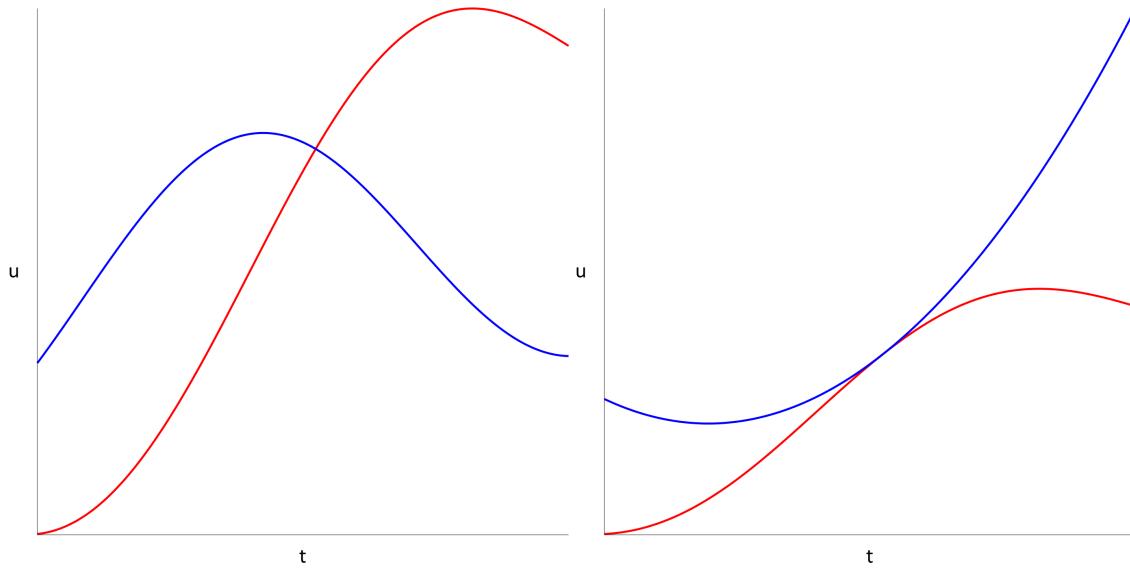


Figure 2.19: “Solution” curves crossing or, more subtly, “touching” with the same slope (right).

Reading Exercise 62 The ODE

$$v'(t) = g - kv^r(t)/m \quad (2.73)$$

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 (2.26) was the special case $r = 2$, while the case $r = 1$ was explored in Exercise 2.2.17. The ODE (2.73) 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 (2.73) with any initial data $v(0) = v_0 > 0$.

Implications for Sketching Solutions

Remark 5 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 situation in Figure 2.19 is impossible, if both curves 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, $f(t_0, u_0)$. It's the situation in the right panel that 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 2.15** 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

$$u' = f(u)$$

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 $u(t) \equiv 0$ (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 $\frac{\partial f}{\partial u} = \frac{2|u|^{2/3}}{3u}$ is not continuous or even defined at $u = 0$.

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

In Exercises 2.4.1 to 2.4.4 use Theorem 2.4.1 to show the 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.

Exercise 2.4.1 $u'(t) = u(t) + 3$, $u(0) = 3$

Exercise 2.4.2 $u'(t) = -u^2(t) + \sin(t)$, $u(1) = 4$

Exercise 2.4.3 $u'(t) = 1/u(t)$, $u(2) = 2$

Exercise 2.4.4 $u'(t) = ru(t)(1 - u(t)/K)$, $u(t_0) = u_0$, the logistic equation.

Exercise 2.4.5 Consider a very general form of Newton's Law of Cooling given by

$$u'(t) = h(u(t) - A) \quad (2.74)$$

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 (2.15), is the special case $h(x) = -kx$.

- (a) Use Theorem 2.4.1 to show that (2.74) 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 $A \leq u_1(0) < u_2(0)$? ■

In Exercises 2.4.6 to 2.4.9 find the solution to the given initial value problem and then use it to find the maximum domain of the solution.

Exercise 2.4.6 $u'(t) = 2 - u(t)$, $u(0) = 2$.

Exercise 2.4.7 $u'(t) = 1 + u^2(t)$, $u(0) = 0$.

Exercise 2.4.8 $u'(t) = e^{u(t)}$, $u(0) = 0$.

Exercise 2.4.9 $u'(t) = 1/u(t)$, $u(0) = 3$

2.5 Modeling Projects

In this section we offer three modeling opportunities, based on projects from the SIMIODE website [7]. 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,” [97].

Borrowing money to pay for things is one of the inescapable privileges and curses of adulthood. A mortgage is, quite likely, the most money any of us will ever borrow. 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 such a home loan are also 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 one can 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 put down 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 you’d owe at the end of the first month, except that at that time you will make a payment. The payments 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 principle.

This process repeats in the next month, with a balance of \$248,898.55 in place of \$250,000. Your payment, \$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)(248898.54) = \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

$$p_k = \left(1 + \frac{r}{12}\right) p_{k-1} - b. \quad (2.75)$$

Modeling Exercise 1 Equation (2.75) is called a *difference equation*. Use (2.75) to compute p_2, p_3, \dots, p_{180} . Of course, you'll need technology; a spreadsheet would do nicely. The code in Maple looks like

```
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;
od
```

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. You should verify this.

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 (2.75) can be replaced by

$$p_k = \left(1 + \frac{r}{360}\right) p_{k-1} - \frac{b}{30}. \quad (2.76)$$

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

$$p_k = \left(1 + \frac{r}{n}\right) p_{k-1} - \frac{12b}{n}. \quad (2.77)$$

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.

In the limit that n approaches infinity this models the situation in which interest is compounded “continuously,” and payments are made at a constant continual rate per unit time. What does (2.77) becomes in this case?

Modeling Exercise 2 Show that (2.77) can be written as

$$\frac{p_k - p_{k-1}}{1/n} = rp_{k-1} - 12b. \quad (2.78)$$

Modeling Exercise 3 The quantity p_k is the balance of the loan at time $t = k/n$ years, since each iteration of (2.77) 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 (2.78) 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

$$\frac{p(t) - p(t - \Delta t)}{\Delta t} = rp(t - \Delta t) - 12b \quad (2.79)$$

where $t = k/n$ and $\Delta t = 1/n$.

Modeling Exercise 4 Argue that in the limit $n \rightarrow \infty$ equation (2.79) becomes

$$p'(t) = rp(t) - 12b \quad (2.80)$$

Assume that p is a continuous function, so $p(t - \Delta t) \rightarrow p(t)$ as $\Delta t \rightarrow 0$.

Equation (2.80) 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 Sketch a phase portrait for (2.80) 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 6 Solve (2.80) 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 10 years, and compare to the balances when compounding is done monthly, \$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 7 Solve $p(t) = 0$ for t , to find that time $t = T$ when you pay off the loan.

Modeling Exercise 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 (2.80). The total interest paid should therefore be the accumulation or “sum”

$$\int_0^T rp(t) dt. \quad (2.81)$$

where T , the time the loan is paid off, is from the last exercise. Show that the expression in (2.81) has the dimension V (“value”), or units of dollars. Compute the integral in this case.

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 9 Show that the solution to the ODE (2.80) with initial condition $p(0) = p_0$ and with r, b , and p_0 undefined, is given by

$$p(t) = p_0 e^{rt} + \frac{12b}{r} (1 - e^{rt}). \quad (2.82)$$

Modeling Exercise 10 Suppose you take out a 30 year mortgage instead of 15 years. 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 (2.82), 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 11 Use the procedure of Exercise 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” [26].

Chemists often use differential equations to model chemical reactions. The rate at which a reaction proceeds is often 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.” [8] In this project we consider reactions involving a reactant “A” in which the reaction rate is dependent upon $[A]$, the concentration (perhaps in units of moles per liter) of A 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,

$$\frac{d[A]}{dt} = -k[A]^m, \quad \text{with} \quad [A](0) = [A_0]. \quad (2.83)$$

where $k > 0$ is the *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 such data. In Section 3.5 of the next chapter we’ll consider even more sophisticated methods for estimating these parameters.

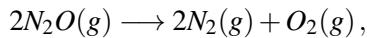
The Order of a Reaction

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 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 [110, 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

$$\frac{dy}{dt} = -ky^0 = -k. \quad (2.84)$$

This would come with an initial condition such as $y(0) = y_0$.

Modeling Exercise 1 Show that the solution to (2.84) is given by $y(t) = y_0 - kt$.

Zeroth-order reactions are fairly self-evident from data, for a plot of $y(t)$ vs. t reveals a linear function starting at $y = y_0$ with a negative slope $-k$. 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 first-order reactions in general,

$$\frac{dy}{dt} = -ky^1 = -ky. \quad (2.85)$$

with initial condition $y(0) = y_0$.

Modeling Exercise 2 Use the separation of variables technique to show that

$$\ln(y) = -kt + \ln(y(0)). \quad (2.86)$$

You will not need to actually solve for $y(t)$ explicitly in the present setting.

The next step in separation of variables would be to solve (2.86) explicitly for $y = y(t)$. But often a chemist might be 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 (2.86) as it stands. Chemists refer to (2.86) as the *integrated form* of the rate law. However, in mathematics it is traditional to push on to a solution that reads “ $y(t)$ is.” Exponentiating both sides of (2.86) shows that

$$y = y(t) = y(0)e^{-kt}. \quad (2.87)$$

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, along with 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 while it rids the cut of germs. 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 . This means for $[H_2O_2]$ moles/L of hydrogen peroxide:

$$\frac{d[H_2O_2]}{dt} = -k \cdot [H_2O_2]^m, \quad (2.88)$$

for some number m . 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 .

In Table 2.3 are time-concentration data pairs for an experiment concerning the decomposition of H_2O_2 .

Modeling Exercise 3 On the book website [6] you will find Matlab, Maple, and Mathematica files that contain this data. In this exercise we will use the various software's capabilities for fitting lines and curves to data to estimate parameters like k and m in (2.88), but we will take a more detailed look at how this can be done in Section 3.5.

- (a) 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`.
- (b) From (2.86) it can be seen that a first-order reaction will produce a linear relationship between $\ln(y)$ and t . Confirm that this is a first-order reaction by plotting $\ln[H_2O_2]$ vs t .
- (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.

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: Collected data [110, p. 682] on the reaction $2H_2O_2(g) \longrightarrow 2H_2O + O_2(g)$.

- (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 are a more equal footing, rather than force 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

$$\frac{dy}{dt} = -ky^2 \quad \text{with} \quad y(0) = y_0. \quad (2.89)$$

Modeling Exercise 4 Use the separation of variables technique to show that

$$\frac{1}{y} = kt + \frac{1}{y(0)}. \quad (2.90)$$

You do not need to actually solve for $y(t)$ explicitly.

Equation (4) can be solved explicitly for $y = y(t)$, and we 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 (2.90). If desired, an explicit form for $y(t)$ can be found by inverting both sides in (2.90),

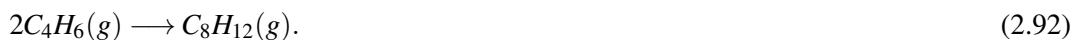
$$y = y(t) = \frac{1}{kt + \frac{1}{y(0)}} = \frac{y(0)}{y(0)kt + 1}. \quad (2.91)$$

Modeling Exercise 5 Revisit the decomposition of hydrogen peroxide in Table 2.3 and show 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.

Decomposition of C_4H_6

Consider the data in Table 2.5 for the reaction of C_4H_6 , butadiene, to form its dimer, a chemical structure formed from two similar sub-units. The reaction is



Modeling Exercise 6 On the book website [6] you will find Matlab, Maple, and Mathematica files that contain the data in Table 2.5 and support the analysis your are asked to do below.

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 obtain possibly a linear plot to confirm the order of the reaction, and a complete solution for a fully developed model.

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 [110, p. 654] for the reaction of C_4H_6 , butadiene, to form its dimer.

- (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 $m = 0, 1, 2$ possibilities.

Decomposition of N_2O_5

Consider the data in Table 2.6 for the reaction describing the decomposition of N_2O_5 , dinitrogen pentoxide:



Modeling Exercise 7 On the book website [6] you will find Matlab, Maple, and Mathematica 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, determining the order and the parameters. Plot the data and the model, being sure to defend your model (and explain what the order is and what the order is not vis-à-vis $m = 0, 1, 2$ orders).

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, quickly rendering them harmless. But is this realistic? See [3, 4] for some actual experiments and footage.

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 [51] for the decomposition of N_2O_5

The Model

The situation can be modeled exactly as 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 (2.26), reproduced here,

$$v'(t) = g - \frac{k}{m} v^2(t). \quad (2.94)$$

Assume that the bullet enters the water at time $t = 0$. The main difference here is that $v(0) = v_0 > 0$ (rather than $v(0) = 0$).

Analysis

Modeling Exercise 1 Sketch a phase portrait for (2.94); you can 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 2.

Modeling Exercise 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 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 authors know the answer. We assume no responsibility for accidents.

Modeling Exercise 3 Show that under the assumptions above the general solution

$$v(t) = \sqrt{\frac{mg}{k}} \left(\frac{1 + Ce^{-2t\sqrt{kg/m}}}{1 - Ce^{-2t\sqrt{kg/m}}} \right). \quad (2.95)$$

to (2.26)/(2.94) that was previously derived in Section 2.2.4 can be expressed as

$$v(t) = \frac{1 + Ce^{-19.62t}}{1 - Ce^{-19.62t}}. \quad (2.96)$$

Modeling Exercise 4 Adjust the constant C in the general solution (2.96) to obtain $v(0) = 1000$ meters per second.

Modeling Exercise 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 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 at $y = 0$) until $t = t^*$. How deep must our heroine be to escape serious injury?

Modeling Exercise 6 Suppose our estimate of the terminal velocity of the bullet is wrong—what if it descends at a terminal velocity of 2 meters per second—how much difference does that make?

Further SIMIODE Projects

Here would go a listing and brief descriptions of other suitable SIMIODE modeling projects for this chapter.

3. Numerical Methods for ODE's

Many ODE's 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 ODE's 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 methods that are implemented in a number of software packages. These ODE solvers 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 3.4 discussing "error control" and "adaptive stepsizing," an essential part of any good ODE solver. The goal is not to make the reader an expert in numerical ODE methods or to write code, but rather to 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 3.5 is devoted to some basic ideas and examples concerning this task, known as "parameter estimation."

3.1 The Need for Numerics

In [50] the authors consider the logistic equation (1.10) as a model for a population under the general circumstance that the 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)

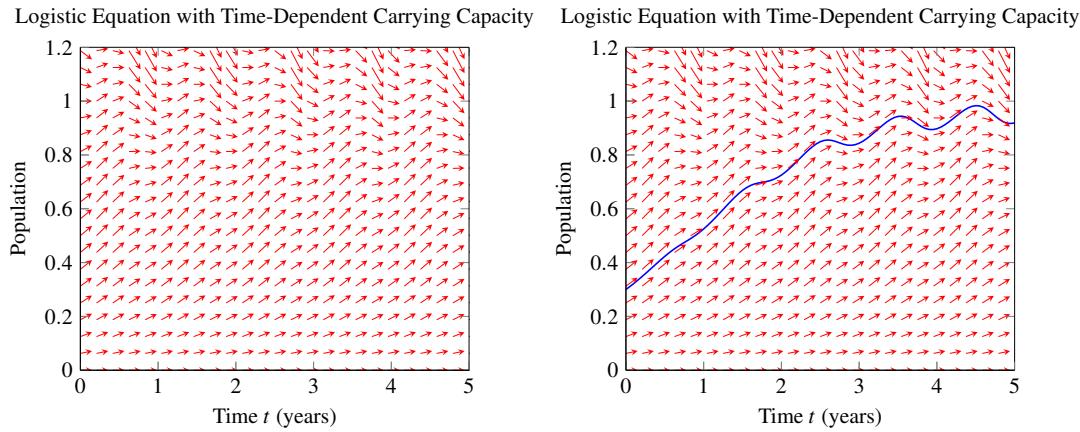


Figure 3.1: Direction field for (3.2) (left) and with solution curve for $u(0) = 0.3$ superimposed (right).

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 allows us to solve this ODE, except in the trivial case that $u(0) = 0$. And since the equation is not separable, it cannot be autonomous, so even the method of phase portraits from Section 2.3 is not applicable.

However, a direction field for (3.2) can be sketched, shown in the left panel of Figure 3.1. It's easy to visualize the solution with, for example, initial condition $u(0) = 0.3$ in the direction field on the left, and this solution curve is traced in the right panel. The hypotheses of the Existence-Uniqueness Theorem 2.4.1 are also straightforward to verify for the ODE (3.2), so as is graphically clear in Figure 3.1, a unique solution with $u(0) = 0.3$ exists. What if the value of $u(2)$ is desired? How can this kind of quantitative information be obtained without drawing pictures? That is the subject of this section.

Reading Exercise 63 Verify that the Existence-Uniqueness Theorem 2.4.1 applies to (3.2) with $u(0) = 0.3$.

3.2 Euler's Method

The Tangent Line Approximation

Recall an elementary fact from Calculus 1, the “tangent line approximation,” or more generally *linearization*. Suppose a function $u(t)$ is 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 red curve. Assume u is continuously differentiable on (a, b) . The tangent line to the graph of u is given by $y = L(t)$, graphed as the dashed blue line in Figure 3.2, where $L(t)$ is the function

$$L(t) = u(t^*) + u'(t^*)(t - t^*). \quad (3.3)$$

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^* .

Consider the choice $t = t^* + h$ as illustrated in Figure 3.2. From (3.3)

$$L(t^* + h) = u(t^*) + u'(t^*)h. \quad (3.4)$$

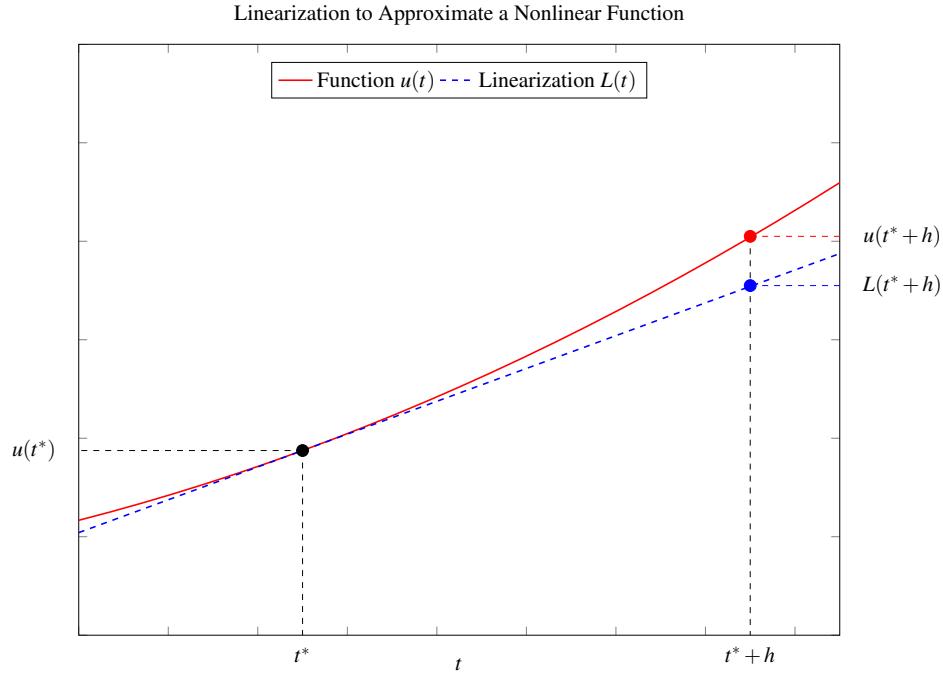


Figure 3.2: Graph of $y = u(t)$ (red) and the linearization $y = L(t)$ (blue) of u at $t = t^*$, with points $(t^*, u(t^*))$ (black dot), $(t^* + h, u(t^* + h))$ (red dot) and $(t^* + h, L(t^* + h))$ (blue dot).

If h is sufficiently close to zero then to good approximation $u(t^* + h) \approx L(t^* + h)$ and so

$$u(t^* + h) \approx u(t^*) + u'(t^*)h. \quad (3.5)$$

Equation (3.5) allows us to approximate $u(t^* + h)$ using knowledge of $u(t^*), h$, and $u'(t^*)$, and is 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 64 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 65 With $u(t)$ and t^* as in Reading Exercise 64, compute and simplify the expression $u(t^* + h) - L(t^* + h)$ (leave h undefined). How does this quantity depend on h ?

3.2.1 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. However, rather than start with (3.2) as a first example, consider instead an ODE

$$u'(t) = u(t) - 3t^2 \quad (3.6)$$

with initial condition $u(0) = 1$. Equation (3.6) can be solved analytically using the integrating factor technique of Section 2.1. The solution is

$$u(t) = 3t^2 + 6t + 6 - 5e^t. \quad (3.7)$$

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 in steps of size 0.25. 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 *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 $t^* = t_1 = 0.25$, while $u_1 = 1.25 \approx u(t_1)$ 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 $f(0.25, 1.25) \approx 1.0625$ equation (3.10) yields $u_2 \approx 1.5156$. The true value is $u(0.5) \approx 1.5064$.

Reading Exercise 66 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 and the Euler Method iterates for (3.6) based on linear extrapolation with step size $h = 0.25$, both superimposed on the direction field for the ODE (3.6). 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, the next iterate is erroneous. This is clearly the case as here in the step from $t_2 = 0.5$ to $t_3 = 0.75$, and on to $t_4 = 1$. One might think of the true solution, shown in red, as the result of taking “infinitesimal” steps of size h forward in time, and so the true solution tracks the direction field perfectly at all points.

Euler's Method In General

Euler's Method in general 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

$$u_{k+1} = u_k + hf(t_k, u_k) \quad (3.11)$$

for some chosen value of N . This “marches” the solution out to time $t_N = t_0 + Nh$. Pseudocode for Euler's Method is show in Figure 3.4.

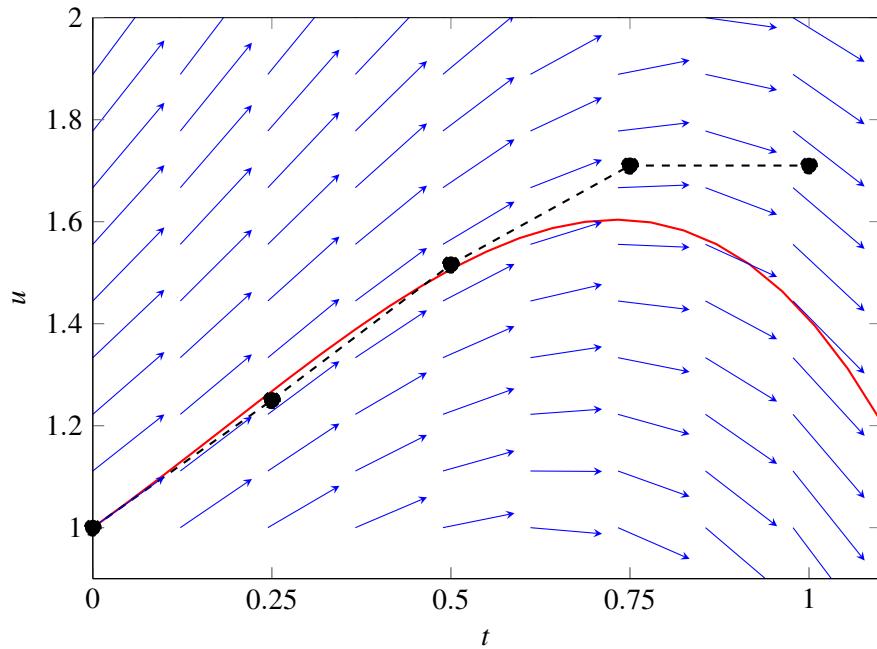


Figure 3.3: True solution $u(t)$ to (3.6) with $u(0) = 1$ (red), Euler iterates with step size $h = 0.25$ (black) and direction field for (3.6) (blue).

```

algorithm Euler's Method:
input initial time t0,
      initial value u0,
      step size h,
      step count N,
      function f(t,u)

begin
  u[0]:=u0;
  t[0]:=t0;
  for k from 0 to N-1 do
    t[k+1] := t[k] + h;
    u[k+1] := u[k] + h*f(t[k],u[k]);
  end do;
return u[N]

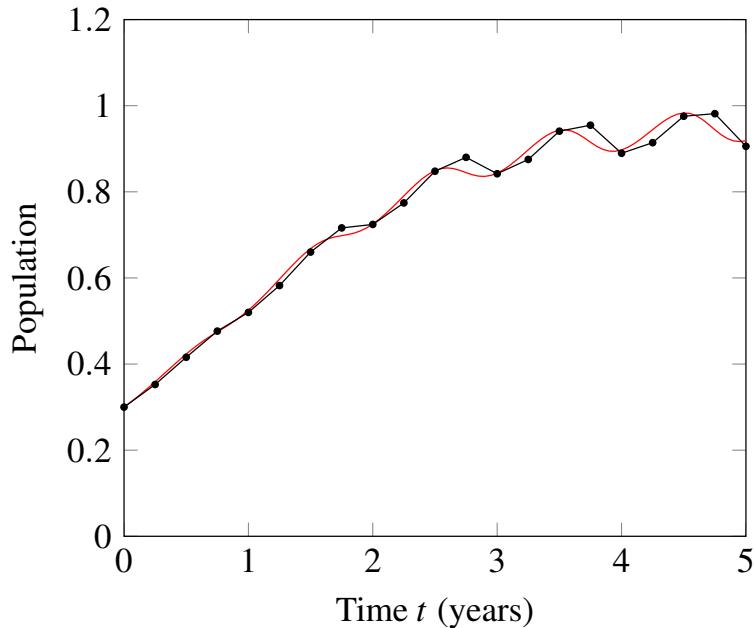
```

Figure 3.4: Pseudocode for Euler's method.

Iteration	t_k	u_k	$f(t_k, u_k)$
0	0	0.3	0.21
1	0.25	0.3525	0.2531
2	0.5	0.4158	0.2429
3	0.75	0.4765	0.1738
4	1.0	0.5199	0.2496
5	1.25	0.5823	0.3110
6	1.5	0.6601	0.2244

Table 3.1: Euler's method approximation for (3.1), step size $h = 0.25$.

Logistic Equation with Time-Dependent Carrying Capacity

Figure 3.5: Euler's method approximation (black) for (3.1), step size $h = 0.25$. True solution shown in red.

■ **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 solution at discrete times $t = t_0, t_0 + h, t_0 + 2h, \dots, t_0 + Nh$. If these points are connected, a polygonal approximation to the true solution's graph is obtained as shown in Figure 3.5 out to $t = 5$, superimposed over the graph of the “true” solution. This true solution was computed numerically using much more accurate methods, discussed in following sections. It looks like Euler's Method does a reasonable job. But what if a more accurate solution is desired? ■

3.2.2 The Accuracy of Euler's Method

The most straightforward method 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 .

■ **Example 3.2** Let's estimate $u(1.0)$ for the solution to $u'(t) = u(t)$ with initial condition $u(0) = 1$,

Step size h	Euler estimate $u_h(1.0)$	Error
1.0	2.0	0.71828
0.1	2.59374	0.12454
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 $u(1) = e$ for various step sizes h , with error $|e - u_h(1.0)|$.

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 $u(1) = e \approx 2.71828128$. This makes it easy to observe how the error in Euler's Method behaves here. 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 $|e - u_h(1.0)|$. ■

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 ODE's 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 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 $t^* \in (a, b)$, 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, this gives a handle on 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 stage). Equation (3.13) shows 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 given by $\frac{1}{2}u''(s)h^2$ 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 $\frac{K}{2}h^2$. 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

$$\frac{NK}{2}h^2 = \frac{K(T - t_0)/h}{2}h^2 = \frac{K(T - t_0)}{2}h \quad (3.14)$$

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 mentioned, the above argument isn't rigorous. For example, we assumed that $u_k = u(t_k)$ at the k th Euler step, that is, 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 demonstrate the following theorem.

Theorem 3.2.1 — 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 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 this see [21]. The constant C in Theorem 3.2.1 is not typically known, but *the main point is that in Euler's Method the error is proportional to the step size h* . If you want an answer that is 10 times more accurate, you will likely need to decrease h by a factor of 10, which means 10 times more work to march the solution out to time $t = T$. This may seem reasonable, but in fact it is possible to do much better!

3.2.3 Exercises

In Exercises 3.2.1 to 3.2.4 apply Euler's Method by hand (but use a calculator!) to the given initial value problem using the indicated step size h and number of steps N . Carry computations to 4 significant figures. Then compute the value of the true (analytic) solution at time $T = t_0 + Nh$ and compare.

Exercise 3.2.1 $u'(t) = u(t) + 3, u(0) = 1$, step size $h = 0.5, N = 2$ steps ■

Exercise 3.2.2 $u'(t) = -u(t) + 3t, u(0) = 2$, step size $h = 0.25, N = 4$ steps ■

Exercise 3.2.3 $u'(t) = 1/u(t), u(0) = 2$, step size $h = 0.25, N = 4$ steps ■

Exercise 3.2.4 $u'(t) = tu(t), u(1) = 3$, step size $h = 0.1, N = 5$ steps ■

In Exercises 3.2.5 to 3.2.8 apply Euler's method using whatever technology you have available to the given initial value problem using 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 $t = T$. Does Theorem 3.2.1 seem to be accurate? With what value of C ?

Exercise 3.2.5 $u'(t) = 1 - u(t)/3, u(0) = 2$. Estimate $u(5)$ using step sizes $h = 1, 0.1, 0.01$. ■

Exercise 3.2.6 $u'(t) = te^{-u(t)}, u(0) = 1$. Estimate $u(3)$ using step sizes $h = 1, 0.1, 0.01$. ■

Exercise 3.2.7 $u'(t) = u^2(t)$, $u(0) = 2$. Estimate $u(0.5)$ using step sizes $h = 0.5, 0.1, 0.01, 0.001$.

■

Exercise 3.2.8 $u'(t) = 1/u(t)$, $u(0) = 2$. Estimate $u(4)$ using step sizes $h = 1.0, 0.01, 0.001$.

■

Exercise 3.2.9 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 3.2.10 A variation on the Hill-Keller ODE (1.3) is to take the resistive force as $F_r = -kmv^r(t)$ for some constants $k > 0$, $r \geq 1$, and leads to the ODE

$$v'(t) = P - kv^r(t) \quad (3.15)$$

under the assumption that $v \geq 0$ and where P still is the same “propulsive effort” as in Section 1.1.2. (Compare (3.15) to (2.73) in Example 62.) Equation (3.15) 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 $F_r = -kv^r$ leads to the ODE (3.15).
- (b) Sketch a phase portrait for (3.15), limiting attention to $v \geq 0$. Label the fixed point in terms of the constants P, k, r .
- (c) Use the Existence-Uniqueness Theorem 2.4.1 to show that (3.15) has a unique solution for $v(t_0) = v_0 \geq 0$.
- (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 to the expected nature of the solution 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 3.2.11 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 $h = 0.1, 0.01, 0.001$. 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 3.2.12 Consider the linear ODE

$$u'(t) = u(t) - \sin(t) + \cos(t).$$

- (a) Find a general solution to this ODE.
- (b) Find the solution with initial condition $u(0) = 0$.
- (c) Sketch a direction field for this ODE 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 “steps off” the analytical solution curve? It might be helpful to plot the Euler iterates for $h = 0.001$. ■

Exercise 3.2.13 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 3.2.10.

Consider the differential equation $u'(t) = -10u(t)$ with $u(0) = 1$.

- Find the analytical solution to this initial value problem, and show that it decays to zero as $t \rightarrow \infty$.
- Apply Euler's method with step size $h = 0.1$ to estimate $u(5)$. Hint: after the first step the computations should be trivial.
- Apply Euler's method with step size $h = 0.2$ to estimate $u(5)$. Hint: a simple pattern should emerge.
- Apply Euler's method with step size $h = 1$ to estimate $u(5)$. Hint: again, there is a pattern!
- 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

$$u_k = (1 - 10h)^k. \quad (3.16)$$

Equation (3.16) should yield results in accord with parts (b)-(d).

- From part (a) the analytical solution decays to zero. How large can we take $h > 0$ in (3.16) to (at least) obtain decay to zero? Interpret the results of parts (b)-(d) in light of this analysis. ■

3.3 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 3.6. For the k th iteration (k doesn't matter) we use $t_k = 0.2, u_k = 0.2$, shown as the black dot, and assume the true solution passes precisely through this point. Euler's Method with step $h = 0.4$ is used to extrapolate to $t = t_{k+1} = 0.6$ via (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 is shown in Figure 3.6 as the blue line segment, with $t_k = 0.6, u_k = 0.32$ as the blue dot. The direction field for $u' = f(t, u)$ is shown in black, the true solution to $u' = f(t, u)$ in red, and the correct solution value $(0.6, 418\dots)$ as the red dot. 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 $u(t_{k+1})$.

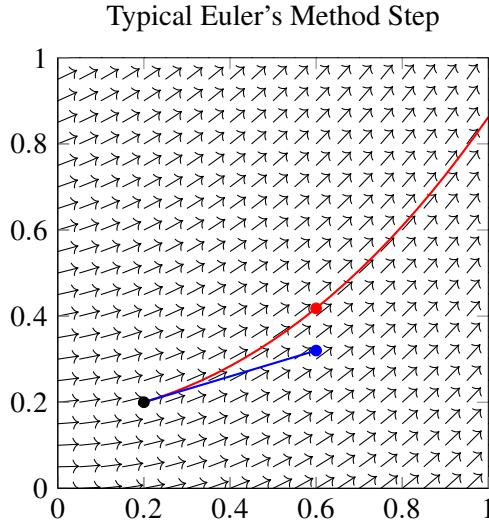


Figure 3.6: A typical Euler's Method step (blue) from $t_k = 0.2$ to $t_{k+1} = 0.6$ compared to true solution (red), and direction field for the ODE $u'(t) = t + u(t)/2$.

3.3.1 Improving Euler's Method

The Euler extrapolation can be improved by recognizing that the direction field changes from $t = t_k$ to $t = t_k + h$. One way to incorporate this observation is as follows:

- From (t_k, u_k) (the solid black dot in Figure 3.7) 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

$$w = u_k + hf(t_k, u_k). \quad (3.17)$$

In Figure 3.7 this step is illustrated by the solid blue line segment; w is the vertical coordinate of the right tip of the solid blue segment, shown as a half-filled blue dot. Notice that this estimate is already significantly off of the true solution curve, shown in red.

- Take a second Euler step of step size h by extrapolating along the line with slope $f(t_k + h, w)$, starting from the point $(t_k + h, w)$. This is illustrated by the dashed blue line segment in Figure 3.7. The resulting point at the right end of this segment is highlighted by the solid blue dot and has t coordinate $t_k + 2h$ and vertical coordinate \tilde{w} given by

$$\tilde{w} = w + hf(t_k + h, w). \quad (3.18)$$

- Construct an improved estimate u_{k+1} for $u(t_k + h)$ by taking the average value of u_k and \tilde{w} . This can be viewed from a geometric perspective as taking u_{k+1} to be the vertical coordinate of the midpoint of a line segment joining (t_k, u_k) to $(t_k + 2h, \tilde{w})$. This segment is shown in black in Figure 3.7, and the midpoint as the solid red dot. The vertical coordinate of this midpoint is the final estimate for u_{k+1} , and is

$$\begin{aligned} u_{k+1} &= \frac{u_k + \tilde{w}}{2} \\ &= \frac{u_k + w + hf(t_k + h, w)}{2} \\ &= u_k + h \left(\frac{f(t_k, u_k) + f(t_k + h, w)}{2} \right), \end{aligned} \quad (3.19)$$

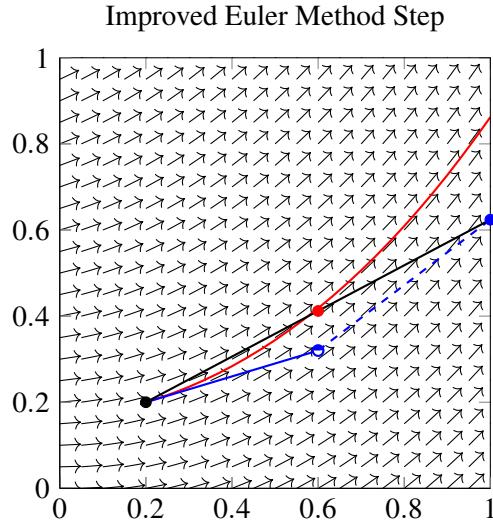


Figure 3.7: Graphic of the Improved Euler Method step. True solution in red, Euler step in blue, Improved Euler step in green.

where we have made use of (3.17) and (3.18). Equation (3.19) can also be viewed as linear extrapolation from $t = t_k$ to $t = t_k + h$

$$u_{k+1} = u_k + hm \quad (3.20)$$

with slope

$$m = \frac{f(t_k, u_k) + f(t_k, w)}{2} \quad (3.21)$$

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 (3.21) 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 3.7 it appears that the improved Euler estimate is much superior to the standard Euler step.

■ Example 3.4 Let's consider some specific numbers relevant to Figure 3.7. This figure is based on the specific choices $t_k = 0.2, u_k = 0.2, h = 0.4$ 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 in red is given by $u(t) = 4.6e^{(5t-1)/10} - 2t - 4$ 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 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 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 (3.20) and (3.21). This is considerably more accurate than the Euler estimate. ■

```

algorithm Improved Euler Method:
input initial time t0,
    initial value u0,
    step size h,
    step count N,
    function f(t,u)

begin
u[0]:=u0;
t[0]:=t0;
for k from 0 to N-1 do
    t[k+1] := t[k] + h;
    w := u[k] + h*f(t[k],u[k]);
    m = 0.5*(f([k],u[k]) + f(t[k+1],w));
    u[k+1] = u[k] + h*m;
end do;
return u[N]

```

Figure 3.8: Pseudocode for Euler's method.

3.3.2 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 Method*. The pseudocode is shown in Figure 3.8.

■ Example 3.5 In this example we perform two steps of the Improved Euler 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 purposes the true solution here is $u(t) = 4e^t - t - 2$ and $u(1) = 4e - 3 \approx 7.873$. Equations (3.20) and (3.21) are used to march the solution out in time (equivalent to the single step formula (3.19), 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 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 estimate $u(1.0) \approx 6.0$, so the Improved Euler Method is substantially better. ■

Reading Exercise 67 Verify that two steps of the standard Euler's Method with step size $h = 0.5$ does in fact yield estimate $u(1.0) \approx 6.0$.

Step size h	Euler estimate	Euler Error	Improved Euler	Improved Euler Error
1.0	2.0	7.183×10^{-1}	2.5	2.813×10^{-1}
0.1	2.593742	1.245×10^{-1}	2.714081	4.201×10^{-3}
0.01	2.704813	1.347×10^{-2}	2.718237	4.497×10^{-5}
0.001	2.716924	1.358×10^{-3}	2.718281	4.522×10^{-7}
0.0001	2.718146	1.359×10^{-4}	2.718281	4.53×10^{-9}

Table 3.3: Euler Method and Improved Euler Method estimate of $u(1) = e$ for ODE $u'(t) = u(t)$ for various step sizes h , with error $|e - u_h(1.0)|$.

Reading Exercise 68 Continue the computations of Example 3.5 to estimate $u(2.0)$. Compare to the true solution value.

Accuracy Of the Improved Euler Method

The Improved Euler Method requires more work for each time step, but the payoff is much better accuracy. The Improved Euler Method is also sometimes known as the *Modified Euler Method* or *Heun's Method*, although the latter term is sometimes used to refer to another closely related method. The Improved Euler Method is also a simple example of a class of methods known as *Runge-Kutta Methods* for numerically solving ODE's. 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 show just how superior the Improved Euler Method is compared to the standard Euler's Method.

■ **Example 3.6** Consider the ODE $u'(t) = u(t)$ and initial condition $u(0) = 1$, with solution $u(t) = e^t$; this was examined for Euler's Method in Example 3.2 in the last section. Consider step sizes $h = 1, 0.1, 0.01, 0.001$, and $h = 0.0001$ for estimating $u(1)$. In each case both Euler's Method and the Improved Euler Method will be used to estimate $u(1) = e \approx 2.718281828$. This requires $N = 1, 10, 100, 1000$ and $N = 10000$ steps, respectively. The results are tabulated in Table 3.3.

An examination of the error for the Improved Euler 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 Method; for most ODE's it is true that

$$|u(T) - u_h(T)| \approx Ch^2 \quad (3.22)$$

for some constant C , which is a rather dramatic improvement over standard Euler's Method. ■

3.3.3 Exercises

In Exercises 3.3.1 to 3.3.4 apply the Improved Euler Method by hand (but use a calculator!) to the given initial value problem using the indicated step size h and number of steps N . Carry computations to 4 significant figures. Then compute the value of the true (analytic) solution at time $T = t_0 + Nh$ and compare.

■ **Exercise 3.3.1** $u'(t) = u(t) + 3$, $u(0) = 1$, step size $h = 0.5$, $N = 2$ steps. ■

■ **Exercise 3.3.2** $u'(t) = -u(t) + 3t$, $u(0) = 2$, step size $h = 0.25$, $N = 4$ steps. ■

Exercise 3.3.3 $u'(t) = 1/u(t)$, $u(0) = 2$, step size $h = 0.25$, $N = 4$ steps. ■

Exercise 3.3.4 $u'(t) = tu(t)$, $u(1) = 3$, step size $h = 0.1$, $N = 5$ steps. ■

In Exercises 3.3.5 to 3.3.8 apply the Improved Euler method using whatever technology you have available to the given initial value problem using 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 analytic solution.

Exercise 3.3.5 $u'(t) = 1 - u(t)/3$, $u(0) = 2$. Estimate $u(5)$ using step sizes $h = 1, 0.1, 0.01$. ■

Exercise 3.3.6 $u'(t) = te^{-u(t)}$, $u(0) = 1$. Estimate $u(3)$ using step sizes $h = 1, 0.1, 0.01$. ■

Exercise 3.3.7 $u'(t) = u^2(t)$, $u(0) = 2$. Estimate $u(0.5)$ using step sizes $h = 0.5, 0.1, 0.01, 0.001$. ■

Exercise 3.3.8 $u'(t) = 1/u(t)$, $u(0) = 2$. Estimate $u(4)$ using step sizes $h = 1.0, 0.1, 0.01$. ■

Exercise 3.3.9 Apply the Improved Euler 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 3.3.10 (Compare the results here to Exercise 3.2.11.) Apply the Improved Euler method to the ODE $u'(t) = u^2(t)$ with $u(0) = 1$, to estimate $u(2)$ using step sizes $h = 1, 0.1, 0.01, 0.001$. 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 3.3.11 (Compare the results here to Exercise 3.2.12.) Consider the linear ODE

$$u'(t) = u(t) - \sin(t) + \cos(t).$$

- Find a general solution to this ODE.
- Find the solution with initial condition $u(0) = 0$.
- Sketch a direction field for this ODE 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.
- Apply the Improved Euler 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 Improved Euler method ever “steps off” the analytical solution curve? It might be helpful to plot the Improved Euler iterates for $h = 0.001$. ■

Exercise 3.3.12 (Compare the results here to Exercise 3.2.13). This problem illustrates that if the step size is too large, the Improved Euler 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 method with step size $h = 0.1$ to estimate $u(5)$.
- (c) Apply the Improved Euler 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 method with step size $h = 1$ to estimate $u(5)$.
- (e) Suppose we apply the Improved Euler method with step size h as detailed in (3.19). 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

$$u_k = (50h^2 - 10h + 1)^k. \quad (3.23)$$

Equation (3.23) should yield results in accord with parts (b)-(d).

- (f) From part (a) the analytical solution decays to zero. How large can we take $h > 0$ in (3.23) to (at least) obtain decay to zero? Interpret the results of parts (b)-(d) in light of this analysis.

■

3.4 Modern Numerical Methods

Numerical methods for solving ODE's is a vast field of research, and 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 ODE's, but many general purpose algorithms exist. They 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, for example, 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 that control the algorithm's behavior. The goal in the section is not to make you an expert in ODE solvers, nor teach you to program your own, but rather help you become a knowledgeable user of modern ODE software.

3.4.1 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 4th order Runge-Kutta algorithm (often abbreviated “RK4”) forms the basis for many computer codes for solving ODE's numerically. Like the Improved Euler 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}

Step size h	RK4 estimate $u_h(1.0)$	Error
1.0	2.708333	9.95×10^{-3}
0.1	2.718279	2.084×10^{-6}
0.01	2.718282	2.247×10^{-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 $|e - u_h(1.0)|$.

using the following formulas:

$$\begin{aligned}
m_1 &= f(t_k, u_k), \\
m_2 &= f(t_k + h/2, u_k + hm_1/2), \\
m_3 &= f(t_k + h/2, u_k + hm_2/2), \\
m_4 &= f(t_k + h, u_k + hm_3), \\
m &= \frac{1}{6}(m_1 + 2m_2 + 2m_3 + m_4), \\
u_{k+1} &= u_k + hm.
\end{aligned} \tag{3.24}$$

These formulas might be seen as a generalization of the philosophy behind the Improved Euler Method (itself a member of the Runge-Kutta family), that is, 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 (3.24), but will illustrate their effectiveness in solving ODE's below. For a derivation of (3.24) see [66].

■ **Example 3.7** Let's apply the RK4 formulas (3.24) to the initial value problem $u'(t) = u(t) + t$ with initial condition $u(0) = 1$ (analytical solution $u(t) = 2e^t - t - 1$) and step size $h = 1$, the largest step size possible here. 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 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 the true solution value using the RK4 method with step size h , and the error is computed as $|e - u_h(1.0)|$. It appears, at least asymptotically, 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 ODE's it is true that for the RK4 method

$$|u(T) - u_h(T)| \approx Ch^4 \tag{3.25}$$

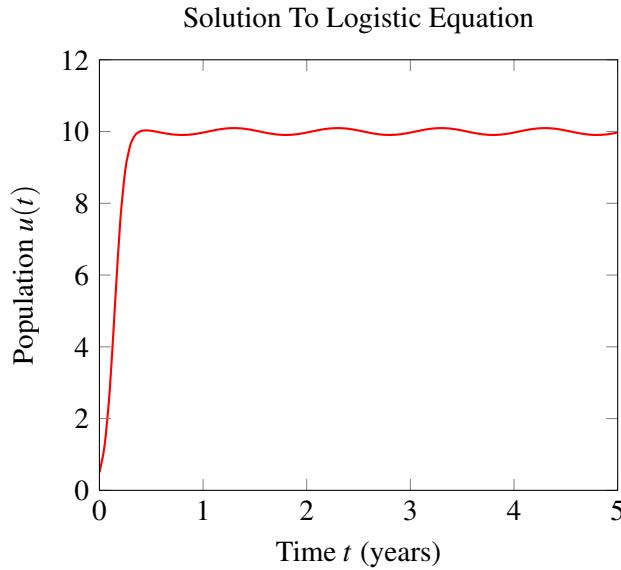


Figure 3.9: Solution to (3.26) with $r = 10$, $K(t) = 10 + 0.1 \sin(2\pi t)$, and initial condition $u(0) = 0.5$.

for some constant C , a huge improvement even over the Improved Euler Method. ■

Numerical ODE solvers can be designed with any desired asymptotic behavior in the error, e.g., Ch^n for any n . However, larger n require more complicated solution formulas and many more computations for each time step. 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 $n+1$ derivative of u . The RK4 method is considered a reasonable balance between accuracy and complexity.

Reading Exercise 69 Use the RK4 method to perform a single step of size $h = 1$ to estimate $u(1)$, where $u'(t) = -2u(t) + t^2$ and $u(0) = 0$. Compare to the true value for $u(1)$.

3.4.2 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 of interest is

$$u'(t) = 20u(t) \left(1 - \frac{u(t)}{10 + 0.1 \sin(2\pi t)}\right). \quad (3.26)$$

We'll use initial condition $u(0) = 0.5$. As in Section 3.1, this $K(t)$ could represent a seasonally varying carrying capacity that fluctuates a bit around $K = 10$. Equation (3.26) has no analytical solution and so must be solved numerically. The solution is shown in Figure 3.9. This solution was computed with an RK4 method using a small time step (on the order of 10^{-4}), and is sufficiently accurate that it may be considered the “true” solution.

The step size used in numerically solving (3.26) determines how closely the method follows the direction field, and so the accuracy of the computed solution. The true solution as shown in Figure 3.9 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 is slowly varying, a much larger step size can be used without unduly sacrificing accuracy. This is illustrated in Figure 3.10 in which an RK4

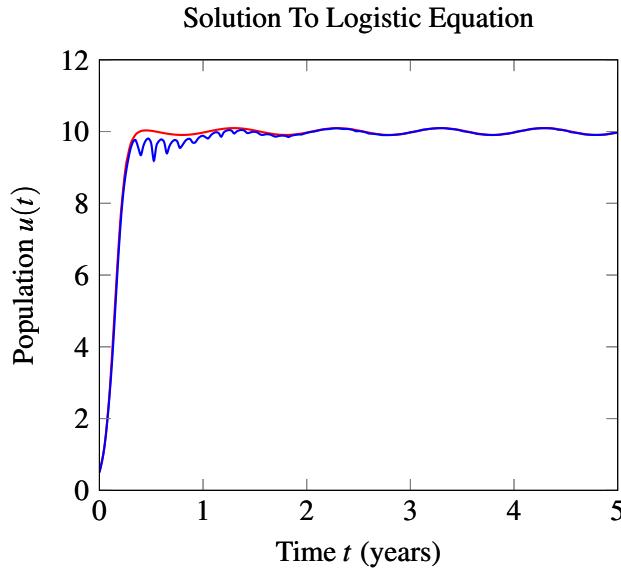


Figure 3.10: Solution to (3.26) with $r = 10$, $K(t) = 10 + 0.1 \sin(2\pi t)$, and initial condition $u(0) = 0.5$, accurate solution in red, RK4 method with fixed step size $h = 0.2$ in blue.

method with fixed step size $h = 0.2$ is used (blue curve), along with the more accurately computed solution (red). The fixed step size procedure is obviously inaccurate for $0.4 < t < 1.5$ or so.

This example illustrates needs that frequently conflict when solving ODE's 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.

But how are we supposed to estimate the accuracy with which the true solution is being tracked when we don't what the true solution is!?

The Idea Behind Adaptive Step Sizing

Let's look at how one might implement adaptive step sizing 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 due to the step size, and then increase or decrease the step size accordingly.

The situation is illustrated in Figure 3.11. 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 $u(t_k + h)$, according to (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

$$u(t_k + h) = u(t_k) + hu'(t_k) + \frac{1}{2}u''(s)h^2. \quad (3.27)$$

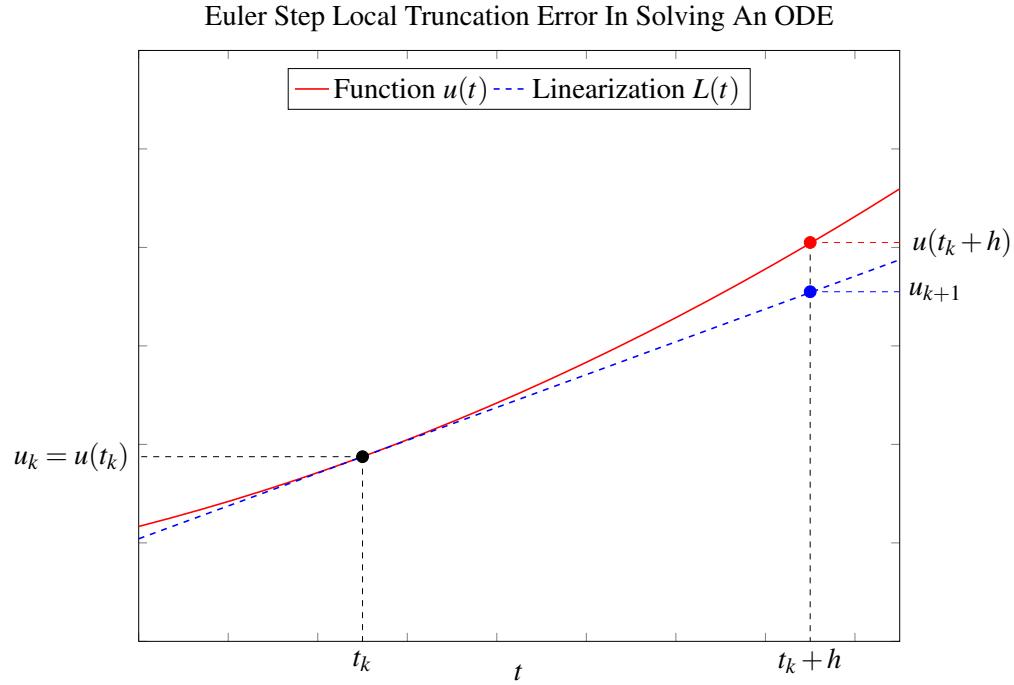


Figure 3.11: Graph of $y = u(t)$ (red, solid) and the linearization $y = L(t)$ (blue, dashed) of u at $t = t_k$, with points (t_k, u_k) (black dot), $(t_k + h, u(t_k + h))$ (red dot) and Euler step $(t_k + h, u_{k+1})$ (blue dot).

Here s lies between t_k and $t_k + h$; in fact 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 (3.27) becomes

$$u(t_k + h) = \underbrace{u_k + hf(t_k, u_k)}_{\text{Euler step } u_{k+1}} + \underbrace{\frac{1}{2}u''(s)h^2}_{\text{local truncation error}} . \quad (3.28)$$

As indicated above, 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 stepping 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.

If h is close to zero then $s \approx t_k$, since $t_k < s < t_k + h$. As a result, the local truncation error will be close to $\frac{1}{2}u''(t_k)h^2$, at least if u'' is continuous. For notational simplicity define $C = \frac{1}{2}u''(t_k)$, so the local truncation error is approximately Ch^2 . What (3.28) shows is that if we start with $u_k = u(t_k)$ (no error at step k) and take a step h to estimate $u(t_{k+1})$ with Euler's Method then the local truncation error $u(t_k + h) - u_{k+1}$ is given approximately by

$$u(t_k + h) - u_{k+1} \approx Ch^2 \quad (3.29)$$

where $C = \frac{1}{2}u''(t_k)$.

Reading Exercise 70 Let $u(t) = 2e^t - t - 1$, which is a solution to $u'(t) = t + u(t)$ (so $f(t, u) = t + u$, 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 (3.29) seem to hold? Based on your table, what is the appropriate value for C here? Is it close to $\frac{1}{2}u''(t_k)$?

Equation (3.29) can be used 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 (3.29) 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 $u(t_k) \approx u_k$, with h as the current “default” step size.

1. Take an Euler step $u_{k+1} = u_k + hf(t_k, u_k)$ of size h . In (3.29) both h and u_{k+1} are known, but not C nor $u(t_k + h)$.
2. Starting at $t = t_k$ again, take two Euler steps, each of size $h/2$, as

$$u_{k+1/2} = u_k + \frac{h}{2}f(t_k, u_k), \quad (3.30)$$

$$\tilde{u}_{k+1} = u_{k+1/2} + \frac{h}{2}f(t_k + h/2, u_{k+1/2}), \quad (3.31)$$

to again march 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 step (3.30) and (3.31) introduces approximate truncation error $C(h/2)^2$, for a total of $2C(h/2)^2 = Ch^2/2$. Then

$$u(t_k + h) - \tilde{u}_{k+1} \approx Ch^2/2. \quad (3.32)$$

4. Equations (3.29) and (3.32) can be considered as two (approximate) equations in unknowns C and $u(t_k + h)$. In particular, subtract (3.32) from (3.29) so the $u(t_k + h)$ terms cancel and a bit of algebra yields

$$|\text{LTE}| = |Ch^2| \approx 2|\tilde{u}_{k+1} - u_{k+1}| \quad (3.33)$$

which is an estimate of $|\text{LTE}|$, the magnitude of the local truncation error.

Equation (3.33) is the punchline: by taking a single Euler step of size h to produce estimate u_{k+1} and then repeating with two steps of size $h/2$ to produce estimate \tilde{u}_{k+1} , (3.33) can be used to estimate the error Ch^2 in for the step of size h , and then a decision made on 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 $f(t, u) = t + u$, initial condition is irrelevant). 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, a single step of size $h = 0.1$ the Euler estimate is

$$u_{k+1} = u_k + hf(t_k, u_k) = 5.95 + 0.1f(1.5, 5.95) = 6.695 \quad (3.34)$$

for $u(1.6)$. Alternatively, take two steps of size $h/2 = 0.05$, as

$$\begin{aligned} u_{k+1/2} &= u_k + \frac{h}{2}f(t_k, u_k) = 5.95 + 0.05f(1.5, 5.95) = 6.3225, \\ \tilde{u}_{k+1} &= u_{k+1/2} + \frac{h}{2}f(t_k + h/2, u_{k+1/2}) = 6.716125. \end{aligned} \quad (3.35)$$

From (3.33) 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 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 71 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 $h = 0.2$.

Adaptive Step Sizing and Error Control

Let’s consider an example to see how (3.33) 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. It might be hoped 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 generally no way to guarantee that the final estimate for $u(1)$ is accurate to within a given 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 10 percent of the current solution value (or better yet, something conservative like 1 percent) and hope this translates into something comparable in the estimate for $u(1)$.

Let’s try this strategy. At the k th iteration in Euler’s Method it will be required that the estimate local truncation error as given by (3.33) satisfies

$$|\text{LTE}| \leq \tau |u_k| \quad (3.36)$$

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 estimate

$$u_1 = u_0 + hf(t_0, u_0) = 1.125.$$

Two Euler steps of size $h/2 = 0.125$ produce 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 (3.33) to estimate the local truncation error $\text{LTE} = Ch^2$ for the step of size $h = 0.25$ as

$$|\text{LTE}| = 2|\tilde{u}_1 - u_1| = 0.039.$$

The current iterate is $u_0 = 1$ and with $|\text{LTE}| = 0.039$ the condition (3.36) is not satisfied. The step of size $h = 0.25$ is rejected.

The value of h must thus be decreased, say by cutting h in half and trying 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 :

$$u_{1/2} = u_0 + \frac{h}{2} f(t_0, u_0) = 1.03125,$$

$$\tilde{u}_1 = u_{1/2} + \frac{h}{2} f(t_0 + h/2, u_{1/2}) = 1.0674.$$

Estimate the local truncation error LTE for the step of size $h = 0.125$ by using (3.33) to find

$$|\text{LTE}| = 2|\tilde{u}_1 - u_1| = 0.00977.$$

In this case the inequality (3.36) is satisfied. The iterate $u_1 = 1.0674$ is accepted and $t_1 = 0.125$.

In the next step of Euler's Method to extrapolate to $t_2 = t_1 + h$ we begin with the current value of $h = 0.125$ and this process continues. 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 (say by a factor of 2) at some iteration if the estimate of the local truncation error is much less than the budgeted amount.

As noted earlier however, despite use of the imposed tolerance (3.36) there is no firm guarantee that the final answer is within any given tolerance of the correct answer, since the errors are not simply additive! An error made at any given 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 use two different ODE solvers to accomplish this. 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 $u' = f(t, u)$. In particular, 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 like (3.36). Additionally, `ode45` accepts an argument “`AbsTol`” that enforces a bound $|\text{LTE}| \leq \varepsilon$ for some tolerance ε . If at any given 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 3.4.9.

For much more information on the numerical solution of ODE's, see [66].

3.4.3 Exercises

In Exercises 3.4.1 to 3.4.4 apply the Runge-Kutta 4th order method (3.24) by hand (but use a calculator!) to the given initial value problem 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 (analytic) solution at time $T = t_0 + Nh$ and compare.

Exercise 3.4.1 $u'(t) = u(t) + 3$, $u(0) = 1$, step size $h = 0.5$, $N = 2$ steps. ▀

Exercise 3.4.2 $u'(t) = -u(t) + 3t$, $u(0) = 2$, step size $h = 0.5$, $N = 2$ steps.

Exercise 3.4.3 $u'(t) = 1/u(t)$, $u(0) = 2$, step size $h = 0.5$, $N = 2$ steps.

Exercise 3.4.4 $u'(t) = tu(t)$, $u(1) = 3$, step size $h = 0.25$, $N = 2$ steps.

In Exercises 3.4.5 to 3.4.8 apply the Runge-Kutta 4th order method using whatever technology you have available to the given initial value problem using 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 analytic solution.

Exercise 3.4.5 $u'(t) = 1 - u(t)/3$, $u(0) = 2$. Estimate $u(5)$ using step sizes $h = 1, 0.1, 0.01$.

Exercise 3.4.6 $u'(t) = te^{-u(t)}$, $u(0) = 1$. Estimate $u(3)$ using step sizes $h = 1, 0.1, 0.01$.

Exercise 3.4.7 $u'(t) = u^2(t)$, $u(0) = 2$. Estimate $u(0.5)$ using step sizes $h = 0.5, 0.1, 0.01, 0.001$.

Exercise 3.4.8 $u'(t) = 1/u(t)$, $u(0) = 2$. Estimate $u(4)$ using step sizes $h = 1.0, 0.1, 0.01$.

Exercise 3.4.9 Consider the logistic ODE (3.26) with initial condition $u(0) = 0.5$. The solution at $t = 1$ is $u(1) = 9.971345698$, to ten significant figures.

- Estimate $u(1)$ by using Euler's Method with steps of size $h = 0.1$, and compute the error.
- Estimate $u(1)$ by using the Improve Euler Method with steps of size $h = 0.1$, and compute the error.
- Estimate $u(1)$ by using the RK4 Method with steps of size $h = 0.1$, and compute the error.
- Estimate $u(1)$ by using whatever numerical ODE solver you have available without specifying the method, and compute the error. Most software, e.g., Mathematica, Maple, Matlab, and Sage, will use a good ODE solver with adaptive stepsizing for error control.
- Explore the error tolerances in the solver your 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 3.4.10 (Compare the results here to Exercises 3.2.11 and 3.3.10.) Apply the RK4 method to the ODE $u'(t) = u^2(t)$ with $u(0) = 1$, to estimate $u(2)$ using step sizes $h = 1, 0.1, 0.01, 0.001$. 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 3.4.11 (Compare the results here to Exercises 3.2.12 and 3.3.11.) Consider the linear

ODE

$$u'(t) = u(t) - \sin(t) + \cos(t).$$

- (a) Find a general solution to this ODE.
- (b) Find the solution with initial condition $u(0) = 0$.
- (c) Sketch a direction field for this ODE 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 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 “steps off” the analytical solution curve? It might be helpful to plot the RK4 iterates for $h = 0.001$.

■

Exercise 3.4.12 (Compare to Exercises 3.2.13 and 3.3.12). 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 Runge-Kutta 4th order method with step size $h = 0.1$ to estimate $u(5)$.
- (c) Apply the Runge-Kutta 4th order method with step size $h = 0.2$ to estimate $u(5)$.
- (d) Apply the Runge-Kutta 4th order method with step size $h = 1$ to estimate $u(5)$.
- (e) Suppose we apply the Runge-Kutta 4th order 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 3.4.13 Suppose we are 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 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 $u(t_k + h)$ (equals $u(0.6)$ here).
- (b) Take two Euler steps of size $h/2$ to produce \tilde{u}_{k+1} , an estimate of $u(t_k + h)$.
- (c) Use (3.33) to estimate the local truncation error. Is it within the accepted tolerance? If not, would $h/2$ work?

■

3.5 Parameter Estimation

3.5.1 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

$$v'(t) = P - kv(t) \tag{3.37}$$

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 (3.37) 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? This 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 (3.37) with $P = 11$ and $v(0.165) = 0$, leaving k as an unspecified constant. The solution can be obtained via separation of variables or the integrating factor technique and is

$$v(t) = \frac{11}{k} \left(1 - e^{-k(t-t_0)} \right) \quad (3.38)$$

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 (3.38) to compute Bolt's position $x(t)$ as a function of time. His position $x(t)$ satisfies $x'(t) = v(t)$ with $x(t_0) = 0$ (he starts at position $x = 0$, the starting line, at time t_0). The function $x(t)$ can be computed, using τ as the variable of integration, as

$$\begin{aligned} x(t) &= \int_{t_0}^t v(\tau) d\tau \\ &= \int_{t_0}^t \frac{11}{k} \left(1 - e^{-k(\tau-t_0)} \right) d\tau \\ &= \frac{11}{k} \left(\tau + \frac{e^{-k(\tau-t_0)}}{k} \right) \Big|_{\tau=t_0}^{\tau=t} \\ &= \frac{11}{k^2} \left(e^{-k(t-t_0)} - 1 + k(t-t_0) \right). \end{aligned} \quad (3.39)$$

The function $x(t)$ in (3.39) gives us the ability 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 meters) in (3.39). The graph of the resulting function $x(t)$ is shown in Figure 3.12, overlayed on the data from Table 1.1. 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 "better" here may be 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 .

3.5.2 Least-Squares Estimation

First, $x(t)$ as defined by (3.39) depends not only on t , but also on k . Let's indicate this dependence by writing

$$x(k, t) = \frac{11}{k^2} \left(e^{-k(t-t_0)} - 1 + k(t-t_0) \right). \quad (3.40)$$

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

$$x(k^*, t_1) - x_1 = 0,$$

$$x(k^*, t_2) - x_2 = 0,$$

$$\vdots = 0,$$

$$x(k^*, t_{10}) - x_{10} = 0.$$

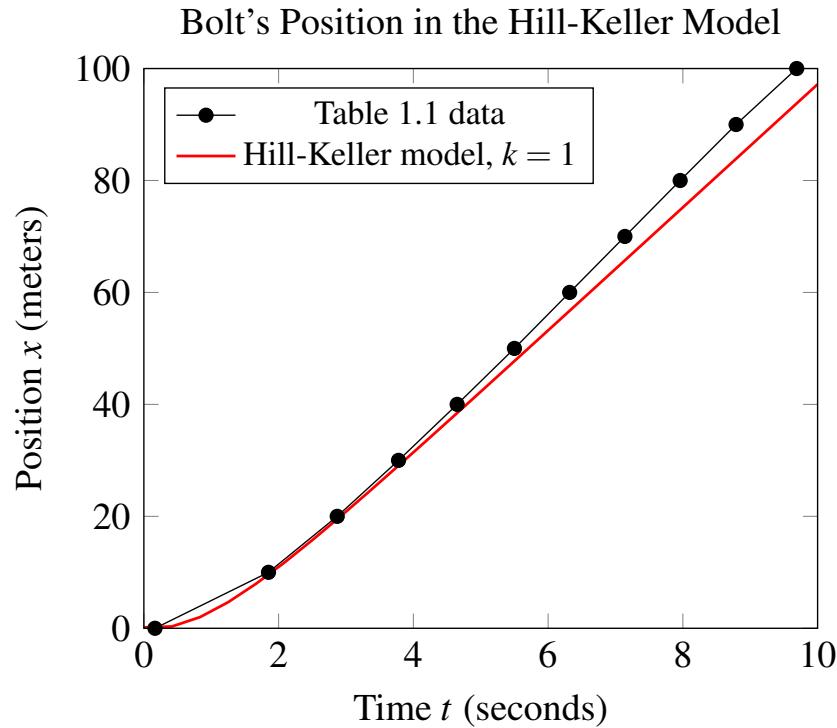


Figure 3.12: Hill-Keller data from Table 1.1 versus equation (3.39) with $k = 1$.

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 “small” and “on average” need to be quantified, and there are any number ways to do this. One common approach is to use

$$s_j(k) = (x(k, t_j) - x_j)^2 \quad (3.41)$$

as a measure of how well a given value of k fits the j th data point. Note that

- $s_j(k) \geq 0$ always
- $s_j(k) = 0$ 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 72 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

$$S(k) = \sum_{j=1}^{10} s_j(k) = \sum_{j=1}^{10} (x(k, t_j) - x_j)^2. \quad (3.42)$$

The function $S(k)$ is called the *sum of squares* function for this problem. Of course $S(k)$, being a sum of non-negative terms, always satisfies $S(k) \geq 0$. If any quantity $x(k, t_j) - x_j \neq 0$ then $s_j(k) > 0$

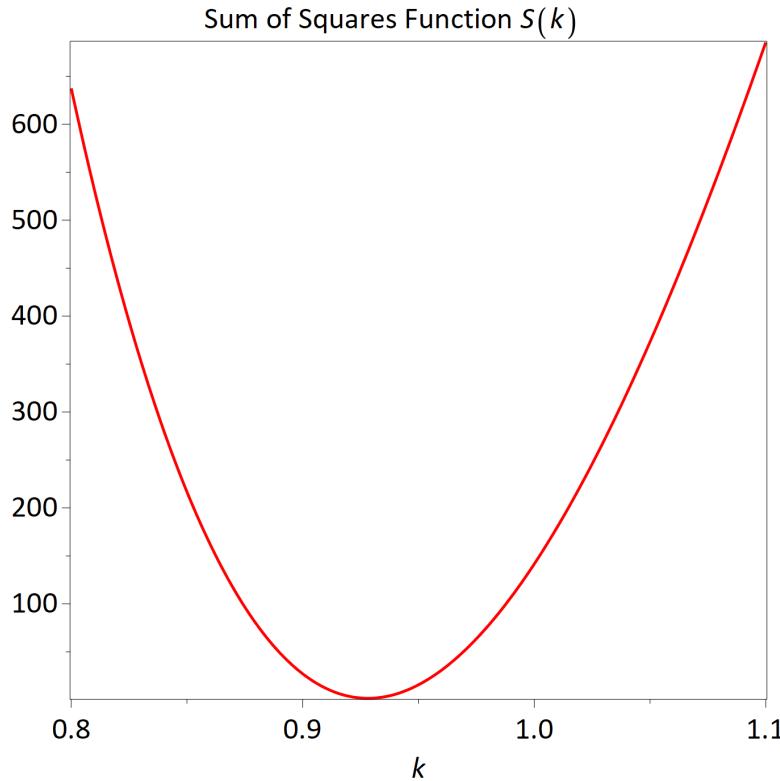


Figure 3.13: Plot of $S(k)$ defined by (3.43).

and this adds a positive contribution to the value of $S(k)$, assuring that $S(k) > 0$. The condition $S(k) = 0$ is satisfied exactly when $x(k, t_j) - x_j = 0$ for each 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, “on average.”

Let’s look at $S(k)$ for the Hill-Keller model and data. Writing out $S(k)$ explicitly by making use of (3.40) yields

$$\begin{aligned}
 S(k) &= \sum_{j=1}^{10} (x(k, t_j) - x_j)^2 \\
 &= (x(k, t_1) - x_1)^2 + (x(k, t_2) - x_2)^2 + \cdots + (x(k, t_{10}) - x_{10})^2 \\
 &= ((10 - 18.535/k - 11e^{-1.685k} - 1)/k^2)^2 \\
 &\quad + ((20 - 29.755/k - 11e^{-2.705k} - 1)/k^2)^2 \\
 &\quad + \cdots \\
 &\quad + ((100 - 104.775/k - 11e^{-9.525k} - 1)/k^2)^2. \tag{3.43}
 \end{aligned}$$

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 3.13. Visually, the best value of k appears to be somewhere around $k = 0.93$.

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, e.g., 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. And finally, it can be helpful to give the root-finding

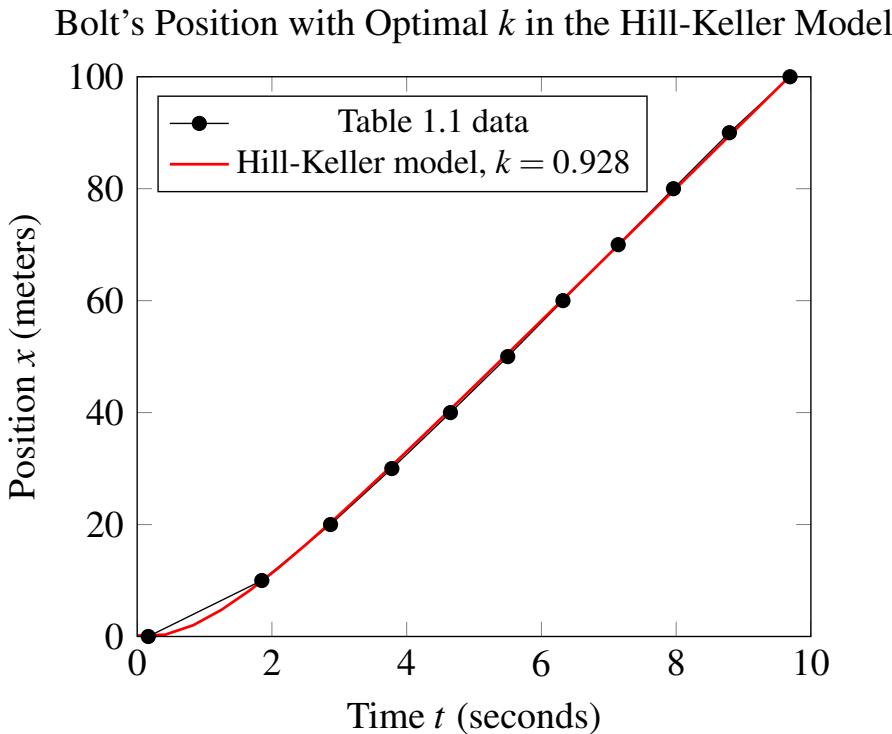


Figure 3.14: Hill-Keller data from Table 1.1 versus equation (3.39) with $k = 0.928$.

algorithm a good starting guess at the best value of k , or perhaps a range in which to seek k . Based on Figure 3.13, 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 3.14. Compare this to Figure 3.12. 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*.

In this case the function $S(k)$ was easy to minimize—it had a single, clear global (absolute) minimum, with no nearby local (relative) extrema. The situation won't always be so simple!

Reading Exercise 73 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) , $1 \leq j \leq 3$, 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.

3.5.3 Hill-Keller Again

The Hill-Keller model as presented in Chapter 1 contained a single unknown parameter k , which was estimated above. 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

$$v'(t) = P - kv(t) \quad (3.44)$$

with initial condition $v(t_0) = 0$, where $t_0 = 0.165$ and P and k are constants. The solution to (3.44) can be obtained via separation of variables or the integrating factor technique and is

$$v(t) = \frac{P}{k} \left(1 - e^{-k(t-t_0)} \right) \quad (3.45)$$

where $t_0 = 0.165$. As before position can be computed as

$$\begin{aligned} x(k, P, t) &= \int_{t_0}^t v(\tau) d\tau \\ &= \int_{t_0}^t \frac{P}{k} \left(1 - e^{-k(\tau-t_0)} \right) d\tau \\ &= \frac{P}{k} \left(\tau + \frac{e^{-k(\tau-t_0)}}{k} \right) \Big|_{\tau=t_0}^{t=t} \\ &= \frac{P}{k^2} \left(e^{-k(t-t_0)} - 1 + k(t-t_0) \right), \end{aligned} \quad (3.46)$$

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 was to be estimated, a sum of squares function $S(k, P)$ can be formed as

$$\begin{aligned} S(k, P) &= \sum_{j=1}^{10} (x(k, P, t_j) - x_j)^2 \\ &= (x(k, P, t_1) - x_1)^2 + (x(k, P, t_2) - x_2)^2 + \cdots + (x(k, P, t_{10}) - x_{10})^2 \\ &= (10 - P(e^{-1.685k} - 1 + 1.685k)/k^2)^2 \\ &\quad + (20 - P(e^{-2.705k} - 1 + 2.705k)/k^2)^2 \\ &\quad + \cdots \\ &\quad + (100 - P(e^{-9.525k} - 1 + 9.525k)/k^2)^2. \end{aligned} \quad (3.47)$$

The function $S(k, P)$ quantifies the fit to the data given by any pair of parameters k and P in the model (3.46). The goal is to find that pair $(k, P) = (k^*, P^*)$ that minimizes $S(k, P)$.

Minimizing the Sum of Squares 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 on the left in Figure 3.15. A contour plot for $S(k, P)$ is also shown on the right, 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 on the left in Figure 3.16, and a contour plot of $\ln(S(k, P))$ is shown on the right. This makes it a bit easier to see the location of the minimum. In this case, visually, $k \approx 0.87$ and $P \approx 10.3$ seem close.

In order to find the minimum precisely some computation is needed. The technique you learned in multivariable calculus is applicable here: Form equations

$$\frac{\partial S}{\partial k} = 0 \text{ and } \frac{\partial S}{\partial P} = 0$$

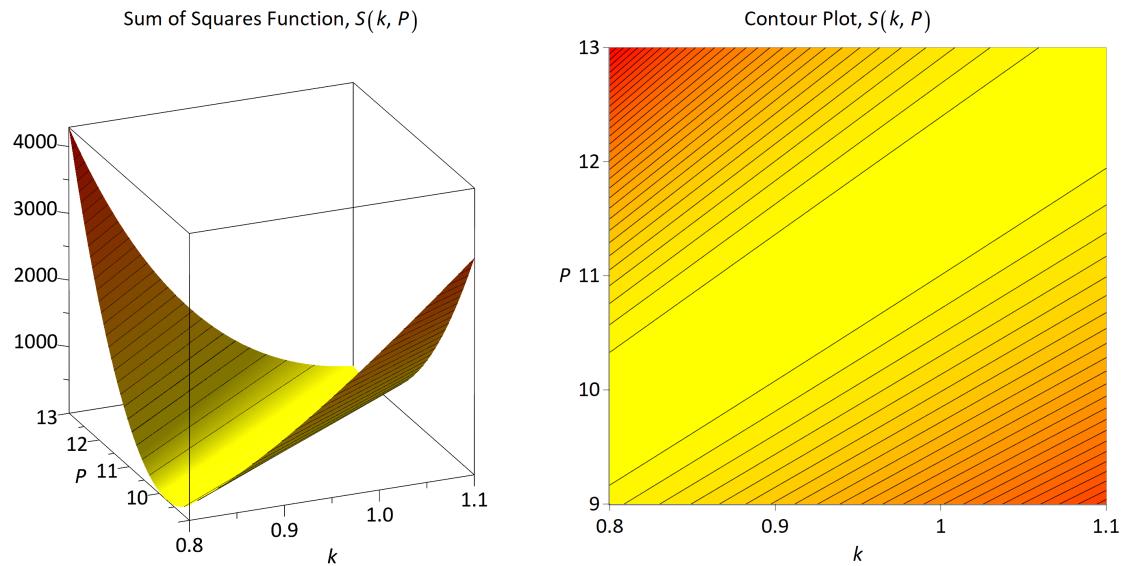


Figure 3.15: Graph of $S(k, P)$ defined by (3.47) (left) and contour plot (right).

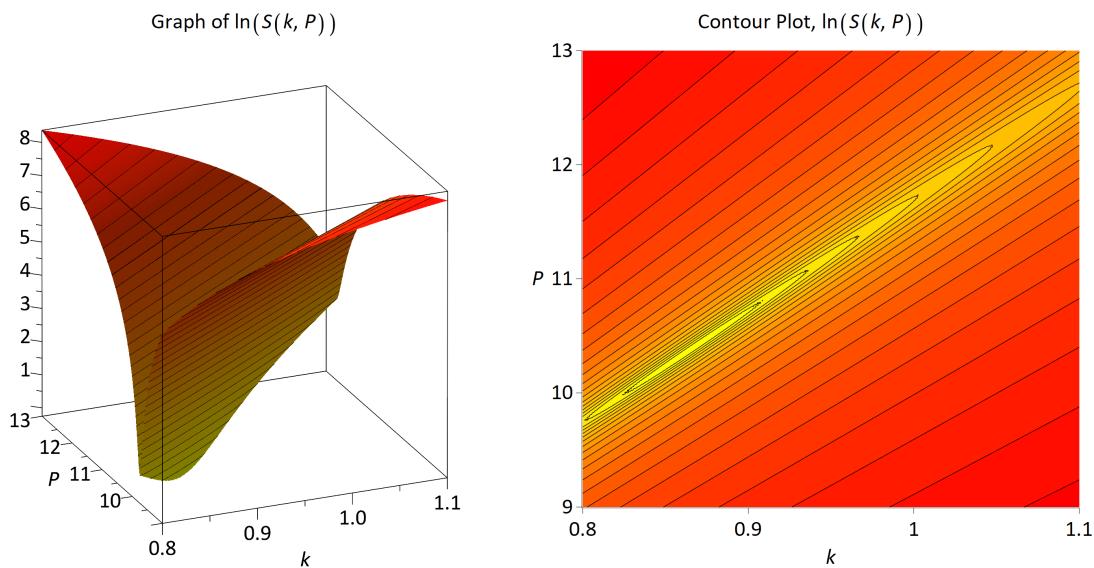


Figure 3.16: Graph of $\ln(S(k, P))$ defined by (3.47) (left) and contour plot (right).

Bolt's Position For Optimal k and P in the Hill-Keller Model

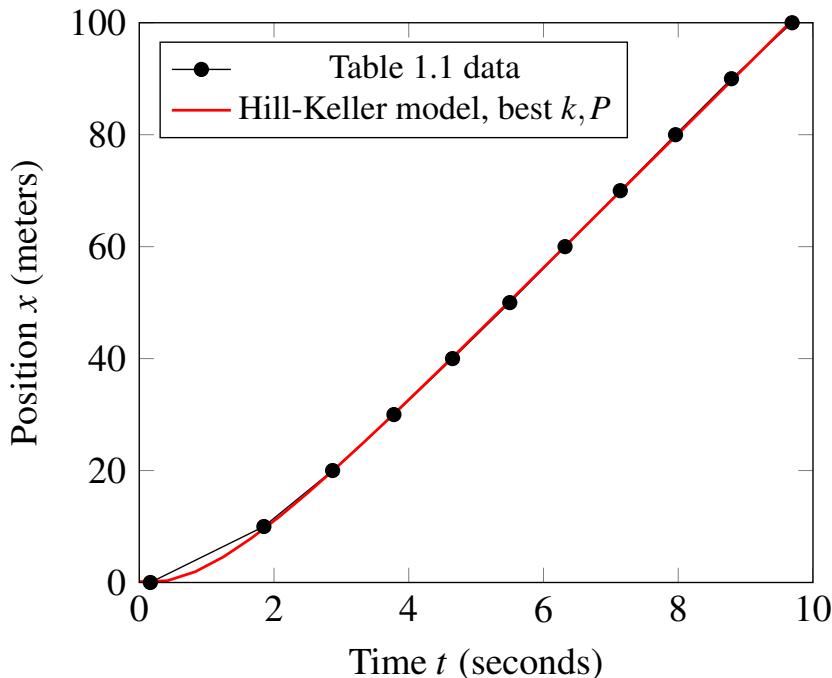


Figure 3.17: Hill-Keller data from Table 1.1 versus equation (3.46) with $k = 0.865$ and $P = 10.38$.

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 method, e.g., 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) compared to fitting only k . The resulting graph of $x(k^*, P^*, t)$ is shown in Figure 3.17, superimposed on the data. Compare this to Figures 3.12 and 3.14.

It seems that little has been gained in Figure 3.17, despite adding the second adjustable parameter P . This is a common phenomena 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 Model Selection" 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 74 This is a variation on Reading Exercise 73. Again, suppose 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) , $1 \leq j \leq 3$, 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

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>

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 compare 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 $z = \ln(S(k, A))$.

3.5.4 Least Squares For ODE Parameter Estimation

The General Setting

This least-squares approach to parameter estimation procedure 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

$$u'(t) = f(\boldsymbol{\alpha}, u(t)) \quad (3.48)$$

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

$$S(\boldsymbol{\alpha}) = \sum_{j=1}^n (u(\boldsymbol{\alpha}, t_j) - u_j)^2. \quad (3.49)$$

In particular, we seek that value $\boldsymbol{\alpha}^* = (\alpha_1^*, \dots, \alpha_m^*)$ that minimizes S .

The estimation procedure for k and/or P in the Hill-Keller model actually involved a slight variation on the above procedure. The sum of squares in that case was not $\sum_{j=1}^n (v(\boldsymbol{\alpha}, t_j) - v_j)^2$ that involves the velocity function itself, but rather the position function obtained from $v(t)$. Nonetheless, the principle is the same.

Remark 6 It may sometimes be convenient or advantageous to work with rescaled versions of the data and function $u(t)$ in (3.49), for example, by taking the logarithm of each. In this case we might use an alternate least-squares function of the form

$$\tilde{S}(\boldsymbol{\alpha}) = \sum_{j=1}^n (\ln(u(\boldsymbol{\alpha}, t_j)) - \ln(u_j))^2, \quad (3.50)$$

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 3.5.9 and 3.5.10.

Minimizing the Sum of Squares

The elementary multivariable calculus approach to minimizing $S(\boldsymbol{\alpha})$ in (3.49) 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 a large number of 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 α^* 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 [25].

3.5.5 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

$$r(t) = r_0 + A \sin(\omega t). \quad (3.51)$$

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

$$u'(t) = r(t)u(t)(1 - u(t)/K) \quad (3.52)$$

with $r(t)$ given by (3.51). The ODE (3.52) 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 (3.52) with initial condition $u(0) = u_0$ can be solved analytically using separation of variables. The solution is

$$u(t) = \frac{K}{e^{-R(t)}(K/u_0 - 1) + 1} \quad \text{where} \quad R(t) = \int r(t) dt \quad (3.53)$$

and 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 (3.53) is shown in Figure 3.18 on the range $0 \leq t \leq 40$.

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

$$S(\omega) = \sum_{j=0}^{20} (u(\omega, t_j) - u_j)^2 \quad (3.54)$$

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 least-squares function $S(\omega)$ is shown in Figure 3.19. Even in this ideal case, Figure 3.19 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 tailored to finding only local minima will very likely find a

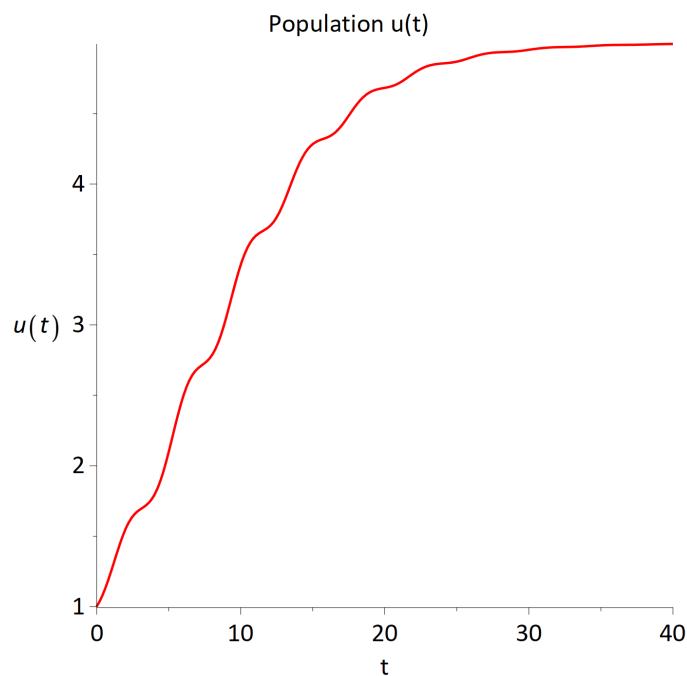


Figure 3.18: Function (3.53) satisfying (3.52), $K = 5, A = 0.7, u_0 = 1, r_0 = 0.2$, and $\omega = 1.5$.

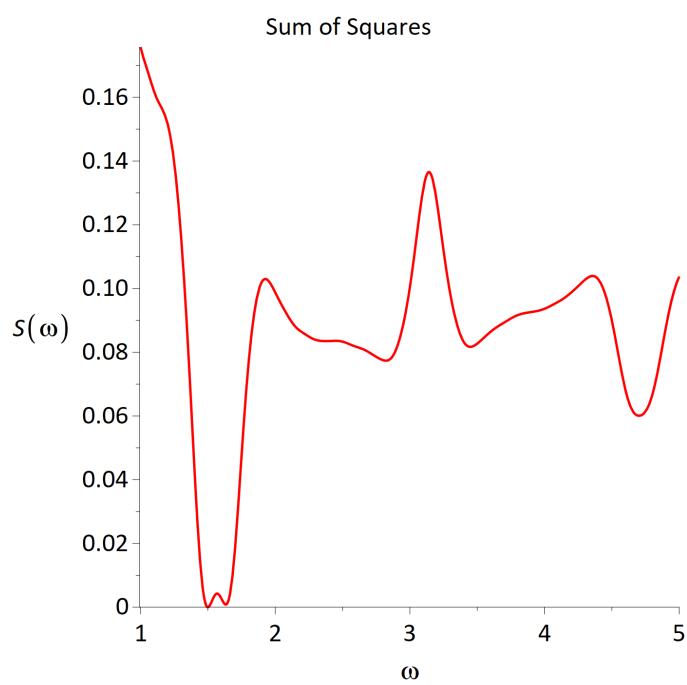


Figure 3.19: Sum of squares $S(\omega)$.

sub-optimal local minimum, not 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 this it pays, at the very least, to have a good estimate of what reasonable parameter values are, and then use this knowledge to inform the minimization algorithm's search.

3.5.6 Exercises

Exercise 3.5.1 Consider the set of data points (t_j, u_j) , $1 \leq j \leq 4$ below:

$$\{(0.1, 0.11), (0.6, 0.5), (1.1, 0.6), (1.4, 0.5)\}.$$

For each function u of the given form in (a)-(d) below, 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, plot the best-fit curve, and compare to a plot of the data.

- (a) $u(a, t) = at$, parameter a .
- (b) $u(a, b, t) = at + b$, parameters a, 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, c . Compare the residual to parts (b) and (c); it should be smaller than the residual for each of these. Why? ■

Exercise 3.5.2 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 (2.85) 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 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

$$\ln(y(t)) = -kt. \quad (3.55)$$

Equation (3.55) is the relation we will use to adjust k to best fit the data.

Using whatever software you have available, form the sum of squares

$$S(k) = (\ln(0.78) + 300k)^2 + (\ln(0.37) + 1200k)^2 + (\ln(0.08) + 3000k)^2$$

appropriate to (3.55) (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? Also plot the linear function $-k^* \ln(t)$ and compare to a plot of the data pairs (time, $\ln([H_2O_2])$) from Table 3.5. ■

Exercise 3.5.3 Table 3.6 contains split data for Tori Bowie's 2017 gold medal women's 100 meter victory in the 2017 IAAF World Championships (see [24]). Use the procedure of Section 3.5.3 to find the best choices k^* and P^* for k and P in the Hill-Keller model for this data; note

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.

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	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.

that the initial condition is $v(0.182) = 0$, and you need to fit her position function $x(t)$, not her velocity $v(t)$ to the data. Plot the function $x(t)$ with these parameters and compare to the data. ■

Exercise 3.5.4 Use the optimal parameter values $k = 0.865$ and $P = 10.38$ found for Usain Bolt in Section 3.5.3 for the Hill-Keller model to predict how fast Bolt could run 200 meters, and compare to the current 200 meter world record. It will be helpful to recall the position function $x(t)$ defined in (3.46). 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 is the prediction so far off? ■

Exercise 3.5.5 Suppose we have data points (x_j, y_j) for $j = 1$ to $j = n$ and 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 convenient property that a function $f(m)$ of a single variable can possess to guarantee that a minimizer for f exists and is unique 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 3.5.6 Show that the result of Exercise 3.5.5 holds if the model is $y_j = mx_j + b$ where b is considered known. Hint: Define $\tilde{y}_j = y_j - b$. ■

Exercise 3.5.7 Another approach to fitting parameters is “ 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 equation (3.49) is

$$S(\boldsymbol{\alpha}) = \sum_{j=1}^n |u(\boldsymbol{\alpha}, t_j) - u_j| \quad (3.56)$$

where $\boldsymbol{\alpha}$ embodies the parameter(s) to be fit. In particular, we seek that value $\boldsymbol{\alpha}^* = (\alpha_1^*, \dots, \alpha_m^*)$ for the parameters that minimizes S in (3.56). 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 (3.56) is not typically differentiable, so it's not straightforward to make use of derivative information in seeking a minimum. Nonetheless, numerical algorithms exist for this type of minimization problem.

Redo Exercise 3.5.2 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 3.5.8 This is an extension of Exercise 2.2.16. In Table 3.7 are some population data concerning the growth of a species of yeast (from a classic study [33].) 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 growth rate r and the carrying capacity K . Based on the data in Table 3.7 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 (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

$$\frac{\partial S}{\partial r} = 0 \quad \text{and} \quad \frac{\partial S}{\partial K} = 0$$

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.

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, (minutes, degree Fahrenheit).

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 to a plot of the data. Does this seem like a good model? ■

Exercise 3.5.9 This exercise is inspired by the SIMIODE project “Potato Cooling” [104]. A medium-sized potato was placed in a microwave 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 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 (2.15), with solution $u(t) = A + (u_0 - A)e^{-kt}$ previously developed in (2.16). 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) , $1 \leq j \leq n$, we expect $u_j \approx 72 + 132e^{-kt_j}$ for the correct value of k .

In accord with (3.49) (where α there is just “ k ” here) and using the data from Table 3.8, form an appropriate sum of squares

$$S(k) = (72 + 132e^{-2k} - 193)^2 + \dots + (72 + 132e^{-30k} - 130)^2$$

(the initial term at time $t = 0$ is always zero.) Then

- (a) Graph $S(k)$ for $0.01 \leq k \leq 0.05$ and visually identify the minimum.
- (b) Find that value k^* that minimizes $S(k)$. What is the residual sum of squares?
- (c) Plot the resulting function $u(t) = 72 + 132e^{-k^*t}$ and compare to the data. Does this seem like a reasonable model? ■

Exercise 3.5.10 This is a variation on Exercise 3.5.9. It may be helpful to review Remark 6. For the case $A = 72, u_0 = 204$ 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) , $1 \leq j \leq n$ 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 $u_0 > A$)

$$\ln(u_j - A) \approx \ln(u_0 - A) - kt_j, \quad (3.57)$$

$1 \leq j \leq n$. 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:

- (a) Plot $\tilde{S}(k)$ for $0 \leq k \leq 0.05$ and visually identify the minimum.
- (b) Find the minimizer k^* by solving $\tilde{S}'(k) = 0$.
- (c) Plot the resulting function $u(t) = 72 + 132e^{-k^*t}$ and compare to the data. Does this seem like a reasonable model? Compare the value obtained in Exercise 3.5.9.
- (d) What advantage might minimizing $\tilde{S}(k)$ have over minimizing $S(k)$ from Exercise 3.5.9?

Exercise 3.5.11

- (a) Newton's Law of Cooling assumes that the rate of temperature change is proportional to $u - A$, the difference in the object's current temperature and the ambient temperature. A more general and flexible model might posit that

$$u'(t) = -F(u(t) - A) \quad (3.58)$$

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 (3.58) for some positive real numbers k and r ($r = 1$ is the usual Newton's Law of Cooling). Verify that F satisfies $F(0) = 0$ and that F is strictly increasing.

- (c) With F as in part (b) the ODE (3.58) becomes

$$u'(t) = \begin{cases} -k|u(t) - A|^r, & v < 0 \\ k|u(t) - A|^r, & v \geq 0 \end{cases} \quad (3.59)$$

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 (3.59) becomes

$$u'(t) = -k(u(t) - A)^r. \quad (3.60)$$

Verify that if $r \neq 1$ the solution to (3.60) with $u(0) = u_0 > A$ is given by

$$u(t) = A + ((u_0 - A)^{1-r} + k(r-1)t)^{1/(1-r)} \quad (3.61)$$

- (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 (3.61). 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 a variety of 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 (3.61) 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 square here compare to that of Exercise 3.5.9 (if you did that one)?

■

3.6 Modeling Projects

In this section we offer four modeling opportunities, based on projects from the SIMIODE website [7] and book website [6]. All of these projects involve parameter estimation in conjunction with ODE models, sometimes only one parameter, sometimes several. In most cases this parameter estimation can be done informally (guess and check/plot) or via a least-squares procedure.

3.6.1 Project: Sublimation of Carbon Dioxide

This modeling project is based on the SIMIODE Modeling Scenario “Sublimation of Carbon Dioxide,” [99].

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 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 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 3.20. 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 at given times. This configuration will not allow condensed moisture realized 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.

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 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 3.21.

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

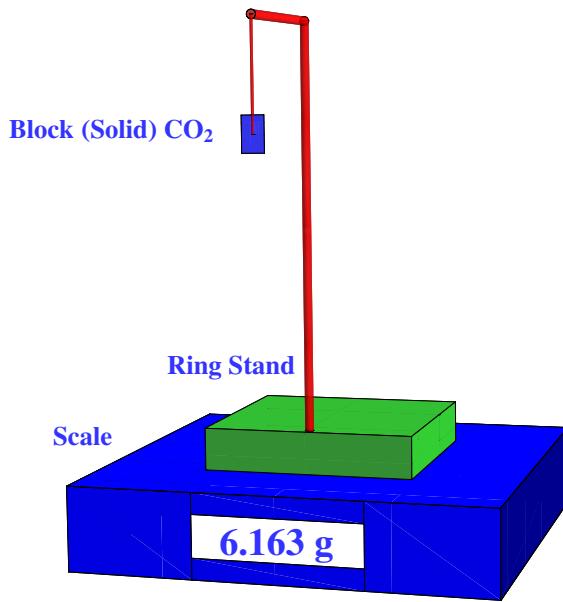


Figure 3.20: Apparatus for measuring the mass of a sublimating block of dry ice.

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).

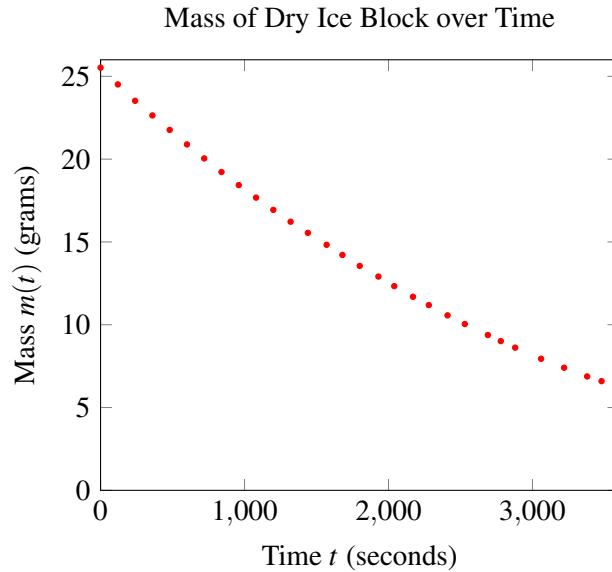


Figure 3.21: Sublimation data from Table 3.9.

models the evolution of $m(t)$ over time, in the form

$$m'(t) = F(t, m(t)) \quad (3.62)$$

with initial condition $m(0) = m_0$.

Modeling Sublimation

Modeling Exercise 1 Consider what form the function $F(t, m)$ in (3.62) should 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 at the surface of the dry ice block—is that relevant? And of course, consider the graph in Figure 3.21. In any case, the ODE you come up with should certainly have at least one “adjustable” constant that can be estimated from the data at hand, either by visual fitting or least-squares.

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.² It might be beneficial, on a first pass, to come with an analytically solvable ODE.

Modeling Exercise 2 Solve the ODE you came up with in Modeling Exercise 1 with an appropriate initial condition. The solution should contain whatever undetermined parameters you introduced in that Modeling Exercise.

Modeling Exercise 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!

3.6.2 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” [40]. We developed an ODE (1.12), reproduced

²Keep It Simple, Stupid.

Year j	u_j	h_j	Year	u_j	h_j	Year	u_j	h_j
0	72,148	0.18847	10	68,702	0.231541	20	20,196	0.189526
1	73,793	0.149741	11	61,191	0.208597	21	25,776	0.170108
2	74,082	0.219209	12	49,599	0.335648	22	23,796	0.156601
3	92,912	0.176781	13	46,266	0.295344	23	19,240	0.281787
4	82,323	0.282033	14	34,877	0.331848	24	16,495	0.252869
5	59,073	0.34528	15	28,827	0.350394	25	12,167	0.255417
6	59,920	0.206545	16	21,980	0.282701	26	21,104	0.081034
7	48,789	0.338185	17	17,463	0.199275	27	18,871	0.0873972
8	70,638	0.147236	18	18,057	0.18781	28	21,241	0.0819517
9	67,462	0.19757	19	22,681	0.193574	29	22,962	0.105181

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 [108].

here for convenience:

$$u'(t) = ru(t) \left(1 - \frac{u(t)}{K}\right) - h(t)u(t). \quad (3.63)$$

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 (3.63) 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 that 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 (3.63).

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 , $0 \leq j \leq n-1$ with data u_j , construct an approximation $u'_j \approx u'(t_j)$ as

$$u'_j = \frac{u_{j+1} - u_j}{\Delta t_j} \quad (3.64)$$

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 a *finite difference* approximation to $u'(t_j)$.

- Form a discrete version of the ODE (3.63) as

$$\frac{u_{j+1} - u_j}{\Delta t_j} = ru_j(1 - u_j/K) - h_j u_j \quad (3.65)$$

for $0 \leq j \leq n-1$, where r and K are to be determined, while the harvest rates h_j are tabulated in Table 3.10. The $u'(t)$ term in (3.64) 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 (3.64) is not likely to hold for any choice of r and K , but a sum of squares function

$$S(r, K) = \sum_{j=0}^{n-1} \left[\frac{u_{j+1} - u_j}{\Delta t_j} - (ru_j(1 - u_j/K) - h_j u_j) \right]^2. \quad (3.66)$$

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 (3.63) in this setting.

Modeling Exercise 1 Using whatever software you have available, form the quantities u'_j for $1 \leq j \leq 30$ using (3.64) with $\Delta t_j = 1$ for all j , and then form the sum of squares function $S(r, K)$ according to (3.66).

Modeling Exercise 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 $\ln(S(r, K))$ instead.

Modeling Exercise 3 Let r^*, K^* denote the least squares estimates for r and K . A function for $h(t)$ isn't given explicitly, so the ODE (3.64) 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 (3.64) numerically using Euler's method with step size 1. Specifically, define $U_0 = u_0 = 72148$ and then define

$$U_{j+1} = U_j + r^* U_j (1 - U_j/K^*) - h_j U_j \quad (3.67)$$

for $j = 0$ to $j = n - 1$. The result is an Euler estimate of the solution to (3.64) using step size 1 with the least-squares estimates for r and K , and using the sampled values of the harvesting rate $h(t)$ at times $t = t_j$.

Plot the "discrete solution" defined by the pairs (t_j, U_j) for $0 \leq j \leq n$ (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?

3.6.3 Project: The Mathematics of Marriage

This modeling project is based on the SIMIODE modeling scenarios "Mathematics of Marriage" [61], "At What Age Do People Get Married?" [95], and a model developed in [52]. We explore the process of entry into marriage by an individual, in particular, the fraction of married people in a given sociological group 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 declines as one gets older.

Modeling Exercise 1 Do these assumptions seem reasonable? Write down several other factors that may effect 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 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

$$P(t) = m(t)/n. \quad (3.68)$$

We will develop a differential equation satisfied by $P(t)$.

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

$$\text{the probability of an individual marrying in time interval } (t, t + dt) = p(t) dt \quad (3.69)$$

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 (3.69), 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

$$\frac{dm}{dt} = (n - m(t))p(t). \quad (3.70)$$

Divide both sides of (3.70) by n and use (3.68) to obtain

$$\frac{dP}{dt} = (1 - P(t))p(t). \quad (3.71)$$

This is an ODE for $P(t)$, but some choice for the function $p(t)$ in (3.69) is needed. This is where assumptions (1) and (2) come into play.

Let us assume that

$$p(t) = q(t)P(t) \quad (3.72)$$

where

$$q(t) = Ab^t \quad (3.73)$$

for constants A and b with $0 < A, b < 1$. 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 initial value $q(0) = A$. Equations (3.72)-(3.73) 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.

Use $q(t)$ as defined in (3.73) in (3.72) to obtain $p(t) = Ab^t P(t)$, and in (3.71) this yields an ODE

$$\frac{dP}{dt} = Ab^t P(t)(1 - P(t)), \quad (3.74)$$

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.

for $P(t)$, the fraction of married persons in the cohort as a function of time. This is the equation to be used 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 2 Review the derivation of (3.74) 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 later.)

Modeling Tip 3 Now that we have a model, the next step is to solve (3.74). 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 3 below.

Modeling Exercise 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 (3.74) satisfies the conditions of the Existence-Uniqueness 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 (3.74) with initial condition $P(0) = 0$ is given by $P(t) = 0$ for all $t > 0$. Verify that the solution with $P(0) = 1$ is given by $P(t) = 1$ 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, with a bit of simplifying algebra and with the assumptions that $0 < A, b, P < 1$, yields solution

$$P(t) = \frac{P_0}{P_0 + e^{-\frac{A(b^t - 1)}{\ln(b)}} (1 - P_0)} \quad (3.75)$$

to (3.74) with $P(0) = P_0$.

Modeling Exercise 4 Verify that $P(t)$ as given by (3.75) satisfies (3.74) with $P(0) = P_0$.

Parameter Estimation and Comparison to Data

In Table 3.11 are some data for two cohorts, U.S. men born between 1940 and 1944, and U.S. men born between 1945 and 1949. This data is from [61] and [64]. 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 $P(0) = P_0 = 0.211$ the solution (3.75) becomes

$$P(t) = \frac{0.211}{0.211 + 0.789e^{-\frac{A(b^t - 1)}{\ln(b)}}}. \quad (3.76)$$

The function $P(t)$ can be used in (3.76) to form a sum of squares for fitting A and b ,

$$S(A, b) = \sum_{j=1}^7 (P(5j - 5) - R_j)^2 \quad (3.77)$$

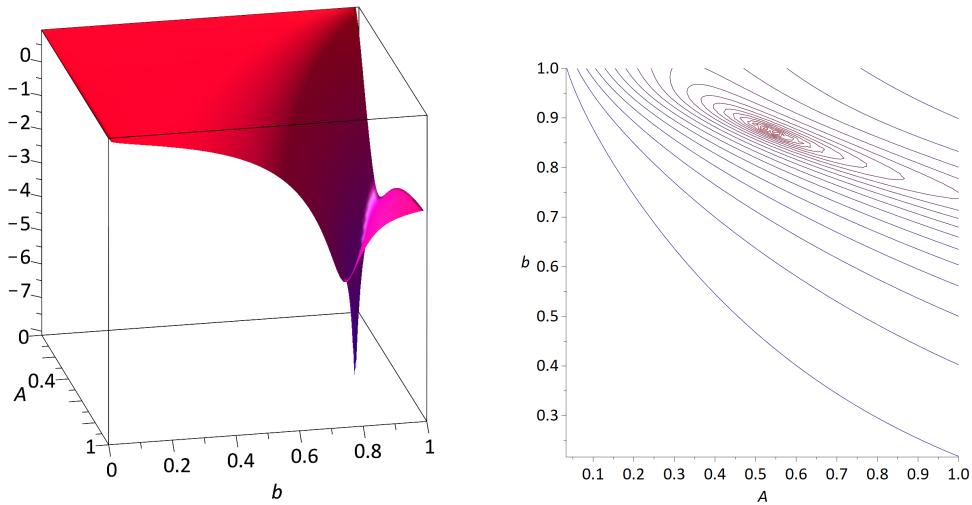


Figure 3.22: Graph (left panel) and contour plot (right panel) for $\ln(S(A,b))$ defined by (3.77).

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. In Figure 3.22 in the left panel is the graph of $\ln(S(A,b))$ on the domain $0 < A, b < 1$, and on the right a contour plot for $\ln(S(A,b))$.

Modeling Exercise 5 Based on Figure 3.22 (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 (3.77) to compute the residual, and in (3.76) to plot the function $P(t)$ that best fits this data. Compare to a plot of the data.

To obtain P_0 in (3.75) we used the initial data point $(0, 0.211)$. This forces the solution $P(t)$ given by (3.75) 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 6 Define a new sum of squares as

$$S(A, b, P_0) = \sum_{j=1}^7 (P(5j - 5) - R_j)^2 \quad (3.78)$$

where $P(t)$ is given by (3.75). 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 4 as initial guesses. Does this change the previous values much? Use these values in (3.78) to compute the residual, and in (3.75) to plot the function $P(t)$ that best fits this data. Compare to a plot of the data. Based on your parameters, what percentage of this cohort will eventually marry?

Modeling Exercise 7 Form and minimize an appropriate sum of squares of the form for men in the 1945-1949 cohort

$$S(A, b, P_0) = \sum_{j=1}^7 (P(5j - 5) - R_j)^2$$

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.

for men in the 1945-1949 cohort. Based on your parameters, what percentage of this cohort will eventually marry?

Modeling Exercise 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?

Note that the procedure we've developed is not mere "curve fitting," but based on the two basic assumptions state above. See [61] for more data and ideas for analyzing this type of data.

3.6.4 Project: Shuttlecocks and Model Selection

This problem is based on the SIMIODE modeling project [101] and extends Exercise 2.2.17.

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 [82]. 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 it 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. That is,

$$F_r = -F_0(|v|)$$

for some function F_0 . The minus sign will be used to incorporate opposition to the direction of motion. Specifically, the interest here is $v > 0$, in which case $|v| = v$. Also, 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 $v > 0$. With these assumptions and the requirement that the force of air resistance

opposes the direction of motion it follows that

$$F_r = -F_0(v). \quad (3.79)$$

Equation (3.79) along with $F_0(v) > 0$ for $v > 0$ and the explicit “minus” sign capture the necessity that F_r is opposed to the object’s motion.

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

$$v'(t) = g - F(v(t)) \quad (3.80)$$

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 leads to the following possibilities:

- (a) $F(v) = 0$, no air resistance.
- (b) $F(v) = kv$ for some constant $k > 0$.
- (c) $F(v) = kv^2$ for some constant $k > 0$.
- (d) $F(v) = k_1v + k_2v^2$ for constants $k_1, k_2 > 0$.
- (e) $F(v) = kv^r$ for some constants with $k > 0, r > 1$.

Let’s consider each possibility in turn. The shuttlecock data was taken in Villanova, PA, altitude 120 meters above sea level, longitude 75.3492 degrees west, latitude 40.0376 degrees north, where $g \approx 9.80136$ meters per second squares, according to the “Local Gravity Calculator” at <https://www.sensorsone.com/local-gravity-calculator/>. Thus gravitational acceleration g in (3.80) will be considered known.

Fitting the Air Resistance Models

Modeling Exercise 1 Solve the ODE (3.80) with $F(v) = 0$ and initial data $v(0) = 0$. Use this to show that the position of the shuttlecock would be given by $x(t) = gt^2/2$ (assuming $x(0) = 0$). Note that there are no parameters to estimate here, since $g = 9.80136$ is known. Compute the sum of squares

$$S = \sum_{j=1}^{17} (x(t_j) - x_j)^2 \quad (3.81)$$

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 2 Consider the ODE (3.80) with $F(v) = kv$ and initial data $v(0) = 0$. Show that in this case the shuttlecock position is given by $x(t) = \frac{g}{k^2}(kt + e^{-kt} - 1)$ (again assuming $x(0) = 0$). In this case the parameter k is to be estimated (we still assume $g = 9.80136$). Form the sum of squares

$$S(k) = \sum_{j=1}^{17} (x(t_j) - x_j)^2 \quad (3.82)$$

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 3 Repeat Modeling Exercise 2 using $F(v) = kv^2$. Note that in this case the ODE (3.80) effectively becomes (2.39) but with k replacing k/m . As such, $v(t)$ is given by (2.52) but with k/m there replaced by just k here. Use this to show that (with $x(0) = 0$)

$$x(t) = \frac{\ln(\cosh(t\sqrt{kg}))}{k}.$$

Consulting Exercise 2.2.17 may be useful. Form the sum of squares and find the optimal value of k . Compute the residual sum of squares.

Modeling Exercise 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 (3.80) with $v(0) = 0$ is given by

$$v(t) = \frac{\alpha \tanh(\alpha t/2 + \operatorname{arctanh}(k_1/\alpha)) - k_1}{2k_2}$$

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 $g = 9.80136$)

$$S(k_1, k_2) = \sum_{j=1}^{17} (x(t_j) - x_j)^2 \quad (3.83)$$

and then minimize S in the variables k_1, k_2 . Confine your attention to $k_1, k_2 \geq 0$; 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, k_2 values along with the data.

The Case $F(v) = v^r$

The case in which $F(v) = v^r$ in the ODE (3.80) yields

$$v'(t) = g - kv^r(t) \quad (3.84)$$

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 (3.84) numerically to compute $v(\tau_i)$ at equispaced times $\tau_i = T/N$, $0 \leq i \leq N$ where $T = 1.873$ (the largest time for which we have position data) and N is reasonably large, e.g., $N = 1000$.
2. Use 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, using any numerical integration rule, e.g., the trapezoidal rule.

The above steps allow us to compute the residual sum of squares $S(k, r)$ for any given 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 [6]. 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 3.23.

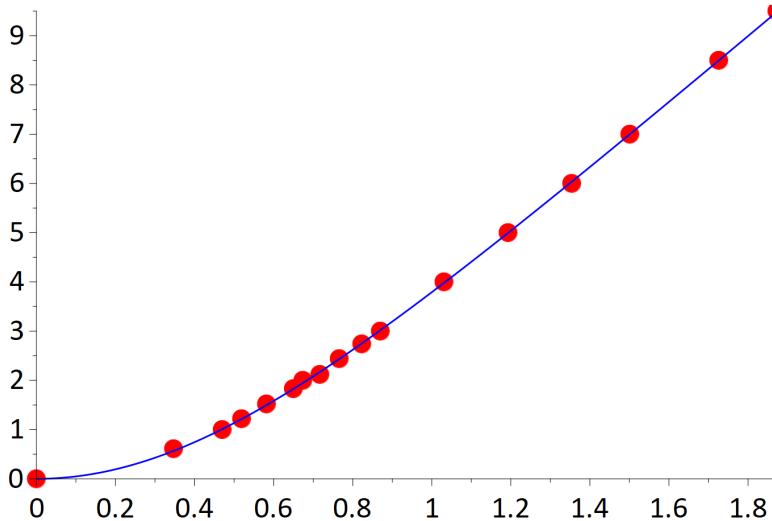


Figure 3.23: Estimated position $x(t)$ from model $v' = g - kv^r$ (blue) and data from Table 2.2 (red), optimal values $k \approx 0.205$ and $r \approx 2.02$.

Model	$F(v) = 0$	$F(v) = kv$	$F(v) = kv^2$	$F(v) = k_1v + k_2v^2$	$F(v) = kv^r$
Residual					1.01×10^{-2}

Table 3.13: Residual sum of squares for each model.

Modeling Exercise 5 Tabulate your answers for models (a) to (d) for the residual sum of squares for each model in Table 3.13, according to which model for air resistance was used. The results for model (e) are already filled in. As a sanity check, make sure the residuals for the models (a) to (e) 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.

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. Moreover, it has a much, much larger residual sum of squares.

The model $F(v) = kv$ is much superior and has a much smaller residual, but again, an examination of $x(t)$ and the data in comparison to the other models and their residuals (the next largest of which is still 50 times smaller!) puts this model out of the running.

But the other three models have residuals 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 the *Akaike Information Criterion* (AIC). Each model above contained a certain number 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 $P = 2$ 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

$$\text{AIC} = 2(P+1) + \frac{2P(P+1)}{N-P-1} + N\ln(\text{RSS}/N) \quad (3.85)$$

where RSS denotes the residual sum of squares after the best least-squares parameters have been found.³ 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 to choose from, the value of the AIC figure of merit is one way to select the “best” model, that 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 relies on certain assumptions being met. An examination of these assumptions would take us too far afield here, but they are reasonable in this setting for this data set. For more information on the AIC see [18].

Modeling Exercise 6 Use the data from Table 3.13 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?

³When the data set is large in comparison to the number of parameters, say $N/P > 40$, one may use $\text{AIC} = 2(P+1) + N\ln(\text{RSS}/N)$.

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 am, an earthquake of magnitude 8.8 struck Chile, one 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, with damage estimated 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. This stands in contrast to a much less powerful magnitude 7.0 earthquake that struck Haiti the previous month, on January 10, 2010, 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 is the strict building code that exists in that country, and its stringent enforcement; see [89]. These building codes were enacted after a massive quake of magnitude 9.5 there in 1960, the strongest earthquake ever recorded. Builders in Chile are held liable for any damages sustained from not adhering to the building codes, for 10 years after the building's construction. The high death toll in Haiti, however, 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 [11].

At the heart of engineering structures that can endure this kind of abuse is a detailed understanding of how mechanical objects vibrate. This 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.

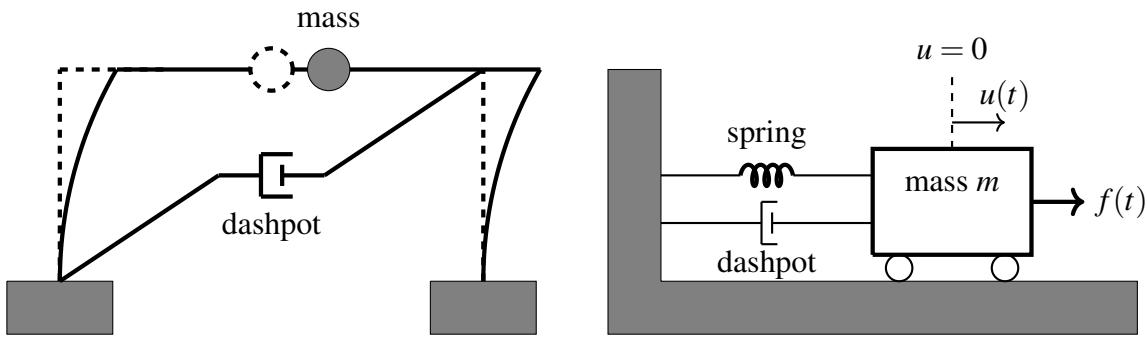


Figure 4.1: Simplified one-story building (left) and mass-spring-damper abstraction (right).

4.1.2 The Harmonic Oscillator

Springs, Dashpots, and Masses

Consider the left panel in Figure 4.1 (adapted from [72]) 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 (e.g., swaying to the right in the left panel of Figure 4.1) the walls exert a force that opposes this displacement, with the magnitude of the force proportional to the magnitude of the displacement. The situation is shown in abstraction 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.

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 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 at which the spring (attached to the vertical wall) exerts no force on the cart, so the spring is at its “natural” 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*¹ and it is quantified as

$$F_{\text{spring}} = -ku \quad (4.1)$$

for some constant $k \geq 0$, known as the *spring constant*. The minus sign indicates the force is opposed to the displacement (recall Modeling Tip 1). Larger values for k embody 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 deformations*. 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 this case.

The dashpot acts as a frictional or damping force of the form

$$F_{\text{damping}} = -cu' \quad (4.2)$$

for some constant $c \geq 0$, so the damping force is always opposed to the direction of motion. Equation (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

¹Hooke first published his law in 1676 as a Latin anagram, and later explicitly as *ut tensio, sic vis* (“as the extension, the force”).

additional “external” time-dependent force $f(t)$ acts on the mass, as indicated in the right panel of Figure 4.1.

Reading Exercise 75 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 75 should be equivalent to

$$mu''(t) + cu'(t) + ku(t) = f(t). \quad (4.3)$$

Equation (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 (4.4) 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

$$mu''(t) + cu'(t) + ku(t) = 0. \quad (4.4)$$

This is the equation of the undriven or *unforced harmonic oscillator*. Equation (4.4) is *homogeneous*, since the right side is zero, while equation (4.3) is the *nonhomogeneous* version of the ODE. 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 76

- (a) Consider (4.5) when $m = 1$ and $k = 1$, 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 .
- (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 77 Suppose that for the undamped spring-mass system of Reading Exercise 76 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 76 to find a function $u(t)$ that satisfies both (4.5) and these initial conditions.

Remark 7 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 78 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 76-78, 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. The necessity of specifying both initial position and initial velocity for second order ODE's was encountered previously, 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. Suppose the rider has a mass of 80 kg, the bike (less wheels) has a mass of 12 kg, and that half of this mass is supported by the front shock absorber, so the an 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. Equation (4.6) is nonhomogeneous and of the form (4.3). ■

Reading Exercise 79 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} . Substitute $u(t) = u_{eq}$ into the ODE (4.6), noting that $u'(t) = u''(t) = 0$ in this case, to solve for u_{eq} . How far does the shock 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 a large area of mechanical engineering. The goal 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. This is 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 Table Shakedown” in Section 5.7. Other techniques for vibration isolation are passive and consist of coil springs, “air springs,” dashpots, rubber pads, etc.

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 blue cylindrical support on the left in Figure 4.3 or spring-dashpot counterpart in the right panel. This leg supports the tabletop of mass m , the gray 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 it’s

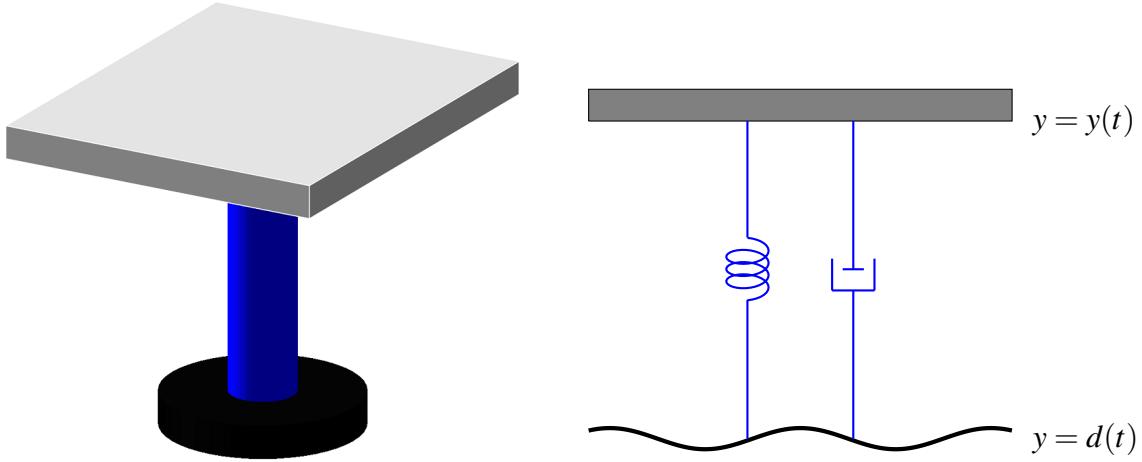


Figure 4.3: Vibration isolation table (left panel) and spring-mass-damper model (right panel).

easy to see that the length $L(t)$ of the spring at time t is given by

$$L(t) = y(t) - d(t). \quad (4.7)$$

The displacement $u(t)$ of the spring from its natural length is then

$$u(t) = L(t) - L_0 = y(t) - d(t) - L_0. \quad (4.8)$$

The force of the spring on the mass m is $F_{spring} = -ku(t)$ or

$$F_{spring} = -ku(t) = -k(y(t) - d(t) - L_0). \quad (4.9)$$

Assume the damping force as 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

$$F_{damping} = -cL'(t) = -c(y'(t) - d'(t)). \quad (4.10)$$

The vertical acceleration of the tabletop mass is given by $y''(t)$. If gravity is the only other force acting on the tabletop then from Newton's Second Law of Motion

$$\begin{aligned} my''(t) &= F_{spring} + F_{damping} + F_{gravity} \\ &= -k(y(t) - d(t) - L_0) - c(y'(t) - d'(t)) - mg \end{aligned} \quad (4.11)$$

with $g > 0$ (hence the explicit minus sign in front of mg on the right in (4.11)). Equation (4.11) can be rearranged to

$$my''(t) + cy'(t) + ky(t) = k(d(t) + L_0) + cd'(t) - mg, \quad (4.12)$$

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 (4.12) are known or given. This equation fits the mold of (4.3). ■

Reading Exercise 80 Suppose $d(t) = 0$ (level ground). Find an equilibrium solution $y(t) = y_{eq}$ to (4.12). What is the physical interpretation of this solution?

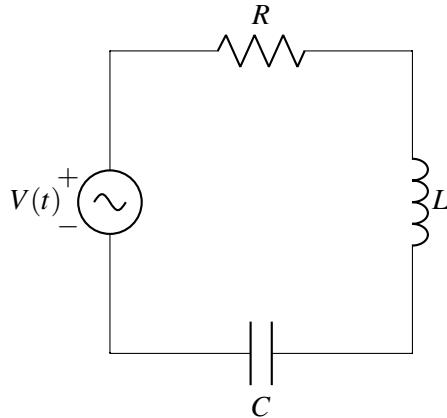


Figure 4.4: Single loop RLC series circuit.

RLC Circuits

■ **Example 4.3** Figure 4.4 shows a single loop “RLC” circuit that contains a voltage source $V(t)$, a resistor R , inductor L , and 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 the inductor 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*.

Reasoning similar to that of Example 2.5 can be used to find an ODE that governs the charge $q(t)$ on the capacitor at any time, and from this 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 voltages changes must sum to zero so that

$$V(t) - RI(t) - LI'(t) - q(t)/C = 0. \quad (4.13)$$

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 (4.13) can be written as

$$Lq''(t) + Rq'(t) + q(t)/C = V(t). \quad (4.14)$$

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 $I(0) = 0$, e.g., if a switch in the circuit was closed at time $t = 0$).

Note the similarity of (4.14) and the spring-mass-damper equation (4.3). 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 81 Consider (4.14) in the case that $R = 0$ and $V(t) = \cos(\omega t)$. Verify that

$$q(t) = \frac{C}{LC\omega^2 - 1} \cos(\omega t)$$

is a solution to (4.14) in this case. Note that $q(t)$ is periodic and at 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 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. ■

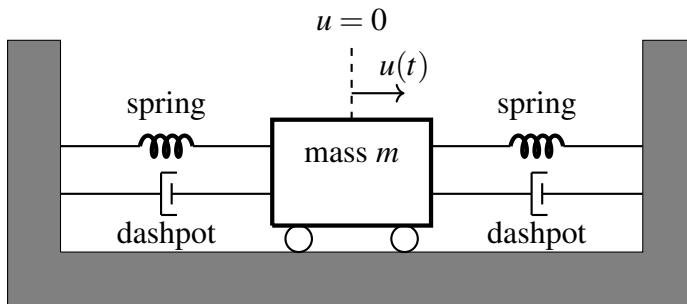


Figure 4.5: Mass acted upon by two springs and dashpots.

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 equation (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 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 as 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 $m = 5000 \text{ kg}$. The walls exert a restoring force modeled by a spring constant $k = 5 \times 10^5 \text{ newtons per meter}$, and a damping constant $c = 2 \times 10^4 \text{ newtons per meter per second}$.

- 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 $u(0) = 0.01 \text{ meters}$ 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 $0 \leq t \leq 5 \text{ seconds}$.

- What is the period of the building's (damped) motion? According to [19], 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 does this correspond to? According to [19], 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 with $u(0) = 0.01, u'(0) = 0$, of the form $u(t) = u_0 \cos(\omega t)$, 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 that figure the water surface is indicated by the light blue horizontal plane. We use $y(t)$ to indicate the depth of the bottom of the buoy 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 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 the water which 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!
- Use Newton's Second Law of Motion to show that $y(t)$ satisfies the constant coefficient,

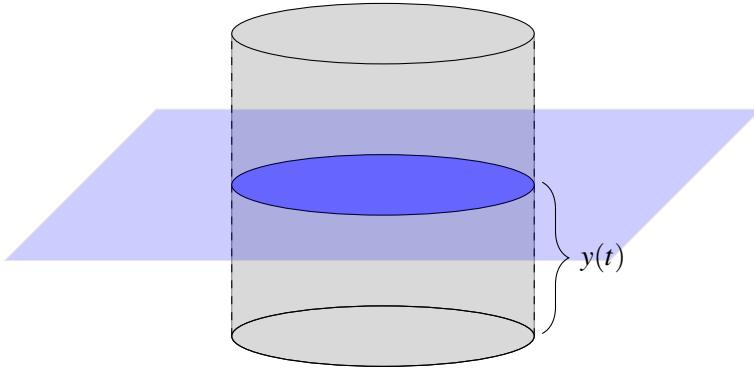


Figure 4.6: Cylindrical buoy with cross-sectional area A ; light blue plane indicates water surface.

nonhomogeneous, second order ODE

$$y''(t) + \frac{\rho g A}{m} y(t) = -g. \quad (4.15)$$

■

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 (4.14). 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.

Newton's Universal Law of Gravitation states that the force of gravitational attraction between point masses m and M is given by

$$F(r) = \frac{GMm}{r^2} \quad (4.16)$$

where r is the distance between the objects and $G \approx 6.674 \times 10^{-11}$ newtons-square meters per kg squared is the gravitational constant. This law also applies to spherical masses of uniform density, where r is the distance between the bodies' center of mass.

Let M denote the mass of the earth, $M \approx 5.972 \times 10^{24}$ kg, and $m \ll M$ the mass of some object, e.g., a satellite. Suppose this object moves on a trajectory, relative to the earth, that is purely radial. That is, the object's position is purely a function of r .

- (a) Suppose an object is launched from the earth surface, $r = R \approx 6.37 \times 10^6$ meters from the center of the earth, with initial radial velocity v_0 , and 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 gravity is the only force acting on the object then $r(t)$ satisfies

$$r''(t) = -\frac{GM}{r^2(t)}. \quad (4.17)$$

What are the initial conditions for this second order ODE?

- (b) Despite its apparent simplicity, the ODE (4.17) is not solvable in any simple analytical form. Try some numerical experiments: solve the ODE numerically with whatever software you have available, with a variety of initial velocities v_0 , say $v_0 = 100, 1000, 10000, 100000$ meters per second. Do the solutions behave as expected?
- (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. Multiply both sides of (4.17) by $r'(t)$ and integrate from $t = 0$ to $t = T$ (here $T > 0$ is some unspecified time.) Show that this yields

$$\frac{1}{2}((r'(T))^2 - (r'(0))^2) = GM \left(\frac{1}{r(T)} - \frac{1}{R} \right)$$

or (multiply through by the object’s mass m)

$$\frac{m(r'(T))^2}{2} + GMm \left(\frac{1}{R} - \frac{1}{r(T)} \right) = \frac{mv_0^2}{2} \quad (4.18)$$

We can interpret the right side of (4.18) as the total energy of the object when it is launched, if we take $r = R$ as the normalization for zero potential energy: $m(r'(0))^2/2$ or $mv_0^2/2$ is its kinetic energy, and it has zero potential energy when $r = R$. Show that the left side of (4.18) can be interpreted as the total energy of the object at time $t = T$, kinetic plus potential.

- (d) Suppose the object is launched with v_0 so $r'(T) \rightarrow 0$ and $r(T) \rightarrow \infty$ as $T \rightarrow \infty$ (so the object has just enough velocity at launch to escape earth’s gravity). 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), the effect of damping on the solution both mathematically and physically, and 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

$$u''(t) + 4u'(t) + 3u(t) = 0. \quad (4.19)$$

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 $u'(0) = 0$. A solution might be expected to look like 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 76-78, 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?

Sines and cosines are really exponentials in disguise! If you’ve studied complex arithmetic you’ve seen this, in the form of Euler’s identity, which states that $e^{i\theta} = \cos(\theta) + i\sin(\theta)$. If not,

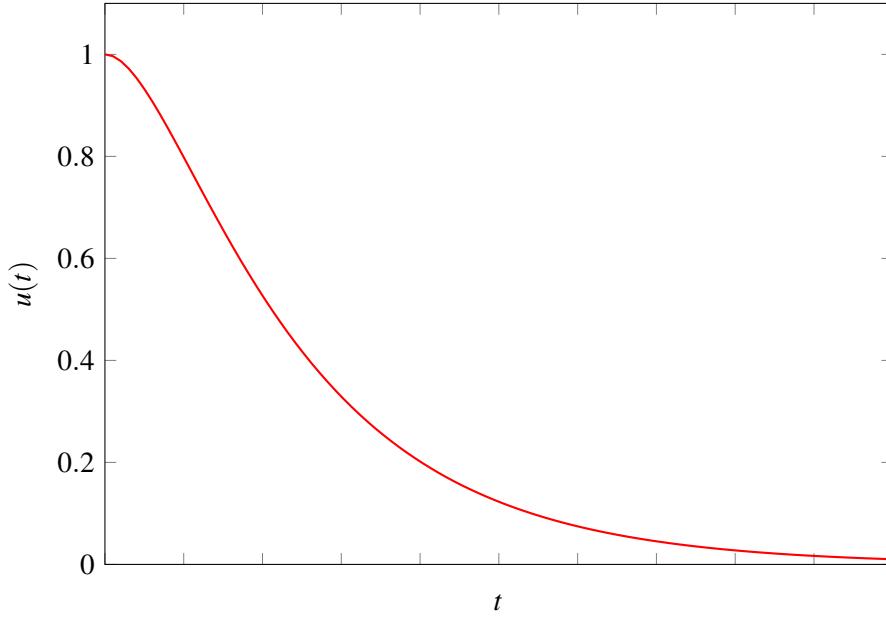


Figure 4.7: Motion $u(t)$ of a heavily damped spring-mass system.

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 (4.19) at hand we will seek a solution of the form $u(t) = e^{rt}$ for an appropriate choice of r , by substituting $u(t) = e^{rt}$ into (4.19) and seeing what falls out. The basic rules for differentiation yield

$$\begin{aligned} u(t) &= e^{rt}, \\ u'(t) &= re^{rt}, \\ u''(t) &= r^2 e^{rt}. \end{aligned}$$

Insert this information into (4.19) and collect terms to obtain

$$e^{rt}(r^2 + 4r + 3) = 0. \quad (4.20)$$

Equation (4.20) must be satisfied identically in t , that is, the left side must be the zero function. Note that e^{rt} is never 0, so divide both sides of (4.20) by e^{rt} to obtain

$$r^2 + 4r + 3 = 0. \quad (4.21)$$

This is equivalent to (4.20), and is necessary and sufficient for e^{rt} to satisfy (4.19). Equation (4.21) shows that r must be a root of $r^2 + 4r + 3 = 0$, a quadratic equation.

Equation (4.21) is called the *characteristic equation* or *auxiliary equation* for (4.19). The roots to (4.21) 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

$$u(t) = e^{-t} \quad \text{and} \quad u(t) = e^{-3t} \quad (4.22)$$

provides a solution to (4.19).

Remark 8 In the analysis above we guessed that there might be solutions to (4.19) of the very specific form $u(t) = e^{rt}$, then successfully adjusted r to make this 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 82 Look up the definition of “*ansatz*.” What is the literal translation from German?

Linearity and Superposition

Equation (4.22) provides two distinct solutions to (4.19). These solutions can be used to construct infinitely many more solutions; the key is to use the linearity of (4.19), an indispensable asset here. Specifically, if $u_1(t)$ and $u_2(t)$ are any solutions to (4.19) then any linear combination of the form

$$u(t) = c_1 u_1(t) + c_2 u_2(t) \quad (4.23)$$

is also a solution. This is easy to verify: with u as in (4.23), a bit of algebra shows that

$$\begin{aligned} u'' + 4u' + 3u &= (c_1 u_1 + c_2 u_2)'' + 4(c_1 u_1 + c_2 u_2)' + 3(c_1 u_1 + c_2 u_2) \\ &= c_1 \underbrace{(u_1'' + 4u_1' + 3u_1)}_0 + c_2 \underbrace{(u_2'' + 4u_2' + 3u_2)}_0 \\ &= 0, \end{aligned} \quad (4.24)$$

so $u(t)$ also satisfies (4.19), for any choice of c_1 and c_2 . The computation leading to (4.24) is an example of the *principle of superposition* or simply, *superposition*, in which an arbitrary linear combinations of solutions to a linear ODE are again a solution. This principle of superposition is one of the most fundamental tools available for analyzing linear ODE's.

Constructing A General Solution and Obtaining Initial Data

With $u_1(t) = e^{-t}$ and $u_2(t) = e^{-3t}$ from (4.22), the principle of superposition shows that any function $u(t)$ of the form

$$u(t) = c_1 e^{-t} + c_2 e^{-3t} \quad (4.25)$$

satisfies (4.19), for any choice of constants c_1 and c_2 . The function $u(t)$ defined by (4.25) is called a *general solution* to (4.19). We say “a” general solution rather than “the” general solution, since a general solution may assume different forms; recall Remark 2.

The function $u(t)$ in (4.25) is called a “general solution” for good reason: any solution to (4.19) is uniquely determined by initial data $u(0) = u_0, u'(0) = v_0$ (more generally, $u(t_0) = u_0, u'(t_0) = v_0$) and any such initial data can be obtained from (4.25) by choosing c_1 and c_2 appropriately. To illustrate, note that $u(0) = u_0, u'(0) = v_0$ yields 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 above two equations $c_1 + c_2 = u_0, -c_1 - 3c_2 = v_0$ 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 (4.19) with initial conditions $u(0) = 1, u'(0) = 0$. With a general solution of the form (4.25), $u(0) = 1$ 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 $c_1 = 3/2, c_2 = -1/2$, and so the solution to (4.19) 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 $0 \leq t \leq 5$. ■

Reading Exercise 83 Find the solution to (4.19) with initial conditions $u(0) = 2, u'(0) = 4$. 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, spring constant $k = 40000$ newtons per meter. Let's compute the displacement of the roof mass if a sudden shock/impact/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

$$10000u''(t) + 50000u'(t) + 40000u(t) = 0 \quad (4.26)$$

with initial conditions $u(0) = 0, u'(0) = 1/4$.

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 with those of (4.26). 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 (4.24) (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.

To obtain $u(0) = 0$ requires $c_1 + c_2 = 0$, while $u'(t) = -2c_1e^{-2t} - 4c_2e^{-4t}$ yields $u'(0) = -2c_1 - 4c_2 = 1/4$. The simultaneous solution to $c_1 + c_2 = 0$ and $-2c_1 - 4c_2 = 1/2$ is $c_1 = 1/12, c_2 = -1/12$. 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 84 The initial conditions for an ODE need not be given at time $t = 0$, though this is common. Solve the ODE (4.19) using the general solution (4.25), but with initial conditions $u(1) = -1$ and $u'(1) = 6$.

4.2.2 Solving Second Order Linear ODE's: The General Case

The solution procedure in the previous section illustrates the typical method for solving ODE's of the form (4.4) with initial conditions $u(0) = u_0, u'(0) = v_0$. The goal is to first find two "independent" solutions $u_1(t)$ and $u_2(t)$ and then use these to construct a general solution, which we now define a bit more formally.

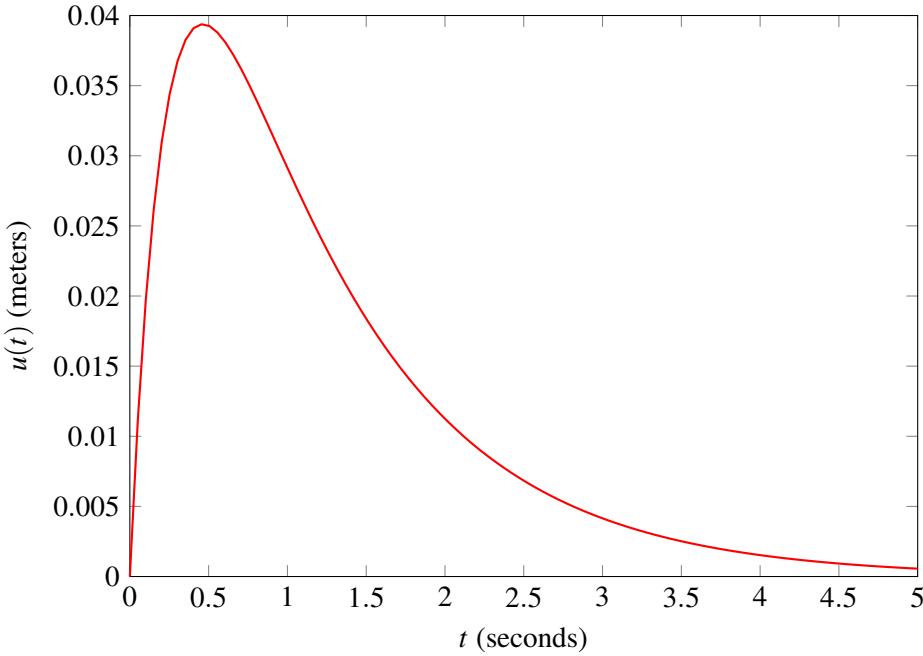


Figure 4.8: Graph of solution to $10000u''(t) + 50000u'(t) + 40000u(t) = 0$ with $u(0) = 0, u'(0) = 1/4$.

Definition 4.2.1 A *general solution* $u(t)$ to $mu''(t) + cu'(t) + ku(t) = 0$ is a function of the form

$$u(t) = c_1u_1(t) + c_2u_2(t) \quad (4.27)$$

where u_1 and u_2 both satisfy $mu''(t) + cu'(t) + ku(t) = 0$, with the property that initial conditions $u(t_0) = u_0, u'(t_0) = v_0$ can be obtained with a suitable choice for c_1 and c_2 , for any initial time t_0 and any u_0 and v_0 . That is, 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 solvable for c_1 and c_2 , for any initial time t_0 and any u_0 and v_0 .

The functions $u_1(t)$ and $u_2(t)$ in the general solution (4.27) are called *basis functions* or a *basis* for the space 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 can be constructed via superposition. As suggested by Reading Exercise 84, if a general solution of the form (4.27) works at a specific time $t = t_0$ then this general solution works for any other initial time, though this will not be proved at this time. It will be clear for the various ODE's we encounter, however.

Reading Exercise 85 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 (4.24).

Reading Exercise 86 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 a basis for the set of solutions to the ODE. Hint: Show that, for given initial data u_0, 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, v_0 happen to satisfy $u'_2(t_0)u_0 = u_2(t_0)v_0$.

The result of Reading Exercise 86 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 ODE's

To find a general solution to $mu'' + cu' + ku = 0$:

1. Find the roots r_1 and r_2 to the characteristic equation

$$mr^2 + cr + k = 0. \quad (4.28)$$

This stems 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, given by

$$\begin{aligned} r_1 &= -\frac{c}{2m} + \frac{\sqrt{c^2 - 4mk}}{2m}, \\ r_2 &= -\frac{c}{2m} - \frac{\sqrt{c^2 - 4mk}}{2m}. \end{aligned} \quad (4.29)$$

2. If $r_1 \neq r_2$, form a *general solution*

$$u(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t}. \quad (4.30)$$

The function $u(t)$ is a solution to (4.4) for any choice of c_1 and c_2 , as shown in Reading Exercise 85. If $r_1 = r_2$ (so the characteristic equation has a double root, meaning $c^2 - 4mk = 0$) then (4.30) is not a general solution; this special case will be considered shortly.

3. The initial conditions $u(0) = u_0$ and $u'(0) = v_0$ lead to equations

$$\begin{aligned} c_1 + c_2 &= u_0, \\ r_1 c_1 + r_2 c_2 &= v_0. \end{aligned}$$

The unique solution to these two equations is

$$\begin{aligned} c_1 &= \frac{v_0 - r_2 u_0}{r_1 - r_2}, \\ c_2 &= \frac{r_1 u_0 - v_0}{r_1 - r_2} \end{aligned} \quad (4.31)$$

if $r_1 \neq r_2$. In this case the solution to $mu'' + cu' + ku = 0$ with initial data $u(0) = u_0, u'(0) = v_0$ 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

Examination of the procedure embodied by (4.28)-(4.31) for solving $mu'' + cu' + ku = 0$ reveals that the procedure works perfectly well when m, c , and k are any real or complex numbers, as long as the roots to the characteristic equation are distinct, since the algebraic operations involved are perfectly legitimate for complex-valued functions.

However, essentially all of our work will focus on the physically relevant situation in which m, c , and k are real with $m, k > 0$ and $c \geq 0$. In this case the problem breaks into a few mathematically and physically distinct cases, depending on the nature of the roots (4.29) of 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 (4.29). This was the case in both Examples 4.4 and 4.5.

Moreover, both roots in this case must be negative. To see why, note that r_2 in (4.29) 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 $c^2 > c^2 - 4mk$ (true since both m and k are positive). Then since $c^2 - 4mk > 0$ it follows that

$$c^2 > c^2 - 4mk \implies c > \sqrt{c^2 - 4mk} \implies -c < -\sqrt{c^2 - 4mk} \implies -c + \sqrt{c^2 - 4mk} < 0,$$

so the numerator of r_1 in (4.29) is negative, hence $r_1 < 0$.

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 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, conjugate to each other. As shown in the next section, solutions in this case oscillate and decay to zero.
- **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 and given by $\pm\sqrt{-4mk}/(2m)$. Solutions in this case oscillate periodically and never stop.
- **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 $r_1 = r_2 = -c/(2m)$; both roots are real since c and m are real.

Although the solution procedure (4.28)-(4.31) works in the underdamped and undamped settings, there is mathematical and physical insight to be gained by a more careful examination of the these cases.

4.2.3 The Underdamped and Undamped Cases

As mentioned in the last section, the solution procedure (4.28)-(4.31) 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 76. The characteristic equation (4.28) here is

$$r^2 + 1 = 0$$

with roots $r = i$ and $r = -i$, where $i = \sqrt{-1}$. As dictated by (4.30) a general solution in this case is

$$u(t) = c_1 e^{it} + c_2 e^{-it} \tag{4.32}$$

where $e^{it} = \cos(t) + i\sin(t)$ and $e^{-it} = \cos(t) - i\sin(t)$ from Euler's identity. See Appendix A for more on Euler's identity.

As a specific example let's consider initial conditions $u(0) = 1$ and $u'(0) = 0$. From the general solution (4.32) 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 $u'(0) = i(c_1 - c_2)$, so that $u'(0) = i(c_1 - c_2) = 0$. The initial conditions then dictate $c_1 + c_2 = 1$ and $i(c_1 - c_2) = 0$. The solution to these two equations is $c_1 = c_2 = 1/2$, and from (4.32) 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 $u(0) = 1, u'(0) = 0$ contain no complex numbers; why are there i 's in the solution?

The answer is, there aren't! Invoking Euler's identity 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 87 Verify that $u(t) = \cos(t)$ does in fact satisfy $u'' + u = 0$ 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/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

$$10000u''(t) + 20000u'(t) + 40000u(t) = 0 \quad (4.33)$$

with initial conditions $u(0) = 0, u'(0) = 1/4$.

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 roots 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 conjugates. Both $e^{r_1 t}$ and $e^{r_2 t}$ are solutions to the ODE, and the linearity of the ODE (4.33) and the principle of superposition yields a general solution

$$u(t) = c_1 e^{(-1+i\sqrt{3})t} + c_2 e^{(-1-i\sqrt{3})t}.$$

To obtain the initial condition $u(0) = 0$ requires $c_1 + c_2 = 0$, while $u'(t) = c_1 r_1 e^{r_1 t} + c_2 r_2 e^{r_2 t}$ yields $u'(0) = r_1 c_1 + r_2 c_2 = 1/4$ 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 is

$$u(t) = -\frac{i\sqrt{3}}{24}e^{(-1+i\sqrt{3})t} + \frac{i\sqrt{3}}{24}e^{(-1-i\sqrt{3})t}. \quad (4.34)$$

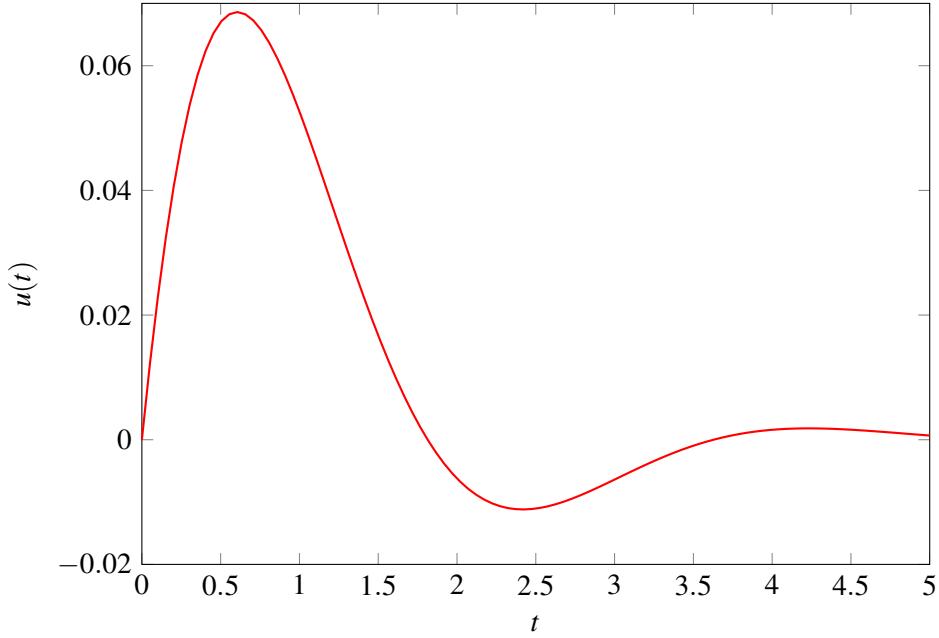


Figure 4.9: Graph of solution to $10000u''(t) + 20000u'(t) + 40000u(t) = 0$ with $u(0) = 0, u'(0) = 1/4$.

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 $u(t)$, as the position of the roof mass, should be a real-valued function of t .

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 identity 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 is in fact real-valued and given by $u(t) = \frac{\sqrt{3}}{12}e^{-t}\sin(t\sqrt{3})$.

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. ■

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^{-1t} and $\sin(t\sqrt{3}), \cos(t\sqrt{3})$. This is no coincidence, as will soon be shown.

4.2.4 The General Underdamped Case

Complex Roots

Examples 4.6 and 4.7 illustrate that the solution procedure (4.28)-(4.31) still works if the roots to the characteristic equation are complex, and despite all the complex-valued arithmetic in those examples, the solutions turn out to be real-valued. Let's examine the situation more generally.

Consider the harmonic oscillator ODE (4.4), $mu'' + cu' + ku = 0$. The characteristic equation is $mr^2 + cr + k = 0$ with roots given by (4.29). 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 (4.29) 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

$$\begin{aligned} \alpha &= \frac{c}{2m}, \\ \omega &= \frac{\sqrt{4mk - c^2}}{2m}. \end{aligned} \tag{4.35}$$

Both α and ω are real-valued quantities with $\omega > 0$ and, since it was assumed $c \geq 0$, $\alpha \geq 0$. 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 (4.29) can then be expressed as

$$r_1 = -\alpha + i\omega \quad \text{and} \quad r_2 = -\alpha - i\omega, \tag{4.36}$$

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

$$u(t) = c_1 e^{(-\alpha+i\omega)t} + c_2 e^{(-\alpha-i\omega)t}. \tag{4.37}$$

Any initial conditions can be obtained using (4.37) with a suitable (and unique) choice for c_1 and c_2 , as given in (4.31).

However, with a little additional work most of the complex arithmetic can be avoided.

A Real-Valued General Solution

Let's write (4.37) in a slightly more insightful form by using Euler's identity 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 solution. 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 (4.37) as

$$\begin{aligned} u(t) &= c_1 e^{(-\alpha+i\omega)t} + c_2 e^{(-\alpha-i\omega)t} \\ &= c_1 e^{-\alpha t} \cos(\omega t) + c_1 i e^{-\alpha t} \sin(\omega t) + c_2 e^{-\alpha t} \cos(\omega t) - c_2 i e^{-\alpha t} \sin(\omega t) \\ &= \underbrace{(c_1 + c_2)}_{d_1} e^{-\alpha t} \cos(\omega t) + \underbrace{i(c_1 - c_2)}_{d_2} e^{-\alpha t} \sin(\omega t) \end{aligned} \quad (4.38)$$

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 (4.38) can also be expressed as

$$u(t) = d_1 e^{-\alpha t} \cos(\omega t) + d_2 e^{-\alpha t} \sin(\omega t). \quad (4.39)$$

Equation (4.39) in conjunction with (4.35) provides an alternate form of a general solution to $mu'' + cu' + ku = 0$, since any solution that can be expressed using (4.30) or (4.37) can be expressed using (4.39) 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 (4.39) can also be expressed via (4.30) 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 (4.39) 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 (4.30) and (4.39) 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

$$u(t) = \left(\frac{1}{2} - \frac{i}{2}\right) e^{(-1+3i)t} + \left(\frac{1}{2} + \frac{i}{2}\right) e^{(-1-3i)t}. \quad (4.40)$$

But if we instead work with (4.39) and (4.35) 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 plus or minus the imaginary part. Then (4.39) 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 solution

$$u(t) = e^{-t} \cos(3t) + e^{-t} \sin(3t). \quad (4.41)$$

Applying Euler's identity to (4.40) yields exactly (4.41). ■

Reading Exercise 88 Write out a complex-valued general solution (4.30) to $3u''(t) + 18u'(t) + 75u(t) = 0$ and use it to find the solution with $u(0) = 0, u'(0) = 4$. Repeat using a general solution in the form (4.39). Verify that the two specific solutions with these initial conditions are, in fact, the same.

Equation (4.39) provides a general solution to $mu'' + cu' + ku = 0$ that does not involve any complex numbers. Even more importantly, it gives the following insights, in the case that the roots of the characteristic equation are complex.

Observations on the Real-Valued Solution (4.39)

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 damping) then $-\alpha = -\frac{c}{2m} < 0$ and the solution decays in time due to the presence of $e^{-\alpha t}$ in (4.39).
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/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 (4.35) 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, in which 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 has a double root at $r = -2$. Using (4.30) 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} , cheap copies of each other. Using this "general" solution to obtain initial conditions $u(0) = u_0, u'(0) = v_0$ produces 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 what is needed is 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

$$u_2(t) = c(t)e^{-2t}. \quad (4.42)$$

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 (4.42) provides a solution to $u''(t) + 4u'(t) + 4u(t) = 0$ if $c''(t) = 0$, or $c(t) = At + B$ 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 $A = 1, B = 0$ to construct solution $u_2(t) = te^{-2t}$.

Reading Exercise 89 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

$$u(t) = c_1u_1(t) + c_2u_2(t) = c_1e^{-2t} + c_2te^{-2t} \quad (4.43)$$

provides a general solution to $u''(t) + 4u'(t) + 4u(t) = 0$ that can be used to obtain any 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 (4.43) dictates $c_1 = u_0$, since $u_2(0) = 0$. Use (4.43) 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 using (4.43), so (4.43) is a truly general solution.

■ **Example 4.11** Let us solve the ODE $u''(t) + 4u'(t) + 4u(t) = 0$ of Example 4.10, with initial conditions $u(0) = 1, u'(0) = -1$. A general solution to this ODE is given by (4.43) and the initial conditions require $u(0) = c_1 = 1, u'(0) = -2c_1 + c_2 = -1$ with solution $c_1 = 1, c_2 = 1$. 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 above 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

$$mu''(t) + 2m\alpha u'(t) + m\alpha^2 u(t) = 0. \quad (4.44)$$

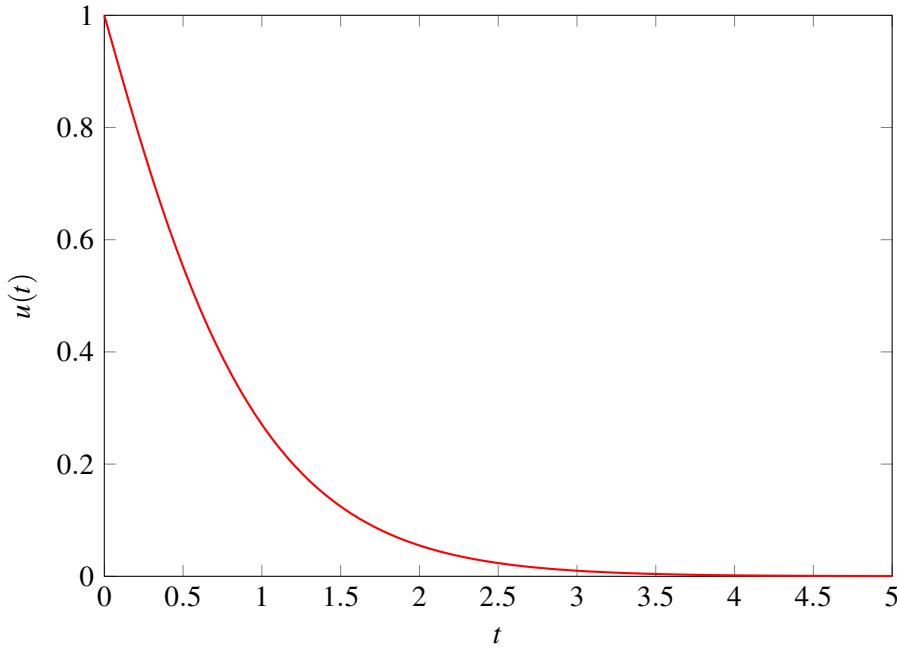


Figure 4.10: Graph of solution to $u''(t) + 4u'(t) + 4u(t) = 0$ with $u(0) = 1, u'(0) = -1$.

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 (4.44) 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 $A = 1, B = 0$ 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

$$u(t) = c_1 e^{-\alpha t} + c_2 t e^{-\alpha t} \quad (4.45)$$

is a general solution to the ODE: any initial conditions can be obtained with an appropriate choice of c_1 and c_2 .

Reading Exercise 90 Verify that inserting $u_2(t) = c(t)e^{-\alpha t}$ into (4.44) yields $me^{-\alpha t}c''(t) = 0$. Then check that $u(t)$ as given by (4.45) can be used to obtain $u(0) = u_0, u'(0) = v_0$, 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 that it can be shown that

Theorem 4.2.1 If $m \neq 0$ then the ODE

$$mu''(t) + cu'(t) + ku(t) = 0$$

with initial conditions $u(0) = u_0, u'(0) = v_0$ 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 [86].

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, critically damped—we produced a general solution as in Definition 4.2.1 from which any initial conditions $u(0) = u_0, u'(0) = v_0$ 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 $m, k > 0$ and $c \geq 0$ there are four distinct cases that have been encountered. 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 (4.30). Any solution decays exponentially in time, and crosses the horizontal axis at most once.
- **The Underdamped Case:** This occurs when $c^2 - 4mk < 0$. In this case both roots r_1 and r_2 of the characteristic equation are complex numbers, conjugate to each other. If $c > 0$ these roots have negative real part. A general solution is given by (4.30), but a real-valued general solution can also be written in the form (4.39) with (4.35). The solution contains sines and cosines, and if $c > 0$ the amplitudes decay 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 (4.39) 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

$$P = 2\pi\sqrt{m/k}. \quad (4.46)$$

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 $m, k > 0$ 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, 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.

- $m = 3, c = 24, k = 60$
- $m = 1, c = 0, k = 20$
- $m = 2, c = 12, k = 10$
- $m = 2, c = 16, k = 64$
- $m = 2, c = 4, k = 10$
- $m = 3, c = 21, k = 36$
- $m = 2, c = 12, k = 18$

- (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, 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 $u(0) = 2, u'(0) = 3$. 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, 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 (4.39). Use each general solution to find a solution with initial conditions $u(0) = 2, u'(0) = 4$, 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, 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 (4.45). Use the general solution to find a solution with initial conditions $u(0) = 2, u'(0) = 4$, 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 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 (4.12) 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 the table top. 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 takes 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 4.31 in Section 4.6, where we derive the ODE

$$\theta''(t) + \frac{g}{L}\theta(t) = 0$$

that the function $\theta(t)$ satisfies approximately, 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?
- Find the general solution to $\theta''(t) + \frac{g}{L}\theta(t) = 0$.
- 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

$$u(t) = A \cos(\omega t) + B \sin(\omega t) \quad (4.47)$$

for constants A and B . However, it is always possible and sometimes useful to exhibit the solution in the form

$$u(t) = C \sin(\omega t + \phi). \quad (4.48)$$

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 (4.48) with $x = \omega t$ and $y = \phi$, then compare the result to the right side of (4.47) to show that these expressions will be identical as functions of t if

$$C \sin(\phi) = A \quad \text{and} \quad C \cos(\phi) = B. \quad (4.49)$$

Thus if we are given $u(t)$ in the form (4.48), we can express $u(t)$ in the form (4.47).

- (b) Use equations (4.49) to show that if given A and B we can solve for C as

$$C = \sqrt{A^2 + B^2}.$$

Note $C \geq 0$.

- (c) Use equations (4.49) to show that if 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 case 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 . Suggested 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 $c_2 = r_1/(r_1 - r_2)$, and then use these values for c_1 and c_2 in $u(t)$ and then adjust r_1 and r_2 to obtain a good fit to the data (perhaps 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 $u(0) = u_0, u'(0) = v_0$ the solution we obtain from the complex-valued general solution (4.30) 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-conjugate.
- (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 (4.36) 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 (one mole is the number 6.02×10^{23}) n of an ideal gas in a closed container satisfies $PV = nRT$ where R is the universal gas constant, $R \approx 8.3145$ joules per degree per mole. 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 the only forces acting on the piston are gravity and the gas pressure from inside the cylinder (so in particular, 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 of Motion 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 (gas pressure and gravity 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 $F = ma$ to show that $u(t)$ obeys the second order differential equation

$$mu''(t) = -mg + \frac{nRT}{u(t) + y_{eq}}. \quad (4.50)$$

Note that (4.50) 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 DE from part(c) using initial conditions $u(0) = 0.2$ meters, $u'(0) = 0$, 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 (4.50) can be approximated by the linear second order ODE

$$u''(t) + \frac{mg^2}{nRT}u(t) = 0. \quad (4.51)$$

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 $u(0) = 0.2$, $u'(0) = 0$ for $0 \leq t \leq 20$ and compare to your answer in part (c). Can you see any difference in the solution to (4.50) and the linearized approximation (4.51)?

- (f) Repeat the solution process for $u(t)$ in parts (c) and (d) but with $u(0) = 2, u'(0) = 0$. Plot and compare.

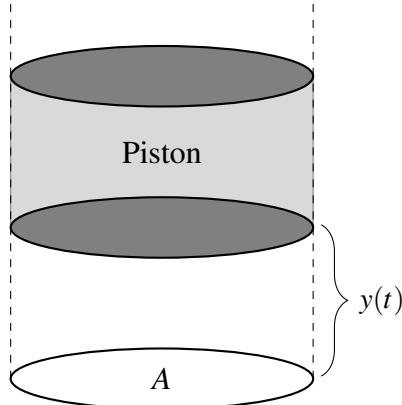


Figure 4.11: Piston in a gas filled cylinder with cross-sectional area A .

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 equation (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,

$$mu''(t) + cu'(t) + ku(t) = f(t). \quad (4.52)$$

As in the homogeneous case (4.4) two initial conditions are needed to specify a unique solution. The focus in this section is the general structure of solutions to (4.52), 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 (4.39) a general solution is given by

$$u(t) = c_1 e^{-t} \cos(9.95t) + c_2 e^{-t} \sin(9.95t).$$

With initial conditions $u(0) = 0.01$ and $u'(0) = 0$ the solution is

$$u(t) = 0.01e^{-t} \cos(9.95t) + 0.001e^{-t} \sin(9.95t).$$

This function is plotted in Figure 4.12. As expected, the oscillations damp out after a few cycles, since there are no external forces acting on the system after time $t > 0$ (though there may have been for $t \leq 0$, in order to initiate the initial nonzero displacement). The sinusoidal portion of the damped oscillations has a frequency of $\frac{9.95}{2\pi} \approx 1.58$ Hz, or a period of about 0.63 seconds.

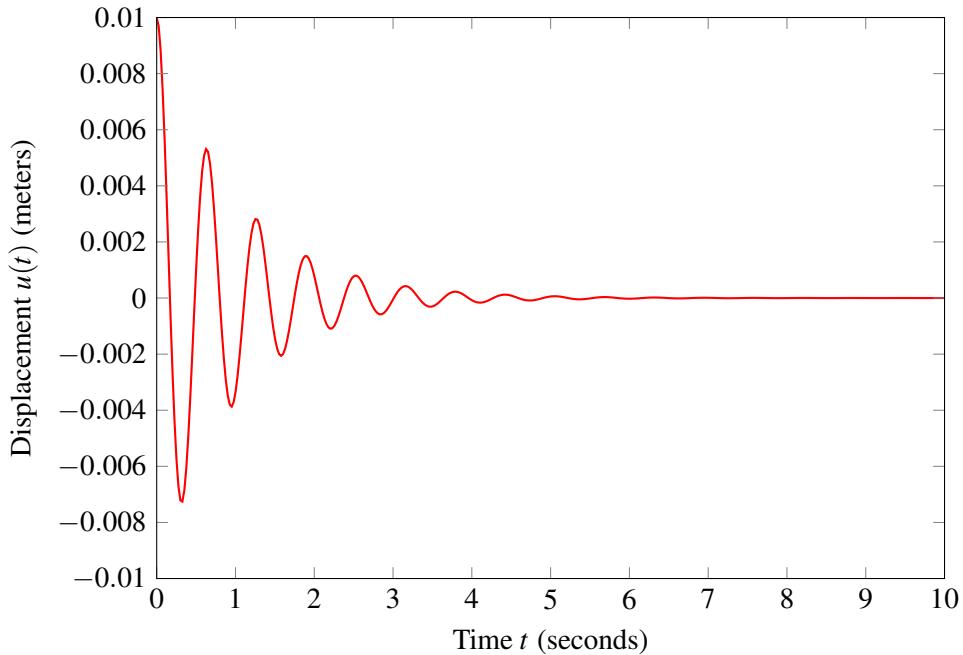


Figure 4.12: Response $u(t)$ of the unforced oscillator $5000u'' + 10^4u' + (5 \times 10^5)u = 0$ with $u(0) = 0.01, u'(0) = 0$.

How would such a structure behave in an earthquake? An earthquake might be modeled as a driving force on the system.² The resulting response may be quite different than 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 equation (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 Hz or higher; see [19]. 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 $5/(2\pi) \approx 0.8$ Hz and amplitude 10^4 newtons, about 2040 pounds. In this case the solution to the ODE with $u(0) = u'(0) = 0$ is, to three significant figures

$$u(t) = \underbrace{-0.0128e^{-t} \sin(9.95t) + 0.0035e^{-t} \cos(9.95t)}_{\text{transient}} + \underbrace{-0.0035 \cos(5t) + 0.0262 \sin(5t)}_{\text{periodic}}. \quad (4.53)$$

How this solution was obtained and its general structure is the focus of this section and will be considered shortly. For the moment, consider a plot of this function, shown in Figure 4.13. The solution (4.53) has a *transient portion* stemming from the terms that contain e^{-t} ; this quickly decays to zero. However, the last two terms on the right in (4.53) involving $\sin(5t)$ and $\cos(5t)$ do not decay in time, and quickly dominate 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. ■

4.3.1 Solving the Forced Harmonic Oscillator Equation

The goal in this section is to 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 $f(t)$, and to understand the nature and structure

²although see the corresponding Project "Earthquake Modeling" in Section 4.6 for a slightly more realistic model.

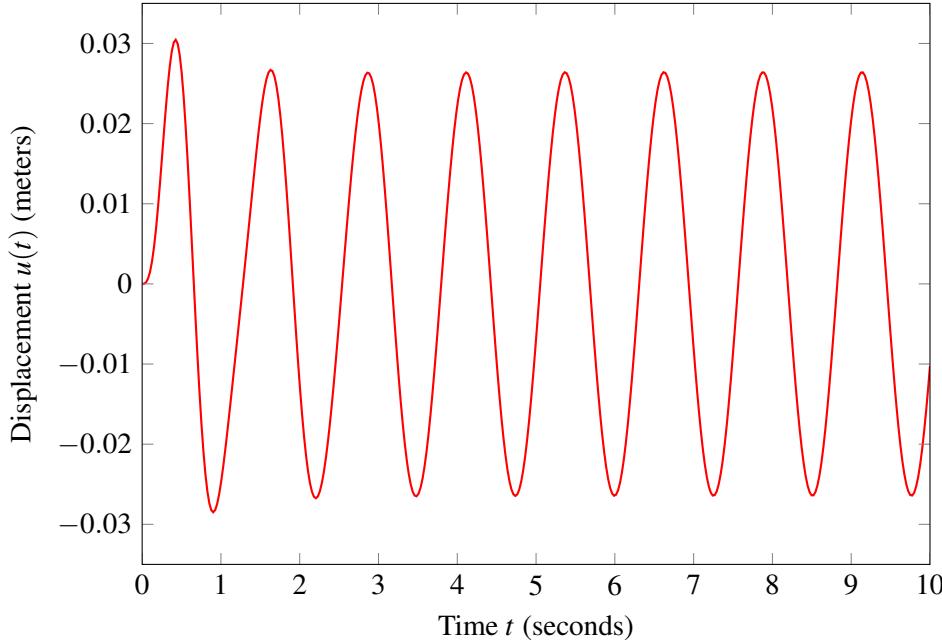


Figure 4.13: Response $u(t)$ of the forced oscillator $5000u'' + 10^4u' + (5 \times 10^5)u = f(t)$ with $f(t) = 10^4 \sin(5t)$.

of the solution. Linearity and the principle of superposition will be essential.

A General Solution to the Forced Equation

At the heart of the computation are 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 obtain any desired initial conditions.

Step 1 we already know how to do. How to accomplish Step 2 we'll consider 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

$$\begin{aligned} mu''(t) + cu'(t) + ku(t) &= m(u_h''(t) + u_p''(t)) + c(u_h'(t) + u_p'(t)) + k(u_h(t) + u_p(t)) \\ &= \underbrace{(mu_h''(t) + cu_h'(t) + ku_h(t))}_0 + \underbrace{(mu_p''(t) + cu_p'(t) + ku_p(t))}_{f(t)} \\ &= f(t) \end{aligned} \tag{4.54}$$

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 above computation in (4.54). 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

$$u(t) = u_h(t) + u_p(t) \quad (4.55)$$

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 $u(0) = 1, u'(0) = 4$. To apply Theorem 4.3.1 first use the procedure developed Section 4.2 to compute a general solution to the homogeneous ODE $u'' + 4u' + 3u = 0$, which is $u(t) = c_1 e^{-t} + c_2 e^{-3t}$. A particular solution $u_p(t)$ to the nonhomogeneous ODE is given by $u_p(t) = -e^{-2t}$, 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 given by

$$u(t) = u_h(t) + u_p(t) = c_1 e^{-t} + c_2 e^{-3t} - e^{-2t}. \quad (4.56)$$

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 $c_1 = 4, c_2 = -2$. The solution to the ODE with the desired initial conditions is then

$$u(t) = 4e^{-t} - 2e^{-3t} - e^{-2t}.$$

■

Reading Exercise 91 Verify that $u(t)$ as given in (4.56) satisfies $u''(t) + 4u'(t) + 3u(t) = e^{-2t}$.

Reading Exercise 92 According to Theorem 4.3.1, any particular solution $u_p(t)$ in (4.55) will yield a valid general solution. The function $u_p(t) = -e^{-2t} - 5e^{-t}$ is a particular solution to the ODE in Example 4.13 (there are infinitely many other particular solutions). Write out the general solution obtained from (4.55) with this choice of $u_p(t)$. Then adjust c_1 and c_2 to obtain $u(0) = 1, u'(0) = 4$. Verify that we obtain exactly the same result as in Example 4.13.

The moral in Reading Exercise 92 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 the 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 (4.3), 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 different examples.

■ **Example 4.14** Let us find a particular solution $u_p(t)$ to

$$u''(t) + 4u'(t) + 3u(t) = 6.$$

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 at 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 an “undetermined coefficient,” 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 at $u_p(t)$ will be of the same form, $u_p(t) = A + Bt + Ct^2$, with A, B , and C as constants to adjust as needed. Then $u'_p(t) = B + 2Ct$ and $u''_p(t) = 2C$. Inserting $u(t)$ 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 above 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 gives flexibility to make the guess work in the ODE. Then $u'_p(t) = -5Ae^{-5t}$ and $u''_p(t) = 25Ae^{-5t}$. Inserting this into the ODE yields

$$(25 - 20 + 3)Ae^{-5t} = e^{-5t}$$

or $8A = 1$ after dividing both sides by e^{-5t} and simplifying, so $A = 1/8$. A particular solution is then

$$u_p(t) = \frac{e^{-5t}}{8}.$$

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 this 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 above equation as

$$(-A - 8B) \sin(2t) + (8A - B) \cos(2t) = \sin(2t). \quad (4.57)$$

Think of the right side in (4.57) as $1 \sin(2t) + 0 \cos(2t)$, which on comparison with the left side of (4.57) makes it clear a particular solution will be obtained if A and B satisfy

$$-A - 8B = 1 \quad \text{and} \quad 8A - B = 0.$$

The solution to these two equations is $A = -1/65$, $B = 8/65$. A particular solution is then given by

$$u_p(t) = -\frac{1}{65} \sin(2t) + \frac{8}{65} \cos(2t).$$

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.

$f(t)$	$u_p(t)$
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.

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 a variety of choices 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 $mu'' + cu' + ku = f_1$ 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 $mu'' + cu' + ku = f_1 + f_2$.

More Undetermined Coefficients Examples

Here are some 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 obtain specific initial conditions.

■ **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 (4.39) 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 $u_p(t) = A\cos(t) + B\sin(t)$. Using $u'_p(t) = -A\sin(t) + B\cos(t)$ and $u''_p(t) = -A\cos(t) - B\sin(t)$ in 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

$$u(t) = u_p(t) + u_h(t) = \frac{1}{10}\cos(t) + \frac{1}{20}\sin(t) + c_1 e^{-t} \cos(2t) + c_2 e^{-t} \sin(2t).$$

To obtain the initial conditions use this general solution to see that $u(0) = 1/10 + c_1 = 1$ and $u'(0) = -c_1 + 2c_2 + 1/20 = -1$. These two equations have solution $c_1 = 9/10$, $c_2 = -3/40$. The

full solution is then

$$u(t) = \frac{1}{10} \cos(t) + \frac{1}{20} \sin(t) + \frac{9}{10} e^{-t} \cos(2t) - \frac{3}{40} e^{-t} \sin(2t).$$

■

Example 4.19 A building is modeled by a spring-mass-damper ODE with $m = 10^4$, $c = 2 \times 10^5$, and $k = 10^7$, with $u(t)$ as the displacement of the mass. 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 (4.39) a general solution to the homogeneous version of the ODE is

$$u_h(t) = c_1 e^{-10t} \cos(30t) + c_2 e^{-10t} \sin(30t).$$

Based on the fact that $f(t) = 10^5 e^{-5t} \sin(30t)$ and the information in Table 4.2, next seek a particular solution of the form

$$u_p(t) = A e^{-5t} \cos(30t) + B e^{-5t} \sin(30t).$$

Inserting $u_p(t)$ into the ODE and collecting like terms produces

$$250000 e^{-5t} (A + 12B) \cos(30t) + 250000 e^{-5t} (-12A + B) \sin(30t) = 10^5 e^{-5t} \sin(30t).$$

Divide both sides above 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 above must be identical as functions of t , 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 $B = 2/725$. That is,

$$u_p(t) = -\frac{24}{725} e^{-5t} \cos(30t) - \frac{2}{725} e^{-5t} \sin(30t)$$

is a particular solution to $mu'' + cu' + ku = f$. From Theorem 4.3.1 a general solution is $u(t) = u_p(t) + u_h(t)$, or

$$u(t) = -\frac{24}{725} e^{-5t} \cos(30t) - \frac{2}{725} e^{-5t} \sin(30t) + c_1 e^{-10t} \cos(30t) + c_2 e^{-10t} \sin(30t). \quad (4.58)$$

The conditions $u(0) = 0$ and $u'(0) = 0$ in (4.58) lead to equations

$$\begin{aligned} -24/725 + c_1 &= 0 \\ -10c_1 + 30c_2 + 36/145 &= 1 \end{aligned}$$

which have solution $c_1 = 24/725$ and $c_2 = 2/725$. The solution we seek 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 $0 \leq t \leq 2$.

■

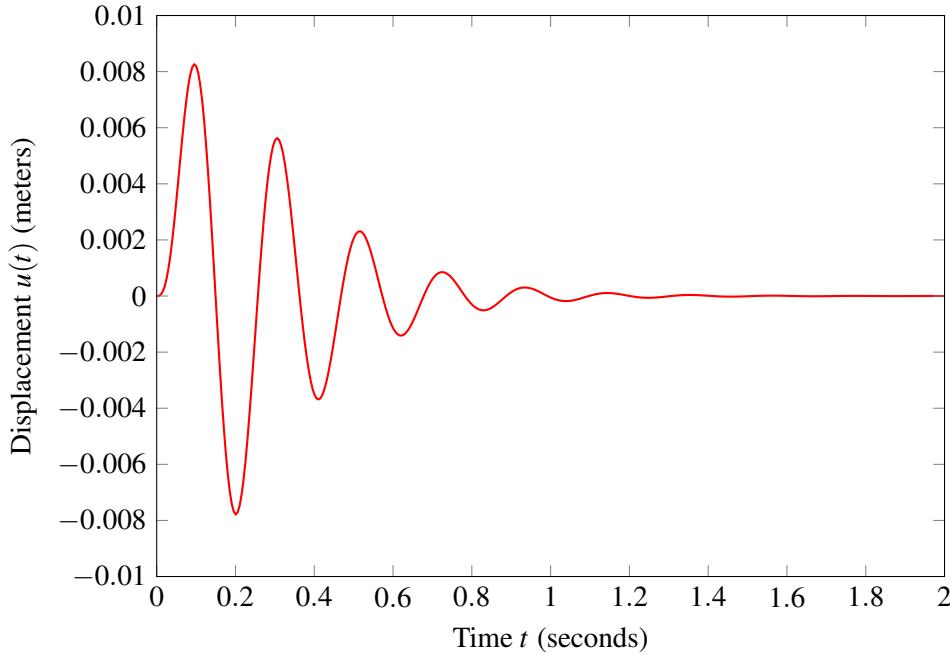


Figure 4.14: Solution $u(t)$ to $10^4u''(t) + (2 \times 10^5)u'(t) + 10^7u(t) = 10^5 \sin(30t)$ with $u(0) = u'(0) = 0$.

■ **Example 4.20** Recall Example 4.1 in which the front shock absorber on a mountain bike was modeled. When the front wheel is in contact with the ground (and so not moving vertically, while supporting the bike and rider's weight), the relevant ODE for the displacement $u(t)$ of the shock-damper is given by (4.6), reproduced here for convenience:

$$46u''(t) + 1700u'(t) + 15000u(t) = -450.8. \quad (4.59)$$

In that example it was assumed that the weight of the rider and bicycle combined is 92 kg, half of which is supported by the front shock, hence the 46 in front of $u''(t)$ above. The spring constant is assumed to be 15000 newtons per meter and the viscous damping coefficient is 1700 newtons-seconds per meter. The constant forcing function on the right side of (4.59) is $-mg = -450.8$, where gravitational acceleration is $g = 9.8$ meters per second squared.

Let us work out a general solution to (4.59). First, the characteristic equation for the homogeneous ODE $46u''(t) + 1700u'(t) + 15000u(t) = 0$ is

$$46r^2 + 1700r + 15000 = 0$$

and 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 then

$$u_h(t) = c_1 e^{-14.56t} + c_2 e^{-22.40t}.$$

Since the forcing function is constant, a particular solution of the form $u_p(t) = u^*$ (constant) is appropriate. In the ODE (4.59) this 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 (4.59) is given by

$$u(t) = -0.03 + c_1 e^{-14.56t} + c_2 e^{-22.40t}. \quad (4.60)$$

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

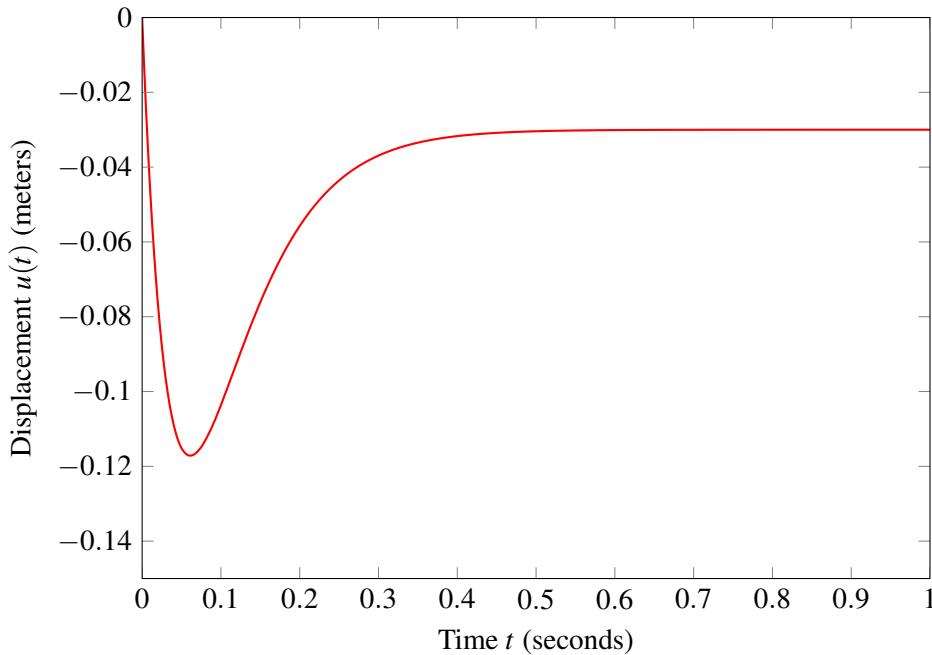


Figure 4.15: Displacement of bicycle front shock after 1.5 meter jump.

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 (4.59) governs the shock's behavior. At this instant $u(0) = 0$, since the shock was not compressed in the air. The behavior of the shock for $t > 0$ can be determined from the additional data $u'(0)$.

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. In the case that means the bike hits the ground at a speed of $v_0 \approx -5.42$ meter per second, negative since the bike is falling. This is $u'(0)$. The initial data $u(0) = 0$ and $u'(0) = -5.42$ along with (4.60) leads to equations

$$-0.03 + c_1 + c_2 = 0 \quad \text{and} \quad -14.56c_1 - 22.4c_2 = -5.42$$

that have solution $c_1 \approx -0.606$, $c_2 \approx 0.636$. The displacement of the shock is given by

$$u(t) \approx -0.03 - 0.606e^{-14.56t} + 0.636e^{-22.40t}.$$

A plot of this function is shown in Figure 4.15. ■

Reading Exercise 93 Compute the maximum compression of the shock that occurs in Example 4.20, to three significant figures. Suppose the bike front shock has a range of motion of about 140 mm before “bottoming out.” Would that be an issue in this scenario?

4.3.3 When the Guess Fails

Under some circumstances the guesses in Table 4.2 fail, but it is generally 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

$$u''(t) + 4u'(t) + 3u(t) = e^{-t}. \tag{4.61}$$

Based on Table 4.2 (and more generally, the intuition that the ansatz should look like the forcing function) seek a particular solution of the form $u_p(t) = Ae^{-t}$. Inserting this into (4.61) 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 , so try a guess of the form

$$u_p(t) = A(t)e^{-t}.$$

Compute

$$\begin{aligned} u'_p(t) &= -A(t)e^{-t} + A'(t)e^{-t}, \\ u''_p(t) &= A(t)e^{-t} - 2A'(t)e^{-t} + A''(t)e^{-t}, \end{aligned}$$

then substitute this information into (4.61) and collect terms to find $e^{-t}(A''(t) + 2A'(t)) = e^{-t}$, or

$$A''(t) + 2A'(t) = 1. \quad (4.62)$$

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 $A''(t) = 0$) and then (4.62) becomes $2A'(t) = 1$ with solution $A(t) = t/2$. This means that a particular solution to (4.61) is given by

$$u_p(t) = A(t)e^{-t} = \frac{t}{2}e^{-t}.$$

This particular solution can then be used to form a general solution to (4.61) 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} dictated 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. This frequently turns out to be the case.

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 $t^n e^{r_1 t}$ and/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. And 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 trigonometric functions.

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 work. The interested reader should consult [27] 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 (though computations can be a bit cumbersome).

■ **Example 4.22** Let us find a particular solution to

$$2u''(t) + 4u'(t) + 10u(t) = e^{-t} \cos(2t), \quad (4.63)$$

then use this to construct a general solution to this ODE, and find a solution with initial data $u(0) = 2, u'(0) = 4$.

The characteristic equation for the homogeneous version of (4.63) is $2r^2 + 4r + 10$ with roots $r = -1 \pm 2i$. A real-valued general solution is therefore given by

$$u_h(t) = c_1 e^{-t} \cos(2t) + c_2 e^{-t} \sin(2t).$$

Next we seek a particular solution $u_p(t)$. Based on Table 4.2 it might expected that $u_p(t) = Ae^{-t} \cos(2t) + Be^{-t} \sin(2t)$ will work, but based on $u_h(t)$ 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 try replacing this guess with

$$u_p(t) = A(t)e^{-t} \cos(2t) + B(t)e^{-t} \sin(2t). \quad (4.64)$$

Inserting $u = u_p$ into (4.63) 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 ODE's 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 above equations becomes $8B'(t) = 1$ 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 (4.64) shows that

$$u_p(t) = \frac{te^{-t} \sin(2t)}{8}$$

is a solution to (4.63).

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 equation $c_1 = 2$, while $u'(0) = 4$ forces $-c_1 + 2c_2 = 4$, and so $c_2 = 3$. The solution to (4.63) 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 ODE's

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 ODE's can be handled in much the same manner as second order. 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 $r^3 + r^2 - 2 = 0$ for this ODE, 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 $u(0) = u_0, u'(0) = u'_0, u''(0) = u''_0$ 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 (4.39). In general an n th order ODE leads to an n degree characteristic equation whose roots must be found. 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 ODE's. Alternatively, second and higher ODE's can be converted into systems of first order ODE's and analyzed using techniques we will develop in Chapters 6 and 7.

Linear ODE's 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 a_k coefficients. See [27] for more on this topic. Second and higher order nonlinear ODE's are kind of the wild west, with techniques specific to small classes of problems. However, the ideas in Chapter 7 may aid analysis.

4.3.4 Exercises

Exercise 4.3.1 For each ODE in parts (a)-(x) below

- Find the 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 the general solution to the nonhomogeneous ODE and use it to obtain initial conditions $u(0) = 2, u'(0) = 3$.

- (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) below

- Find the 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; this answer is not unique.
- Write out the general solution to the nonhomogeneous ODE and use it to obtain the specific solution with initial conditions $u(0) = 2, u'(0) = 3$.

- (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 the characteristic equation for this ODE. 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 $k \neq 0$.

- (a) Suppose $f(t) = a_0$ is constant. Show that an ansatz of the form $u_p(t) = A_0$ will always work for finding a particular solution.
- (b) Suppose $f(t) = a_0 + a_1 t$. Show that an ansatz of the form $u_p(t) = A_0 + A_1 t$ will always work for finding a particular solution.
- (c) Suppose $f(t)$ is an n th degree polynomial. Will taking $u_p(t)$ as an n th degree polynomial with undetermined coefficients always work? Why?

Exercise 4.3.5

- (a) Redo the solution for the ODE (4.59) 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 93 in this setting. What disadvantage might such a value of c have for the rider?
- (b) Redo the solution for the ODE (4.59) 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 93. 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 term of $x(t)$ directly the Hill-Keller ODE becomes

$$x''(t) = P - kx'(t) \quad (4.65)$$

with initial conditions $x(t_0) = 0$ and $x'(t_0) = 0$.

- (a) Equation (4.65) 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 .
- (b) Write out the general solution $x_h(t)$ to the homogeneous ODE.
- (c) Find a particular solution $x_p(t)$ to (4.65).
- (d) Find the solution with initial data $x(t_0) = 0, x'(t_0) = 0$, and verify that we obtain the same result as in (3.46). ■

Exercise 4.3.7 Consider a vibration isolation table as modeled in Example 4.2 and the ODE (4.12). 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.

- (a) Find the position $y(t) = y_{eq}$ assumed by the table if only gravity acts on the tabletop (and $d(t) = 0$).
- (b) Suppose the ground begins to vibrate according to $d(t) = 10^{-4} \cos(40\pi t)$ (20 hz) 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.

- (a) 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 $q(t)$?
- (b) Find the general solution $q_h(t)$ to the homogeneous version of the ODE. Based on your solution, is this system under, over, or critically damped?
- (c) Use the method of undetermined coefficients to find a particular solution $q_p(t)$ to the nonhomogeneous ODE.
- (d) Use your work from parts (b) and (c) to find a general solution to the nonhomogeneous

ODE, and then obtain the required initial conditions.

- (e) 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 phenomena 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 (a building in an earthquake) and sometimes it's an essential part of how the system functions, as is the case for the classic RLC tuner circuit in a radio; see the project "Stayed 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, and $c = 10^4$ newtons per meter per second, and forcing function $f(t) = 10^4 \sin(5t)$ newtons. This forcing function represents a frequency of $5/(2\pi) \approx 0.8$ Hz at an amplitude of 10^4 newtons. With initial conditions $u(0) = 0$ and $u'(0) = 0$ the solution was

$$u(t) = \underbrace{-0.0128e^{-t} \sin(9.95t) + 0.0035e^{-t} \cos(9.95t)}_{\text{transient}} + \underbrace{-0.0035 \cos(5t) + 0.0262 \sin(5t)}_{\text{periodic}}. \quad (4.66)$$

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 a 5 seconds. The periodic portion on the right in (4.66) 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 Hz. The solution with $u(0) = u'(0) = 0$ can be found using the techniques of Section 4.3 and is given by

$$u(t) = \underbrace{0.01e^{-t} \sin(9.95t) + 0.1e^{-t} \cos(9.95t)}_{\text{transient}} - \underbrace{0.1 \cos(10t)}_{\text{periodic}}. \quad (4.67)$$

The response of the building again has a transient portion, but after that the solution consists primarily of the periodic term on the right in (4.67). 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 Hz.

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.

Despite the fact that in each case the sinusoidal driving function had amplitude 10^4 newtons, driving this system at 1.59 Hz elicited a much more vigorous response, compared to 0.8 Hz. By no small coincidence, the undriven version of this system has a natural frequency of about $10/(2\pi) \approx 1.59$ Hz.

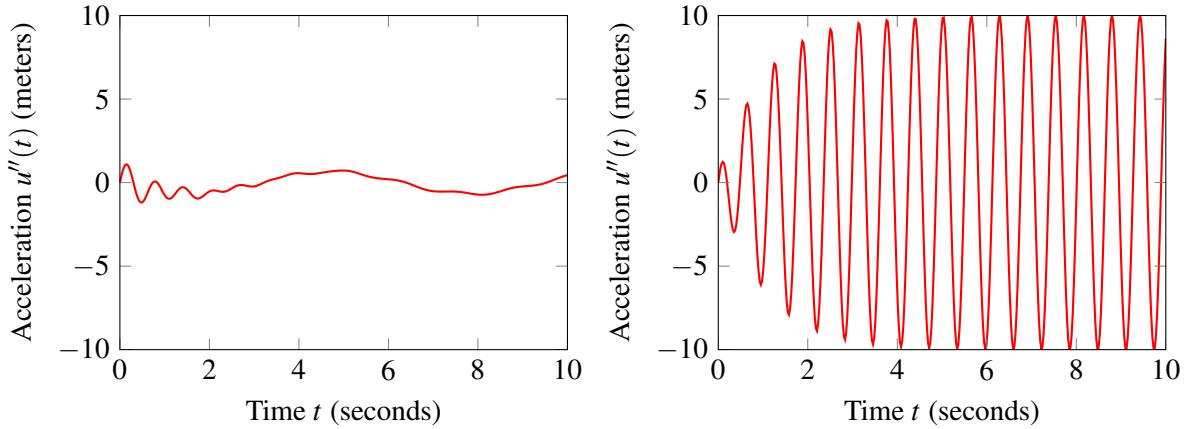


Figure 4.16: Acceleration $u''(t)$ of the forced oscillator $5000u'' + (10^4u' + (5 \times 10^5)u = f(t)$ with $f(t) = 10^4 \sin(5t)$ (left) and $f(t) = 10^4 \sin(10t)$ (right).

4.4.2 Periodic Forcing

Let's analyze this situation more generally and consider how a spring-mass system responds to sinusoidal driving. We return to the driven harmonic oscillator (4.3)

$$mu''(t) + cu'(t) + ku(t) = f(t), \quad (4.68)$$

reproduced here for convenience. One of the most common and important types of forcing functions $f(t)$ in (4.68) are those of the form

$$f(t) = C \sin(\omega t + \phi) \quad (4.69)$$

where C , ω , and ϕ are specified constants. We assume $\omega \geq 0$. The function $f(t)$ in (4.69) is sinusoidal with C as the amplitude, ω the frequency, and ϕ acting as a phase shift. As it turns out, most “reasonable” forcing functions can be written as a linear combination (possibly with infinitely many terms) of functions of the form $C \sin(\omega t + \phi)$, but that's another story. In any case, by analyzing the ODE (4.68) with a sinusoidal forcing term (4.69) we gain insight into solutions with other forcing functions as well.

As shown in Exercise 4.2.9, such a function $f(t)$ can also be written in the form

$$f(t) = A_1 \sin(\omega t) + A_2 \cos(\omega t) \quad (4.70)$$

by taking $A_1 = C \cos(\phi)$ and $A_2 = C \sin(\phi)$. We will work with whichever form, (4.69) or (4.70), is most convenient at the time.

The general solution to (4.68) with $f(t)$ as in (4.69) or (4.70) is of the form

$$\begin{aligned} u(t) &= u_h(t) + u_p(t) \\ &= \underbrace{c_1 u_1(t) + c_2 u_2(t)}_{u_h(t), \text{ transient}} + \underbrace{A \cos(\omega t) + B \sin(\omega t)}_{u_p(t), \text{ periodic}}. \end{aligned} \quad (4.71)$$

The transient portion of the solution, $c_1 u_1(t) + c_2 u_2(t)$, is simply the general solution to the homogeneous ODE $mu'' + cu' + ku = 0$. If $c > 0$ then both $u_1(t)$ and $u_2(t)$ involve decaying exponentials and quickly limit to zero as t increases, so only the periodic portion $u_p(t)$ of the solution remains. The periodic response satisfies

$$mu_p''(t) + cu_p'(t) + ku_p(t) = C \sin(\omega t + \phi) \quad (4.72)$$

and is of the form

$$u_p(t) = A \cos(\omega t) + B \sin(\omega t) \quad (4.73)$$

for a suitable choice of A and B . This is the portion of the solution which is of interest. Notice that $u_p(t)$ never decays, so long as the driving function $f(t)$ remains active. Let's examine how $u_p(t)$ depends on the parameters C , ω , and ϕ . Of particular interest is the amplitude $\sqrt{A^2 + B^2}$ of $u_p(t)$.

Reading Exercise 94 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 is given by $1/\sqrt{\omega^4 - 16\omega^2 + 100}$. Plot this amplitude as a function of ω for $0 \leq \omega \leq 10$.

Computing the Periodic Response

Inserting $u_p(t)$ in the form (4.73) into the ODE (4.72) leads to

$$\begin{aligned} & (-Am\omega^2 + Bc\omega + Ak)\cos(\omega t) + (-Bm\omega^2 - Ac\omega + Bk)\sin(\omega t) \\ &= C \sin(\phi) \cos(\omega t) + C \cos(\phi) \sin(\omega t) \end{aligned} \quad (4.74)$$

after performing all the differentiations and collecting the $\cos(\omega t)$ and $\sin(\omega t)$ terms separately on the left in (4.74), as well as expanding the right hand side of (4.72) as $C \sin(\omega t + \phi) = C \sin(\phi) \cos(\omega t) + C \cos(\phi) \sin(\omega t)$. In order for (4.74) 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 (4.74) match, which yields equations

$$\begin{aligned} (k - m\omega^2)A + c\omega B &= C \sin(\phi), \\ -Ac\omega + (k - m\omega^2)B &= C \cos(\phi). \end{aligned} \quad (4.75)$$

This is a system of two linear equations in two unknowns A and B , and some mildly tedious algebra yields solution

$$\begin{aligned} A &= C \left(\frac{\sin(\phi)(k - m^2\omega^2) - \cos(\phi)c\omega}{(m\omega^2 - k)^2 + c^2\omega^2} \right), \\ B &= C \left(\frac{\cos(\phi)(k - m^2\omega^2) + \sin(\phi)c\omega}{(m\omega^2 - k)^2 + c^2\omega^2} \right). \end{aligned} \quad (4.76)$$

In (4.73) this yields the periodic solution $u_p(t)$ to (4.72).

The Amplitude of the Periodic Response

Our real interest is 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 (4.76), the quantity $\sqrt{A^2 + B^2}$ simplifies considerably and is given by

$$\sqrt{A^2 + B^2} = \underbrace{\frac{C}{\sqrt{(m\omega^2 - k)^2 + c^2\omega^2}}}_{\psi(\omega)} \quad (4.77)$$

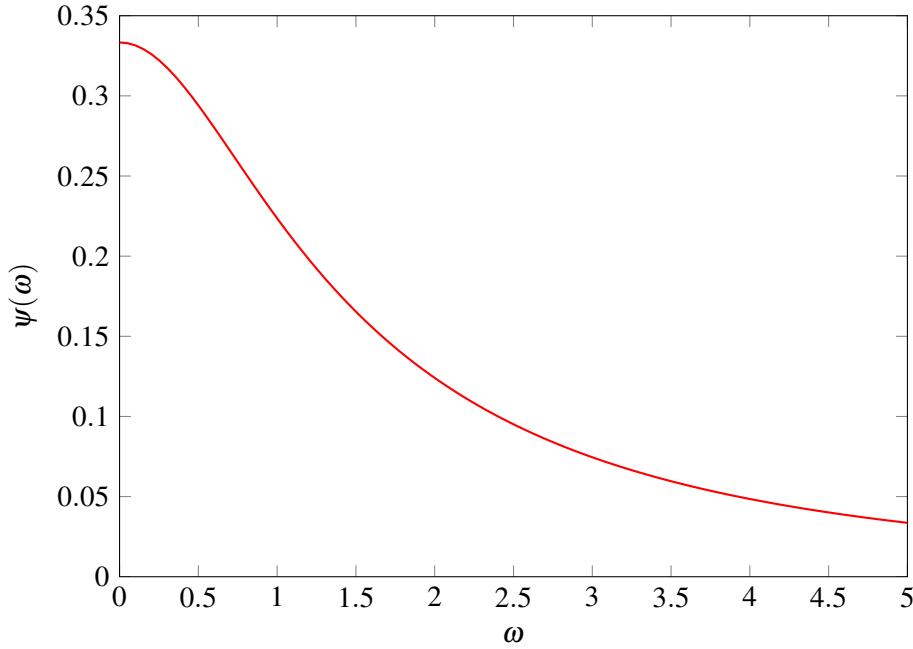


Figure 4.17: Periodic response amplitude for $u''(t) + 4u'(t) + 3u(t) = \sin(\omega t)$.

after simplification, 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 , the amplitude of the driving force.
- The response amplitude $\psi(\omega)$ does not depend on the input phase ϕ .
- If $c > 0$ then 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 the driving amplitude $C = 1$ and ω and ϕ are unspecified in the forcing function $f(t)$ of (4.69). According to the analysis above, the periodic response $u_p(t)$ of the system will be of the form (4.73) and have amplitude

$$\psi(\omega) = \frac{1}{\sqrt{(\omega^2 - 3)^2 + 16\omega^2}}$$

that 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.

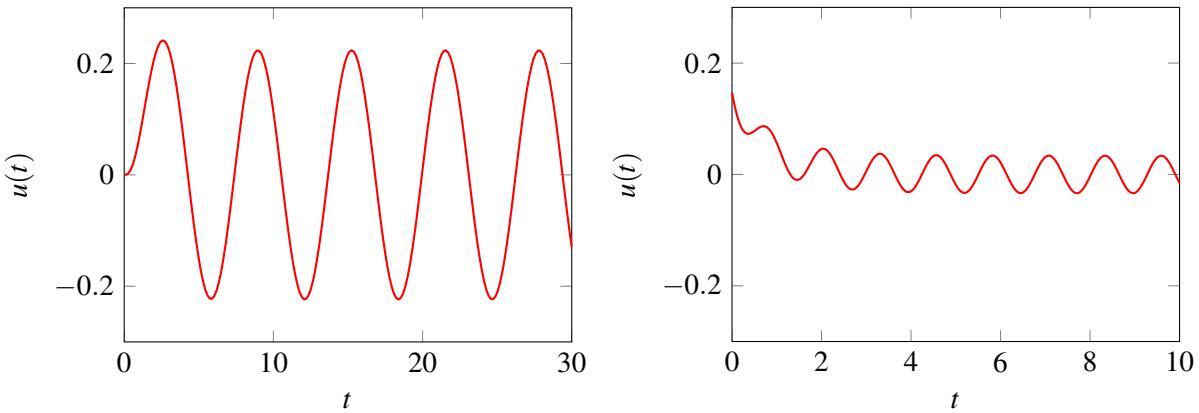


Figure 4.18: Solution to $u''(t) + 4u'(t) + 3u(t) = \sin(\omega t)$ with $u(0) = u'(0) = 0$ and $\omega = 1$ (left panel) and $\omega = 5$ (right panel).

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, in each case with initial conditions $u(0) = u'(0) = 0$. 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. ■

■ **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 (4.69). According to the analysis above, the periodic response $u_p(t)$ of the system will be of the form (4.73) 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 4.20 shows the response $u(t)$ of the system to forcing at $\omega = 3$ in the left panel and $\omega = 5$ in the right panel, in each case with initial conditions $u(0) = u'(0) = 0$. 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 driving force in each case has amplitude $C = 1$.

Reading Exercise 95 Show that the maximum value of $\psi(\omega) = 1/\sqrt{(\omega^2 - 10)^2 + \omega^2}$ graphed in Figure 4.19 is assumed at $\omega = \sqrt{38}/2 \approx 3.08$.

Here is something of which to take 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

$$2u''(t) + 18u(t) = 5 \sin(\omega t + \phi).$$

We've taken $C = 5$ in the forcing function $f(t)$ of (4.69), 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 (4.73) and have

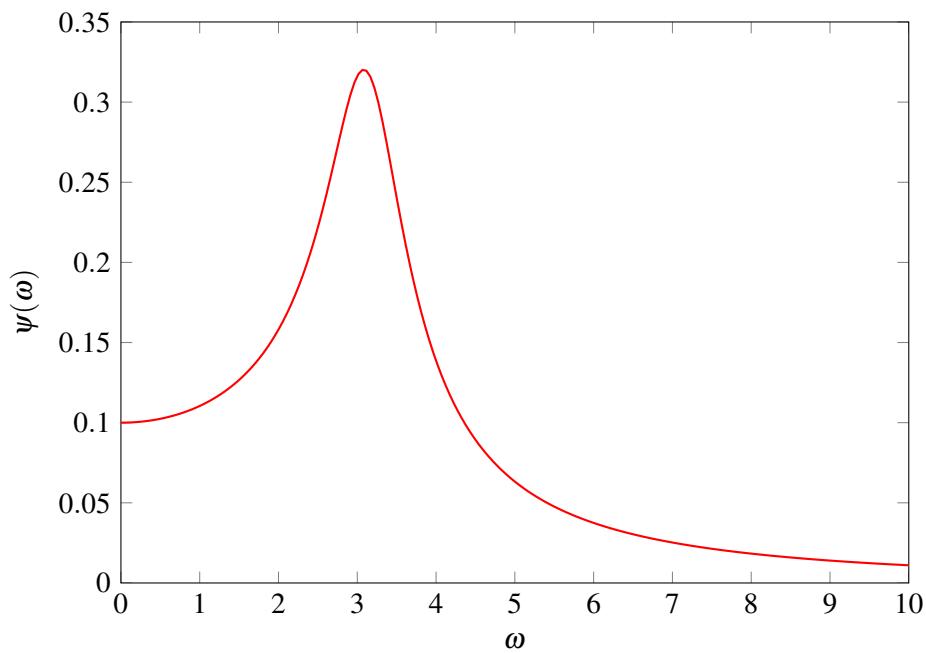


Figure 4.19: Periodic response amplitude for $u''(t) + u'(t) + 10u(t) = \sin(\omega t)$.

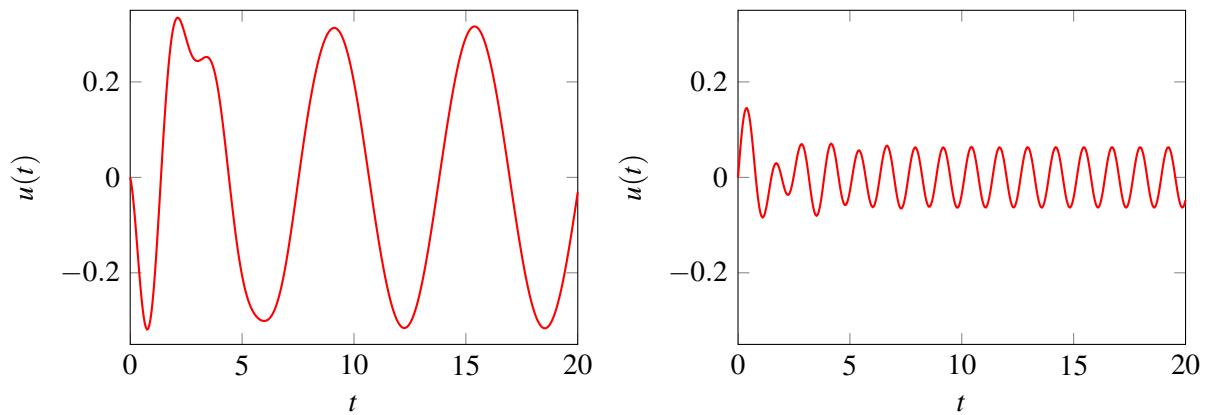


Figure 4.20: Solution to $u''(t) + u'(t) + 10u(t) = \sin(\omega t)$ with $u(0) = u'(0) = 0$ and $\omega = 3$ (left panel) and $\omega = 5$ (right panel).

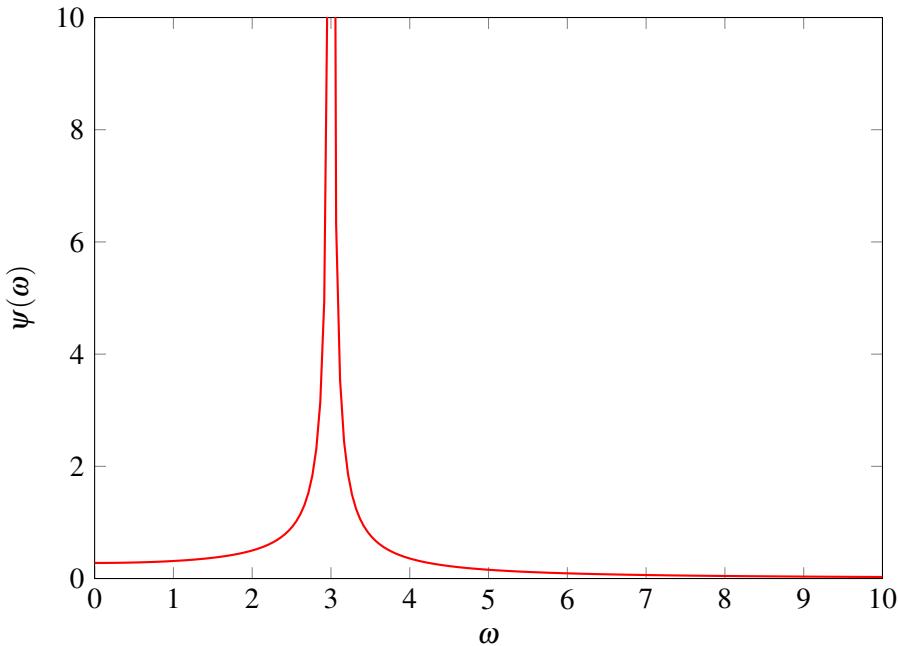


Figure 4.21: Periodic response amplitude for $2u''(t) + 18u(t) = 5 \sin(\omega t)$.

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. Note that 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.

What does the system do when actually driven with a forcing function of the form $f(t) = C \sin(3t)$? In Figure 4.22 is shown the solution to $2u''(t) + 18u(t) = 5 \sin(3t)$ with $u(0) = 1$, $u'(0) = 0$. 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 Example 4.22 succeeds and yields 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. ■

Resonance

The phenomena in which a “lightly” damped (or undamped) system responds vigorously at one or more frequencies as illustrated in Examples 4.24 and 4.25 is called *resonance*. Resonance may be undesirable in some systems, for example, a building in a earthquake, for it greatly increases the system’s response to periodic input. In other cases it is an essential component of how the system functions, for example the RLC tuning circuit in an analog radio. 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 given by $\psi(\omega)$ in (4.77), 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

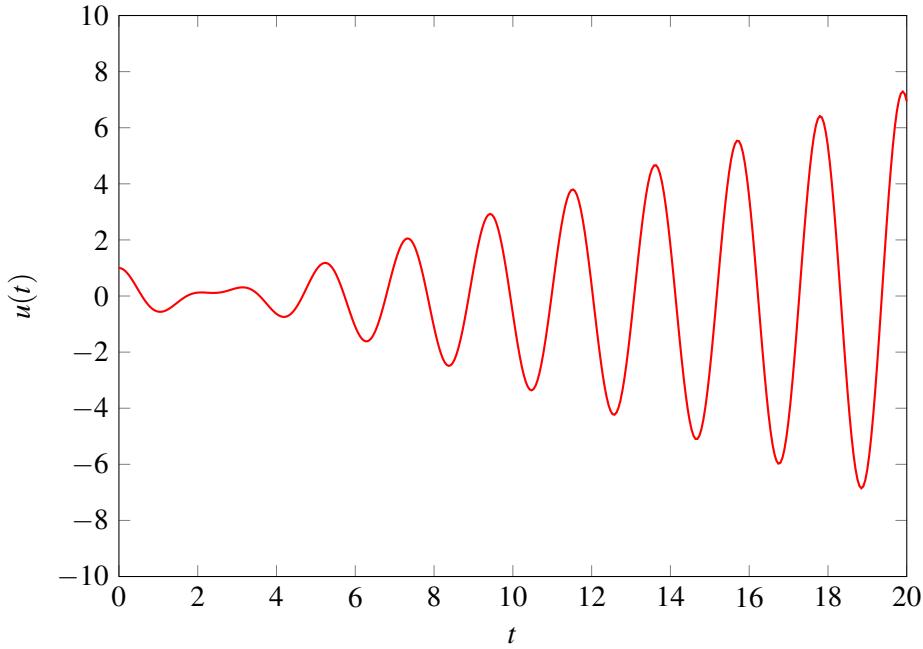


Figure 4.22: Solution to $2u''(t) + 18u(t) = 5 \sin(3t)$ with $u(0) = 1$, $u'(0) = 0$.

can focus on the effect of ω , by defining

$$G(\omega) = \frac{1}{\sqrt{(m\omega^2 - k)^2 + c^2\omega^2}}. \quad (4.78)$$

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 $\psi(\omega) = G(\omega)C$. This 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 (4.78) that $G(\omega) \geq 0$ in all cases and $\lim_{\omega \rightarrow \infty} G(\omega) = 0$. For sufficiently heavily damped systems this graph looks like that of Figure 4.17, and simply decreases monotonically with increasing ω . But for less heavily damped systems the graph of $G(\omega)$ has a unique peak for $\omega > 0$. 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 given by $\omega = 0$; look back 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

$$\omega_{res} = \frac{\sqrt{4km - 2c^2}}{2m} = \sqrt{\frac{k}{m} - \frac{1}{2}\left(\frac{c}{m}\right)^2}. \quad (4.79)$$

If $4km - 2c^2 > 0$ then ω_{res} exists as 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. To summarize

Definition 4.4.1 A system governed by $mu'' + cu' + ku = f(t)$ has a *resonant frequency* ω_{res} given by (4.79) if $c^2 < 2km$.

A few observations concerning resonance can be made:

- 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 $\sqrt{4km - 2c^2}$ in c at $c = 0$ shows that

$$\sqrt{4km - 2c^2} = 2\sqrt{km} - \frac{c^2}{2\sqrt{km}} + O(c^4).$$

Using this in (4.79) shows that

$$\omega_{res} = \sqrt{\frac{k}{m} - \frac{c^2}{4\sqrt{km^3}}} + O(c^4). \quad (4.80)$$

Thus for lightly damped systems ($c \approx 0$) the resonant frequency is very close to $\sqrt{k/m}$.

Reading Exercise 96 Show that $G(0) = 1/k$. Also show that

$$\lim_{\omega \rightarrow \infty} m\omega^2 G(\omega) = 1.$$

Thus if ω is large we are justified in writing $G(\omega) \approx \frac{1}{m\omega^2}$.

Reading Exercise 97 Suppose a spring-mass-damper system with mass m , damping c , and spring constant k is underdamped (that is, $c^2 < 4mk$). Will resonance as defined above 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 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

$$mu''(t) + ku(t) = C \sin(\omega t + \phi). \quad (4.81)$$

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 $\omega \neq \omega_0$ in (4.81) then one can use the method of undetermined coefficients to find that a particular solution is given by

$$u_p(t) = \frac{C \sin(\omega t + \phi)}{\omega^2 - \omega_0^2}$$

(or just verify that $u_p(t)$ as presented satisfies the ODE). This means a general solution to (4.81) is given by

$$\begin{aligned} u(t) &= u_h(t) + u_p(t) = c_1 \cos(\omega_0 t) + c_2 \sin(\omega_0 t) + \frac{C \sin(\omega t + \phi)}{\omega^2 - \omega_0^2} \\ &= c_1 \cos(\omega_0 t) + c_2 \sin(\omega_0 t) + \frac{C \sin(\phi)}{\omega^2 - \omega_0^2} \cos(\omega t) + \frac{C \cos(\phi)}{\omega^2 - \omega_0^2} \sin(\omega t) \end{aligned} \quad (4.82)$$

by making use of $\sin(a+b) = \sin(a)\cos(b) + \cos(a)\sin(b)$. The constants c_1 and c_2 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. This can have some peculiar consequences.

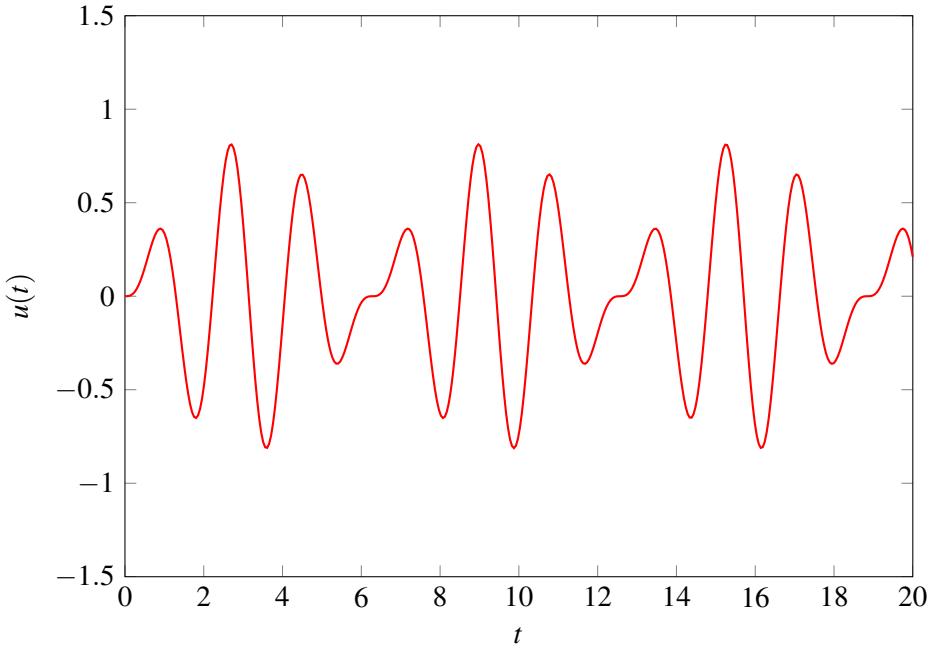


Figure 4.23: Solution to $2u''(t) + 18u(t) = 5 \sin(4t)$ with $u(0) = 0, u'(0) = 0$.

■ **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 $u(0) = 0, u'(0) = 0$. 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 initial conditions, but since there is no damping, this contribution never decays. The $\sin(4t)$ term is due to the driving force. This solution 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 itself periodic. ■

Reading Exercise 98 What is the period of the solution in Example 4.26?

■ **Example 4.27** Consider now what happens if the driving frequency is close 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 in the left panel of Figure 4.24, on the range $0 \leq t \leq 15$. This plot looks very much like the pure resonance of Example 4.25, but plotting out to $t = 100$ shows that the amplitude does not continue to increase. Instead we see here a peculiar “beat” phenomena. This 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. This phenomena occurs in lightly damped systems too, at least for some period of time until the transient portion of the solution dies out. This illustrates a common auditory phenomena 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 phenomena, see [10].

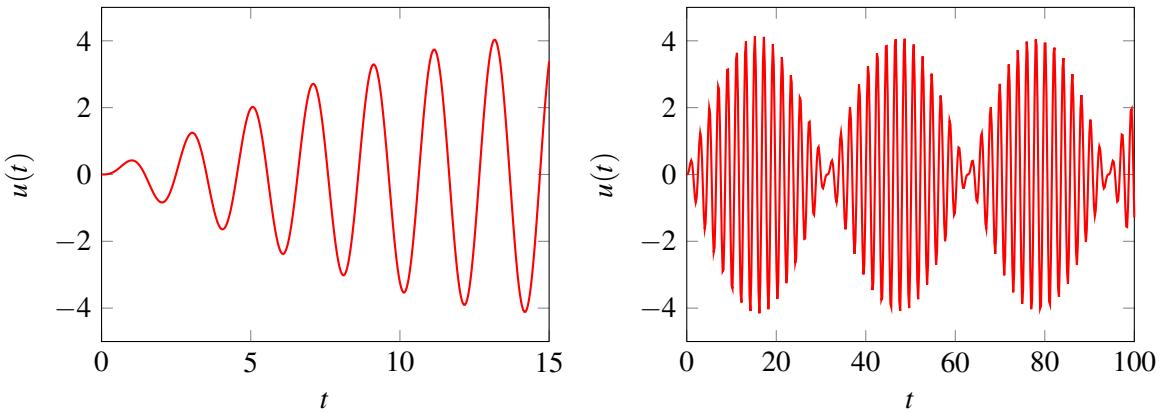


Figure 4.24: Solution to $2u''(t) + 18u(t) = 5 \sin(3.2t)$ with $u(0) = 0$, $u'(0) = 0$ for $0 \leq t \leq 15$ (left) and $0 \leq t \leq 100$ (right).

Analysis of the Beat Phenomena

To understand the beat phenomena 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 (4.82), a general solution is

$$u(t) = c_1 \cos(\omega_0 t) + c_2 \sin(\omega_0 t) + A \cos(\omega t) + B \sin(\omega t) \quad (4.83)$$

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 $u(t) = -A \cos(\omega_0 t) + A \cos(\omega t)$. And in fact let us further specify that $A = 1$ so that

$$u(t) = \cos(\omega t) - \cos(\omega_0 t). \quad (4.84)$$

This function is graphed in Figure 4.25 for the case $\omega_0 = 2$, $\omega = 2.2$ (the red curve).

Reading Exercise 99 What initial conditions does $u(t)$ in (4.84) satisfy?

To understand the behavior of $u(t)$ in (4.84) 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

$$\begin{aligned} u(t) &= \cos(\omega t) - \cos(\omega_0 t) \\ &= 2 \sin((\omega_0 - \omega)t/2) \sin((\omega + \omega_0)t/2). \end{aligned} \quad (4.85)$$

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 (4.85) can be written as

$$u(t) = \underbrace{2 \sin(\delta t/2)}_{\text{amplitude}} \sin((\omega_0 - \delta/2)t). \quad (4.86)$$

The right side of (4.86) can be interpreted as a cosine wave, $\sin((\omega_0 - \delta/2)t)$, that oscillates with a frequency close to ω_0 , but with slowly varying amplitude $2 \sin(\delta t/2)$ (since δ is close to zero). It is this slowly varying amplitude that gives rise to the “beat” phenomena.

In Figure 4.25 is shown the graph of $u(t)$ in red for the case $\omega_0 = 2$, $\omega = 1.8$, so $\delta = 0.2$. Here

$$u(t) = \cos(1.8t) - \cos(2t) = 2 \sin(0.1t) \sin(1.9t).$$

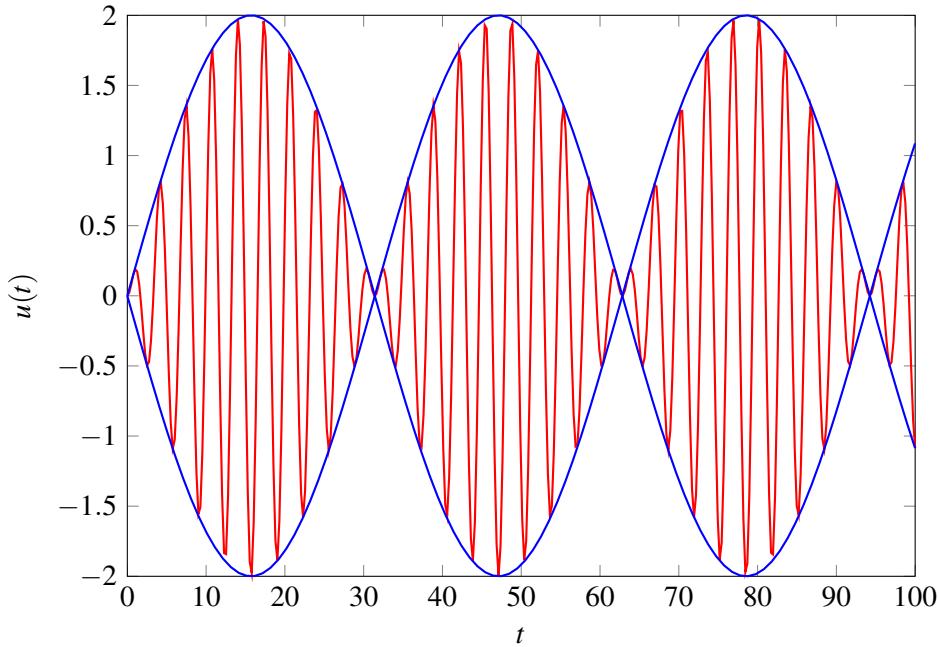


Figure 4.25: Plot of $u(t) = \cos(1.8t) - \cos(2t)$ (red) and amplitude envelope $2\sin(0.1t)$ (blue).

Also shown in blue is the graph of $\pm 2\sin(\delta t/2)$ (which here is $2\sin(0.1t)$) that delineates the slowly varying amplitude of the $\sin(1.9t)$ piece. The quantity $\pm 2\sin(\delta t/2)$ that delineates the amplitude of the rapidly varying sinusoidal piece is usually 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

$$P_{beat} = \frac{2\pi}{|\delta|} = \frac{2\pi}{|\omega_0 - \omega|}. \quad (4.87)$$

time units.

4.4.3 Exercises

Exercise 4.4.1 For each ODE in parts (a)-(h) below

- Plot the gain function $G(\omega)$ as defined by (4.78) 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.

- (a) $2u''(t) + u'(t) + 8u(t) = f(t)$, $f(t) = 3\sin(4t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 10$.
- (b) $2u''(t) + 6u'(t) + 8u(t) = f(t)$, $f(t) = 3\sin(4t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 10$.
- (c) $2u''(t) + 4u'(t) + 20u(t) = f(t)$, $f(t) = 5\cos(2t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 10$.
- (d) $20u''(t) + 2u'(t) + 100u(t) = f(t)$, $f(t) = \cos(2.25t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 10$.
- (e) $20u''(t) + 2u'(t) + 100u(t) = f(t)$, $f(t) = \sin(10t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 10$. Compare the amplitude of the periodic response to that of (d).
- (f) $u''(t) + u'(t) + 10000u(t) = f(t)$, $f(t) = 5\cos(100t) + \sin(100t)$. Plot $G(\omega)$ for $0 \leq \omega \leq$

200.

- (g) $u''(t) + 10u'(t) + u(t) = f(t)$, $f(t) = 2\cos(2t)$. Plot $G(\omega)$ for $0 \leq \omega \leq 5$.
 (h) $u''(t) + u(t) = f(t)$ (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, in 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 it in the range $0 \leq \omega \leq 10^6$, and then 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 $u(0) = 0, u'(0) = 0$.

- (a) Plot the gain function $G(\omega)$ as defined by (4.78) 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 $G(t)$.

■

Exercise 4.4.5 Suppose a spring-mass-damper system with mass m , damping c , and spring constant k exhibits resonance. Show if resonance occurs then the maximum value for the gain function is $\frac{1}{c\omega_{nat}}$, where $\omega_{nat} = \frac{\sqrt{4km - c^2}}{2m}$ is the system's natural frequency (recall (4.35)). ■

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 $u''_p(t)$ of 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 below.)

As shown previously in this section, the periodic response is $u_p(t) = A\cos(\omega t) + B\sin(\omega t)$ (this was (4.73)) where A and B are given by (4.76).

1. Show that magnitude of $u'_p(t)$ is given by $C\omega G(\omega)$ where $G(\omega)$ is defined by (4.78).
2. Show that the value of $\omega > 0$ that maximizes the amplitude of $u'_p(t)$ is $\omega'_{res} = \sqrt{k/m}$ (doesn't depend on c !) Compare to the resonant frequency $\omega_{res} = \sqrt{k/m - c^2/2m^2}$ at which the amplitude of $u_p(t)$ is maximized, in particular, when c is close to zero.
3. Repeat this analysis to find that 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 Engineer's 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

$$Q = \frac{G(\omega_{res})}{\omega_+ - \omega_-} \quad (4.88)$$

where ω_{res} is the resonant frequency of the system, $G(\omega)$ is the gain function defined in (4.78), and ω_-, ω_+ are those frequencies that are, respectively, less than or greater than ω_{res} such that

$$G(\omega_-) = G(\omega_+) = \frac{G(\omega_{res})}{\sqrt{2}}. \quad (4.89)$$

The larger the value of Q , the sharper the resonance peak.

For each system below, compute $G(\omega)$, plot it on the given range, then solve (4.89) for each of ω_- and ω_+ , then compute Q for the system using (4.88).

- (a) $m = 25, c = 10, k = 26, 0 \leq \omega \leq 5$.
- (b) $m = 2, c = 1, k = 20, 0 \leq \omega \leq 10$.
- (c) $m = 2, c = 0.1, k = 20, 0 \leq \omega \leq 10$.
- (d) $m = 2, c = 0.01, k = 20, 0 \leq \omega \leq 10$.
- (e) $m = 2, c = 0, k = 20, 0 \leq \omega \leq 10$. (Based on (b)-(d), how should Q here be defined?)

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 roof mass, suppose initial data $u(0) = 0, u'(0) = 0$ 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 Hz.)

- (a) What is the resonant frequency for this building? What is the maximum possible amplitude of the building's displacement? Hint: start by computing and plotting $G(\omega)$.
- (b) What is the maximum possible amplitude of the building's acceleration for frequencies in this 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.
- (c) Suppose the acceleration of the building's periodic motion must be kept below 6 meters per second squared for frequencies between 0 and 3 Hz. 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 (4.87) to explain the graph, in particular, the period of the beats.

- (a) $u''(t) + u(t) = \cos(0.9t), u(0) = u'(0) = 0$, on the range $0 \leq t \leq 250$.
- (b) $u''(t) + u(t) = \cos(1.2t), u(0) = u'(0) = 0$, on the range $0 \leq t \leq 100$.
- (c) $u''(t) + 4u(t) = \cos(1.9t), u(0) = 0, u'(0) = 0$, on the range $0 \leq t \leq 200$.
- (d) $u''(t) + 4u(t) = \cos(1.99t), u(0) = 0, u'(0) = 0$, 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.

- (a) 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 as

$$|u(t)| \leq \frac{C}{m\omega_0|\omega_0 - \omega|}.$$

Thus the amplitude of the beat is inversely proportional to $1/|\omega_0 - \omega|$.

- (c) Verify this result by checking against the graph of $u(t)$ from Example 4.27. ■

Exercise 4.4.11 Lightly driven systems can exhibit the beat phenomena, 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 $u(0) = 0, u'(0) = 0$. 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 for the undamped system predicted by (4.87)? ■

4.5 Scaling and Nondimensionalization for ODE's

4.5.1 Motivation: Nonlinear Springs

The spring-mass models 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. But for all materials this model has limits, beyond which the relationship becomes nonlinear. Consider, for example, a so-called *hard spring* in which the force-displacement relationship is

$$F = k_1x + k_2x^3, \quad (4.90)$$

where k_1 and k_2 are nonnegative constants. The case $k_2 = 0$ in (4.90) 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 (4.90) 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 (4.90) are always of the same sign, so when $|x|$ is large the force from (4.90) substantially exceeds that predicted by the linear Hooke's law. This is the basis for the term "hard spring."

Reading Exercise 100 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 (4.90) 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.

Let's replace the linear Hooke's Law with (4.90) in the undamped spring-mass-damper model (4.5) developed in Section 4.1. The same reasoning that led to (4.4) now yields a nonlinear second

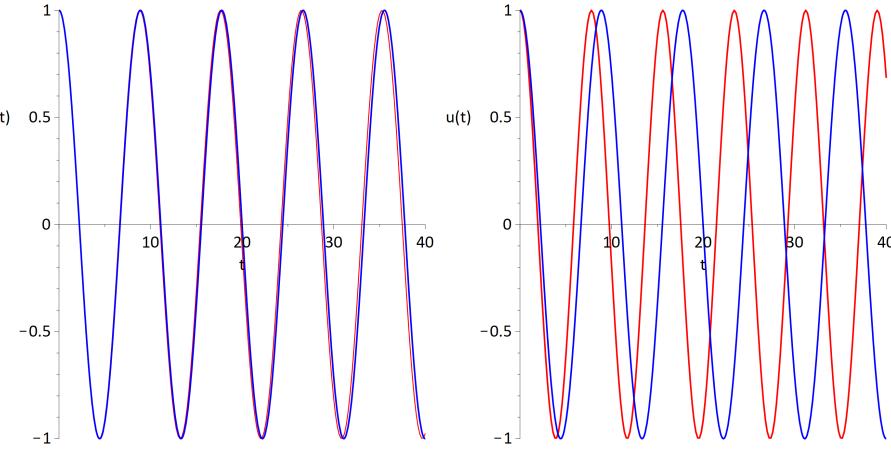


Figure 4.26: Left panel: Solution to (4.91) with $m = 10, k_1 = 5, k_2 = 0.1$ and initial data $u(0) = 1, u'(0) = 0$ (red) and solution to (4.91) with cubic term omitted (blue). Right panel: Same, but with $k_2 = 2$.

order ODE

$$mu''(t) + k_1u(t) + k_2u^3(t) = 0 \quad (4.91)$$

with $u(t)$ as the displacement of a mass m at the end of a spring that obeys (4.90). In some settings the model (4.91) might be more accurate than the ODE (4.4) based on Hooke's Law, but with one big drawback: The ODE (4.91) has no analytical solution. As a result, we would be forced to use numerical or qualitative techniques to analyze (4.91). But if the nonlinear term $k_2u^3(t)$ is sufficiently small as in Reading Exercise 100, perhaps the linear ODE $mu''(t) + k_1u(t) = 0$ 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 (4.91), m, k_1 , and k_2 . Any specific solution to (4.91) 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 the solution is desired may come into play.

Example 4.28 Consider the case $m = 10, k_1 = 5$, and $k_2 = 0.1$, with initial data $u(0) = 1, u'(0) = 0$. In the left panel in Figure 4.26 is shown the solution $u(t)$ to (4.91) in red (computed numerically) and in blue the solution to the linear ODE $mu'' + k_1u = 0$ (cubic term omitted) with the same initial data $u(0) = 1, u'(0) = 0$, on the time interval $0 \leq t \leq 40$. The agreement on this time interval might be considered good enough for some purposes, so the analytical solution to the linear ODE could be used, at least for these parameter values.

In the right panel of Figure 4.26 is shown the solutions to the nonlinear and linear ODE's with the same parameters as above except with $k_2 = 2$. In this case the solutions are actually 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 these circumstances. ■

With the six parameters m, k_1, k_2, u_0, v_0 (we could have thrown in damping too!) 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 is needed 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 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 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 $u(0) = u_0, u'(0) = 0$. The solution to this ODE is $u(t) = u_0 \cos(t\sqrt{k/m})$. This function is oscillatory with period $P = 2\pi\sqrt{m/k}$. In this case we can say that the *characteristic time scale* for this problem is $\sqrt{m/k}$ (dimensionless constants like 2π are usually omitted). For example, if $m = 4$ and $k = 1$ then the time scale is $\sqrt{m/k} = 2$; if m is in kilograms and k is in Newtons per meter this characteristic time scale is in seconds. If we “measured” the solution $u(t) = \cos(t/2)$ (period $4\pi \approx 12.6$) every 2 seconds we’d see the smooth and gradual change in $u(t)$. On the other hand, measuring $u(t)$ every 1000 seconds would produce a meaningless cloud of data with no pattern. Measuring $u(t)$ every 10^{-6} seconds would make it seem that $u(t)$ isn’t changing at all from one data point to the next and nothing at all would be seen without taking a lot of data. Measuring $u(t)$ every 2 seconds is about right to see “what’s going on” with this function.

This time scale 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 or solution also has its own scale. In this example $u(t) = u_0 \cos(t\sqrt{k/m})$ 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.³

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 numerics. 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 to decide how much data to collect. The spring-mass system in the Modeling Project “Parameter Estimation with Second Order ODE’s” of Section 4.6 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 magnitude, this can mean trouble for numerical algorithms. Knowing this can help head off difficulty.

But 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. This 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 $u(0) = u_0, u'(0) = 0$. 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

³A parsec is a unit of length, not time.

hypothesize

$$t_c = m^\alpha k^\beta u_0^\gamma \quad (4.92)$$

for some constants α, β , and γ . Since $[t_c] = T$, dimensional consistency in (4.92) demands that $[t_c] = [m]^\alpha [k]^\beta [u_0]^\gamma$ or (note $T = M^0 L^0 T^1$)

$$M^0 L^0 T^1 = M^\alpha M^\beta T^{-2\beta} L^\gamma = M^{\alpha+\beta} L^\gamma T^{-2\beta}.$$

Matching exponents leads to $\alpha + \beta = 0$, $\gamma = 0$, and $-2\beta = 1$, three equations in unknowns α, β, γ . The unique solution is $\alpha = 1/2, \beta = -1/2, \gamma = 0$. We conclude from (4.92) 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 101 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.

A problem 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. The relevant dimensions are

$$[m] = M, \quad [c] = MT^{-1}, \quad [k] = MT^{-2}.$$

For a characteristic time scale $t_c = m^\alpha c^\beta k^\gamma$ the constants α, β , and γ must satisfy

$$M^0 L^0 T^1 = (M^\alpha)(M^\beta T^{-\beta})(M^\gamma T^{-2\gamma}) = M^{\alpha+\beta+\gamma} L^0 T^{-\beta-2\gamma}.$$

This forces

$$\begin{aligned} \alpha + \beta + \gamma &= 0, \\ -\beta - 2\gamma &= 1. \end{aligned}$$

There are infinitely solutions to these equations. Specifically, we may (for example) choose α arbitrarily and then solve for β and γ in terms of α as

$$\beta = 1 - 2\alpha \quad \text{and} \quad \gamma = \alpha - 1. \quad (4.93)$$

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 (4.93), this problem has infinitely many time scales! With β and γ as in (4.93) any characteristic time scale t_c has the form

$$t_c = m^\alpha c^{1-2\alpha} k^{\alpha-1} = \frac{c}{k} \left(\frac{mk}{c^2} \right)^\alpha \quad (4.94)$$

for any choice of α . Note that on the right in (4.94) the quantity c/k has dimension time, while the parenthesized quantity mk/c^2 in (4.94) is dimensionless, and so is $(mk/c^2)^\alpha$ for any α . This makes it easy to see that the right side of (4.94) has the dimension of time for any α .

That there are infinitely many time scales to choose from might seem a bit disconcerting, but this is not uncommon. Different choices for α in (4.94) emphasize different aspects of how the solution to $mu''(t) + cu'(t) + ku(t) = 0$ evolves in time. ■

Reading Exercise 102 Show that the choice $\alpha = 1/2$ in (4.94) yields the previous time scale $t_c = \sqrt{m/k}$ that captures the period of the oscillatory portion of the mass motion.

Reading Exercise 103 Show that the choice $\alpha = 1$ in (4.94) yields time scale $t_c = m/c$. What aspect of the mass's motion does this emphasize? Hint: The general solution to $mu'' + cu' + ku = 0$ is of the form

$$u(t) = c_1 e^{-\frac{ct}{2m}} \cos(\omega t) + c_2 e^{-\frac{ct}{2m}} \sin(\omega t).$$

4.5.3 Rescaling Variables and Nondimensionalizing ODE's: Examples

Scaling or *nondimensionalizing* the variables in an ODE can greatly simplify matters and reduce the number of parameters involved. It can highlight commonality in problems of vastly different scales, e.g., a skyscraper swaying in the breeze and a quartz crystal in an electronic circuit. It may also give insight into what terms in an ODE are important and which might be ignored, potentially simplifying analysis of the equation. This process makes use of the characteristic variable scales we can find using the techniques above. Let's begin with two examples.

■ **Example 4.31** Recall the logistic equation (1.10) for population growth,

$$\frac{du}{dt} = ru(t)(1 - u(t)/K). \quad (4.95)$$

Here $u(t)$ is the population of some species as a function of time, r is the species growth rate, and K is the carrying capacity of the environment, the maximum number of individuals that can be supported. Let us rescale/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 (4.95) dimensionless, so that $(1 - u/K)$ is well-defined and dimensionless (since 1 is dimensionless). Then $[du/dt] = NT^{-1}$ and dimensional consistency in (4.95) demands $[r] = T^{-1}$.

Characteristic Scales

Based on the above analysis there are two parameters, r and K in (4.95), with dimensions $[r] = T^{-1}$ and $[K] = N$. To form characteristic time and population scales t_c and u_c , respectively, consider the expressions

$$t_c = r^\alpha K^\beta \quad \text{and} \quad u_c = r^\gamma K^\delta.$$

For $[t_c] = T$ dimensional consistency leads to $T = (T^{-1})^\alpha N^\beta = T^{-\alpha} N^\beta$ which is easily seen to force $\alpha = -1$ and $\beta = 0$, so $t_c = 1/r$. For $[u_c] = N$ this leads to $N = (T^{-1})^\gamma N^\delta = T^{-\gamma} N^\delta$, which is easily seen to force $\gamma = 0$ and $\delta = 1$, so $u_c = K$. The characteristic scales for the independent (time) and dependent (population) variables here are

$$t_c = 1/r \quad \text{and} \quad u_c = K. \quad (4.96)$$

This may seem rather clear after the fact.

By performing the analysis that led to (4.96), we've already deduced something important about this ODE: it can be expected 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.

Nondimensional Variables

The next step is to nondimensionalize both the independent (time) and dependent (population) variables. Define a new “time like” independent variable τ as

$$\tau = t/t_c = rt \quad \text{or} \quad t = t_c \tau = \tau/r \quad (4.97)$$

since $t_c = 1/r$. 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 “nondimensional time.” For example, if the ODE has 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 .

Define a new nondimensional dependent variable, a “nondimensional population” \bar{u} , by defining

$$\bar{u} = u/u_c = u/K \quad \text{or} \quad u = u_c\bar{u} = K\bar{u} \quad (4.98)$$

since $u_c = K$. Again, this simply changes the scale for measuring population. If $u_c = 10^6$ then \bar{u} is the population measured “in millions.”

Nondimensionalizing the ODE

The function $u(t)$ will be replaced by its equivalent version with nondimensional variables, $\bar{u}(\tau)$. The relation between $\bar{u}(\tau)$ and $u(t)$ is simple: $\bar{u}(\tau) = u(t)/u_c$, or

$$\bar{u}(\tau) = u(t)/K. \quad (4.99)$$

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 (4.99) by u_c and find

$$u(t) = K\bar{u}(\tau). \quad (4.100)$$

The next step is to replace $u(t)$ in the logistic ODE (4.95) 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 (4.100) and differentiate both sides with respect to t to find

$$\begin{aligned} \frac{du}{dt}(t) &= \frac{d}{dt}(K\bar{u}(\tau)) \\ &= K \frac{d\bar{u}}{d\tau}(\tau) \frac{d\tau}{dt} \quad (\text{use the chain rule}) \\ &= \frac{K}{t_c} \frac{d\bar{u}}{d\tau}(\tau) \quad (\text{since } d\tau/dt = 1/t_c) \\ &= rK \frac{d\bar{u}}{d\tau}(\tau) \quad (\text{since } t_c = 1/r). \end{aligned} \quad (4.101)$$

To nondimensionalize the ODE (4.95) use (4.100) to replace each $u(t)$ with $K\bar{u}(\tau)$ and use (4.101) 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 simple algebra shows that

$$\frac{d\bar{u}}{d\tau}(\tau) = \bar{u}(\tau)(1 - \bar{u}(\tau)). \quad (4.102)$$

All of the physical parameters have disappeared!

The ODE (4.102) is a nondimensional rescaling of the original ODE (4.95). We can move back and forth between solutions $\bar{u}(\tau)$ to (4.102) and solutions $u(t)$ to (4.95) via 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 to (4.102) is fairly easy to find via separation of variables, and is given by $\bar{u}(\tau) = 1/(1 + Ce^{-\tau})$. Then 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, another advantage of this nondimensionalization is this: Aside from the rescaling in time and population, the solution to (4.95) will behave, qualitatively, like the solution to (4.102), and the latter equation might be easier to analyze. But the real power of rescaling is best illustrated by the following extension of this example. ■

Reading Exercise 104 Suppose $u(t) = t^2$, $r = 1/3$ (so $t_c = 3$), and $u_c = K = 7$. Use (4.99) to write out $\bar{u}(\tau)$ explicitly as a function of τ . Then verify (4.101).

■ **Example 4.32** Let's reconsider the logistic ODE of the last Example 4.31, but this time with constant harvesting at a rate of h individuals per unit time, which can be modelled by subtracting an h from the right side of (4.95) to obtain

$$\frac{du}{dt} = ru(t)(1 - u(t)/K) - h. \quad (4.103)$$

Unlike equation (1.12), the harvesting here is not proportional to $u(t)$, but at a flat rate. The ODE (4.103) does have an analytical solution, though it is significantly more complicated than the solution to the standard logistic equation (4.95) in which $h = 0$. Under what conditions can the harvesting rate h be taken as zero and the solution to (4.95) used as a reasonable approximation to the solution to (4.103)? This is can be illuminated by nondimensionalizing (4.103).

Characteristic Scales

We begin by finding characteristic scales for the independent and dependent variables. As in Example 4.31, $[r] = T^{-1}$ 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 $T = T^{-\alpha-\gamma} N^{\beta+\gamma}$. This forces $-\alpha - \gamma = 1$ and $\beta + \gamma = 0$. There are infinitely many solutions to these two equations in three unknowns. Let us take γ as a free variable that we can choose and then solve for $\alpha = -1 - \gamma$ and $\beta = -\gamma$. Thus anything of the form

$$t_c = r^{-1-\gamma} K^{-\gamma} h^\gamma = \frac{1}{r} \left(\frac{h}{rK} \right)^\gamma \quad (4.104)$$

can be used as a characteristic time scale, for any γ . Note on the right side in (4.104) the quantity $1/r$ has the dimension time, while $h/(rK)$ and any power thereof is dimensionless; a similar structure was present in equation (4.94). 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 105 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 105 yields the same characteristic population scale $u_c = K$ that was found in Example 4.31. Let's go with these choices for t_c and u_c , and see where it leads.

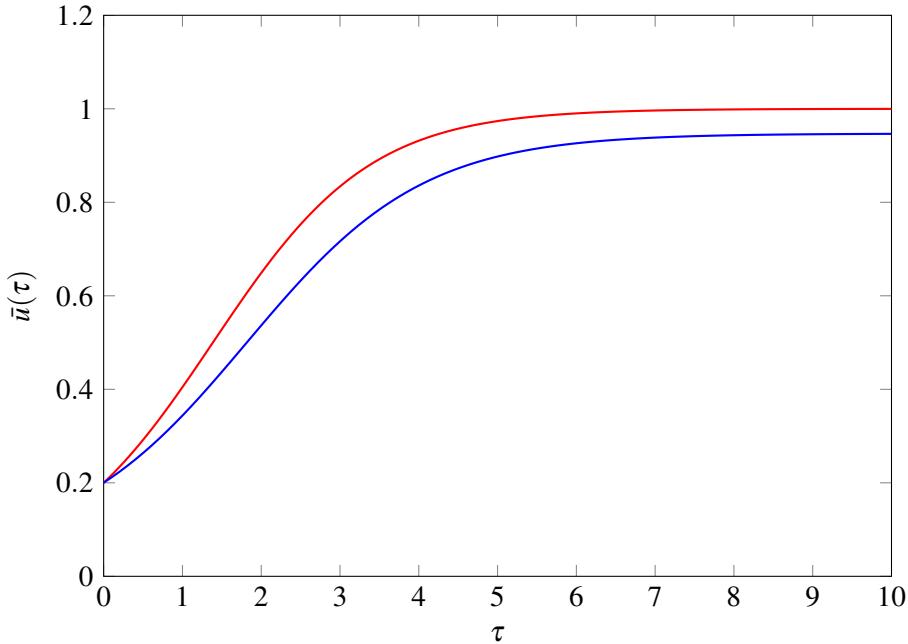


Figure 4.27: Solution to (4.105) with $\varepsilon = 0$ (red) and with $\varepsilon = 0.05$ (blue), both with $\bar{u}(0) = 0.2$.

Rescaling the ODE

With the choices $u_c = K$ and $t_c = 1/r$ equations (4.99), (4.100), and (4.101) are still 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 (4.103) 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

$$\frac{d\bar{u}}{d\tau}(\tau) = \bar{u}(\tau)(1 - \bar{u}(\tau)) - \varepsilon. \quad (4.105)$$

where $\varepsilon = \frac{h}{rK}$; note ε is dimensionless, as it must be in order to be subtracted from \bar{u} . Compare (4.105) to (4.102). The solutions to (4.105) are in a correspondence to those of (4.103) via $t = t_c\tau = \tau/r$ and $u(t) = u_c\bar{u}(\tau) = K\bar{u}(\tau)$.

Interpretation of the Nondimensional ODE and Conclusions

The nondimensionalized ODE (4.105) let's 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 (4.105) should be close to the solution in which $\varepsilon = 0$, which corresponds to no harvesting. To illustrate, Figure 4.27 shows the solution to (4.105) when $\varepsilon = 0.05$ (in blue) and $\varepsilon = 0$ (red), both with $\bar{u}(0) = 0.2$.

Based on Figure 4.27, 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 (4.103) the condition $\varepsilon \leq 0.05$ becomes the condition

$$\frac{h}{rK} \leq 0.05 \quad (4.106)$$

since $\varepsilon = \frac{h}{rK}$. Equation (4.106) 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 . And if $\varepsilon \leq 0.05$ does not produce sufficiently good agreement then a smaller value of ε can be used, but in any case it is a bounded on the dimensionless quantity $\frac{h}{rK}$ that is relevant for ignoring the harvesting term. ■

Reading Exercise 106 Suppose $K = 10^6$ and $r = 0.03$ (units reciprocal years). What value of h will satisfy (4.106)? 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?

4.5.4 The General Outline for Nondimensional Rescaling

The techniques used in Examples 4.31 and 4.32 work more generally. 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, e.g., masses, growth rates, etc. The initial condition(s) may also come into play, if any are specified. Find the dimension of each of these quantities.
2. Find the possibilities for the characteristic time scale t_c of the problem, with t_c as a product of powers of the ODE parameters. There may be a unique time scale, as in Example 4.31, or many, as in Example 4.32. Do the same to 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 already provides valuable insight.
3. Define a nondimensional time variable τ according to

$$\tau = t/t_c \quad \text{or} \quad t = t_c\tau. \quad (4.107)$$

Also define a nondimensional dependent variable \bar{u} as

$$\bar{u}(\tau) = u(t)/u_c \quad \text{or} \quad u(t) = u_c\bar{u}(\tau). \quad (4.108)$$

From $u(t) = u_c\bar{u}(\tau)$, $\tau = t/t_c$, and the chain rule it follows that

$$\frac{du}{dt} = \frac{u_c}{t_c} \frac{d\bar{u}}{d\tau}. \quad (4.109)$$

If the second derivative d^2u/dt^2 is needed in terms of \bar{u} then differentiating both sides of (4.109) with respect to t again and using the chain rule shows that

$$\frac{d^2u}{dt^2} = \frac{u_c}{t_c^2} \frac{d^2\bar{u}}{d\tau^2}. \quad (4.110)$$

Higher derivatives can be computed too, of course.

4. In the original unscaled ODE use (4.108) and (4.109) to replace $u(t)$ by $u_c\bar{u}(\tau)$ and du/dt by $\frac{u_c}{t_c} \frac{d\bar{u}}{d\tau}$. 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 nondimensional.
5. Simplify the nondimensional ODE. Examine the remaining terms in the nondimensional ODE (like ε in (4.105)). 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.

It should be noted that this process is applicable to ODE’s that have independent variables other than t , and in fact is applicable to partial differential equations as well.

4.5.5 Back to the Hard Spring

As a last example, let's return to the "hard spring" model (4.91), 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, 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 (4.91). 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 $[k_2]$ simply note that from (4.90), $[k_2]$ must have dimension force divided by length cubed. Based on Figure 4.26 it should also be expected 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 just work under the assumption that $v_0 = 0$.

The first step is to construct a characteristic time scale of the form

$$t_c = m^\alpha k_1^\beta k_2^\gamma u_0^\delta. \quad (4.111)$$

It's straightforward to verify (see Exercise 4.5.10) by matching dimensions on the left and right in (4.111) that any characteristic time scale is of the form

$$t_c = m^{1/2} k_1^{-1/2-\gamma} k_2^\gamma u_0^{2\gamma} = \sqrt{\frac{m}{k_1}} \left(\frac{k_2 u_0^2}{k_1} \right)^\gamma \quad (4.112)$$

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 (4.112) is dimensionless. This same general structure for t_c was present in (4.94) and (4.104), in which there were infinitely many choices for t_c , each of the form $t^* q^\gamma$ where t^* has the dimension of time, γ 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 (4.112). Moreover, $\sqrt{m/k}$ captures the period of the linear system, and, based on Figure 4.26, 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 (4.112)

$$t_c = \sqrt{m/k_1}. \quad (4.113)$$

If this time scale provides no insight we'll amend it and try again.

For a characteristic length scale u_c consider

$$u_c = m^\alpha k_1^\beta k_2^\gamma u_0^\delta. \quad (4.114)$$

Analysis similar to that for t_c leads to

$$u_c = u_0 \left(\frac{k_2 u_0^2}{k_1} \right)^\gamma \quad (4.115)$$

for any choice of γ , with the same dimensionless quantity $k_2 u_0^2 / k_1$ again—there is a deeper reason! One obvious choice to use $\gamma = 0$ and so take

$$u_c = u_0 \quad (4.116)$$

as before. If the rescaled ODE yields no insight we can go back and try something else for u_c .

The Nondimensional ODE

In accord with the general procedure of (4.107) set $\tau = t/t_c = t\sqrt{k_1/m}$ (or $t = \tau\sqrt{m/k_1}$). Use (4.108) 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 (4.110) it follows that

$$\frac{d^2u}{dt^2}(t) = \frac{u_0 k_1}{m} \frac{d^2\bar{u}}{d\tau^2}(\tau). \quad (4.117)$$

With these substitutions the ODE (4.91) for $\bar{u}(\tau)$ can be written as

$$\frac{d^2\bar{u}}{d\tau^2} + \bar{u} + \varepsilon \bar{u}^3 = 0 \quad (4.118)$$

where

$$\varepsilon = \frac{k_2 u_0^2}{k_1}. \quad (4.119)$$

The quantity ε in (4.119) is the same dimensionless quantity that appeared in (4.112) and (4.115). 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 (4.91) with $u(0) = u_0$ and the solution $\bar{u}(\tau)$ to (4.118) with $\bar{u}(0) = 1$ are in precise correspondence, via 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 (4.118) is $u(t) = 0$ for all t , uninteresting! With this assumption, $\varepsilon = 0$ corresponds to $k_2 = 0$, so there is no cubic term in (4.91) or (4.118). But if $\varepsilon > 0$ and is sufficiently close to zero then the solution to (4.118) should be close to the solution to $\frac{d^2\bar{u}}{d\tau^2} + \bar{u} = 0$ in some reasonable sense, and so the solution to (4.91) 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 4.26 shown as the red curve is the solution to (4.91) with $m = 10, k_1 = 5$, and $k_2 = 0.1$, which corresponds to $\varepsilon = 0.02$, and in blue is the solution with $k_2 = 0$, corresponding of course to $\varepsilon = 0$. In the right panel is the same but with $k_2 = 2$, corresponding to $\varepsilon = 0.4$.

This gives a quantitative and efficient way to determine when the cubic term in (4.91) 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 (4.118), or equivalently the condition

$$\frac{k_2 u_0^2}{k_1} \leq \varepsilon_0 \quad (4.120)$$

for the ODE (4.91), where ε_0 is some threshold that depends on the level of agreement that is desired. This threshold can be determined numerically, by solving (4.118) 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 (4.120). 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 (4.118) is $0 \leq \tau \leq T/t_c$.

■ **Example 4.33** Here's an illustration. Consider (4.91) 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 (4.91) 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 (4.118) is $\bar{u}(\tau) = \cos(\tau)$. A numerical solution to (4.118) with $\varepsilon = 0.01$ is shown in Figure 4.28 in red, with If this agreement is deemed good enough then $\varepsilon_0 = 0.01$ can be used in (4.120). With $m = 200, k_1 = 10$, and $u_0 = 2$ the condition for k_2 in (4.91) becomes $\frac{k_2 u_0^2}{k_1} \leq 0.01$, or $k_2 \leq 0.025$. Notice that in all the analysis above for the hard spring, the mass m was not relevant! ■

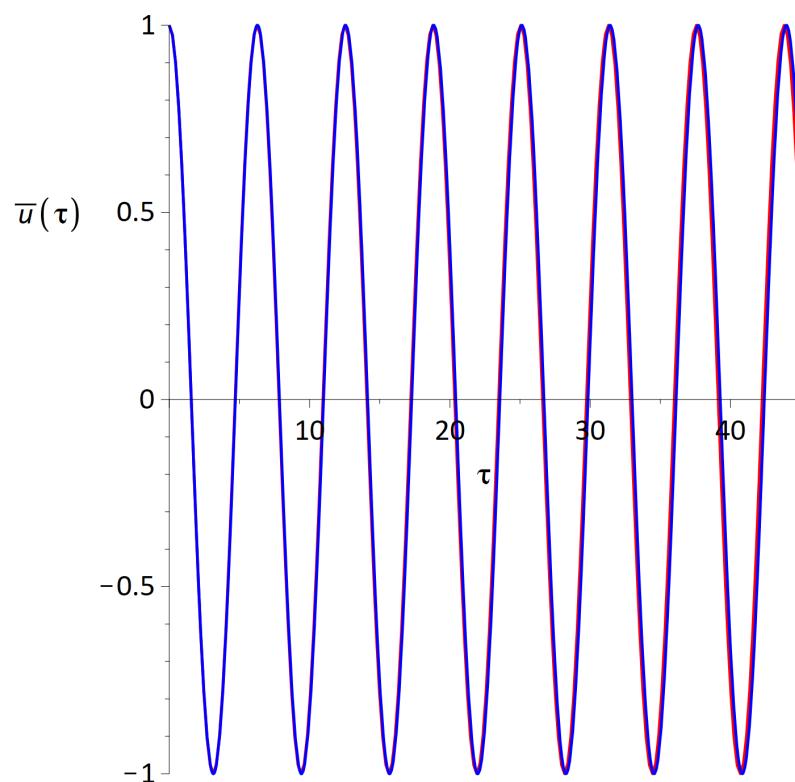


Figure 4.28: Solution to (4.118) with $\varepsilon = 0.01$ and initial data $u(0) = 1, u'(0) = 0$ (red) and solution to (4.118) with cubic term omitted (blue).

Summary

Rescaling via the introduction of dimensionless variables can make the analysis and solution of ODE's (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 the problem. For example, when solving (4.91) 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 (4.118) 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 scale 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 size jet? Dimensional analysis of the type we've done can help assure that it will. See [23] for more on this topic and dimensional analysis in general.

4.5.6 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 an 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 temperature scale u_c for this ODE in terms of the parameters k and A , then nondimensionalize the ODE. What does an initial condition $u(0) = u_0$ becomes, 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 scales t_c and velocity scale v_c for this ODE in terms of P and k , and then nondimensionalize the ODE. To what does the initial condition $v(t_0) = 0$ become? 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/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 substance in the tank at time t , measured on a per mass basis. Find the dimensions of r, c_1 , and V , then use this 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 this 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, γ . Find the dimension of each of R, g , and m , and show that the characteristic time scale for this problem is given by $t_c = \sqrt{R/g}$ (and so 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 equation for $\theta(t)$ that the pendulum makes with 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 $\theta_c = 1$ 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 charge scale q_c for this circuit in terms of R, C , and q_0 , 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 is begin 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 the 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

$$\frac{du}{dt} = -ru \left(1 - \frac{u}{K}\right) \left(1 - \frac{u}{P}\right) \quad (4.121)$$

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 (4.121) and verify that if $0 < u < P$ 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 $t_c = 1/r$. 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 (4.121) 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 (4.121) 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 (4.121), of the form

$$\frac{du}{dt} = -ru \left(1 - \frac{u}{K}\right) \left(1 - \frac{u}{P}\right) - hu \quad (4.122)$$

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

$$\frac{d\bar{u}}{d\tau} = -\bar{u}(1-\bar{u}) \left(1 - \frac{K}{P}\bar{u}\right) - \varepsilon\bar{u} \quad (4.123)$$

where $\varepsilon = h/r$.

- (h) Suppose that in the harvested ODE (4.122) we have $P = K/10$. Write out the ODE (4.123) in this case. Then show that if $\varepsilon > 2.025$ then all solutions to (4.123) that start with $u(0) > 0$ converge to zero; it might help to find the fixed points for (4.123) and then sketch a phase portrait that shows the dependence on ε . What bound does this put on h/r in order to avoid extinction in (4.122)?

■

Exercise 4.5.10 In Example 4.33, show that any characteristic time scale t_c is of the form given in (4.112), and any characteristic length scale is of the form given in (4.115). ■

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 (which is the negative direction) when $v > 0$. From $F = ma$ we find

$$m \frac{dv}{dt} = mg - k_1v(t) - k_2v^2(t). \quad (4.124)$$

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

$$t_c = \frac{m}{k_1} \left(\frac{mgk_2}{k_1^2} \right)^\delta$$

for some choice of γ (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

$$v_c = \frac{mg}{k_1} \left(\frac{mgk_2}{k_1^2} \right)^\delta$$

for some choice of δ .

- (d) Nondimensionalize (4.124) using $t_c = \frac{m}{k_1}$ ($\delta = 0$ in (b)) and $v_c = \frac{mg}{k_1}$ ($\delta = 0$ in (c)). Show that this leads to an ODE for \bar{v} of the form

$$\frac{d\bar{v}}{d\tau} = 1 - \bar{v} - \varepsilon\bar{v}^2 \quad (4.125)$$

where $\varepsilon = mgk_2/k_1^2$ (dimensionless).

Find the analytical solution to (4.125) 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 (4.125) with $\bar{v}(0) = 0$ for each of $\varepsilon = 0.01, 0.1, 1.0$ and

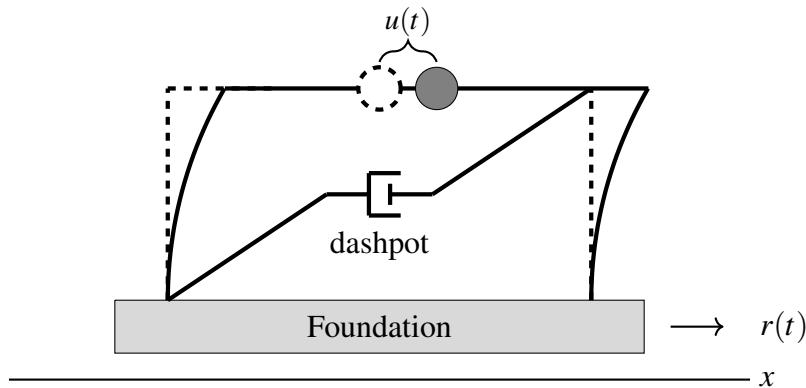


Figure 4.29: Simplified one-story building.

plot each along with $\bar{v}_0(\tau)$ for $0 \leq \tau \leq 5$. Experiment. How small must ε be before the solution to (4.125) agrees well with \bar{v}_0 on this nondimensional time interval?

Based on this, what is a reasonable quantitative criteria (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 $\delta = -1/2$ for t_c and $\delta = -1/2$ for v_c leads to a rescaled ODE of the form

$$\frac{d\bar{v}}{d\tau} = 1 - \varepsilon\bar{v} - \bar{v}^2. \quad (4.126)$$

What is ε here? Mimic the computations in part (d) above to find a reasonable quantitative criteria (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 a variety of ideas seen in this chapter. The last two concern two different approaches to modeling a swinging pendulum, one based on conservation of energy, the second based on Newton's Second Law of Motion. The second approach also 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 motion of the building will be ground motion, instead of a force directly applied to the building roof mass.

See Figure 4.29 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 this case, however, the building's foundation (the light gray rectangle) can move with respect to the horizontal (x) axis; this horizontal axis is a fixed inertial frame of reference. 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 given by $u(t) + r(t)$. The 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 $F = ma$.

Modeling Exercise 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.

Modeling Exercise 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.

Modeling Exercise 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 1 and 2 to justify the ODE

$$mu''(t) + cu'(t) + ku(t) = -mr''(t). \quad (4.127)$$

Modeling Exercise 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 (4.127)? Repeat this thought experiment if $r(t) = vt + b$ (the foundation is moving a constant speed).

Modeling Exercise 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 (4.127) 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 Show that this structure is underdamped. What is its natural frequency? Then repeat Modeling Exercise 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 5?

Modeling Exercise 7 Compute the roof's periodic displacement response $u_p(t)$ to foundation motion $r(t) = 0.02 \cos(\omega t)$ with 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 Hz). What is the maximum displacement in this range?

Modeling Exercise 8 Suppose the building should not experience a displacement of more than 0.05 meters when being driven at any frequency in the range 0 to 5 Hz. What is the smallest damping coefficient (with m and k as already given) that will suffice?

Modeling Exercise 9 The displacement $u(t)$ of the building roof may not be the only issue; the acceleration experienced by the building (and occupants) might be a concern. Recall that the position of the roof mass is given by $u(t) + r(t)$. 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 in this frequency range?

Modeling Exercise 10 Suppose that in the setting of Modeling Exercise 9 the acceleration experienced by the building/occupants should not exceed 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: Stayed Tuned—RLC Circuits and Radio Tuning

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 kHz to 1700 kHz. 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. But if 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. However, 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 as a whole.

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 kHz to 1700 kHz. We wish to "tune in" to 910 AM, that is, to a signal being carried at 9.1×10^5 Hz (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, e.g. amplification.

Modeling Exercise 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 capacitor so that the resonant frequency of the circuit is 910 kHz (5.713×10^6 radians per second). The result of Exercise 4.4.3 may be helpful.

Modeling Exercise 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 $1/\sqrt{LC}$.

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

$$\frac{\sqrt{4L/C - 2R^2}}{2L} \approx \frac{1}{\sqrt{LC}} - \frac{\sqrt{C}}{4L^{3/2}}R^2. \quad (4.128)$$

How does this justify the use of $1/\sqrt{LC}$ as an approximation to the resonant frequency when R is small?

Modeling Exercise 3 With the values of L, R , and the value of C you found from Modeling Exercise 1, 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 4 Find a general solution $q_h(t)$ to the unforced/homogeneous ODE you found in Modeling Exercise 3, in a real-valued form as per (4.39); 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 5 Compute the gain function $G(\omega)$ for this circuit using the values of R, L, C from Modeling Exercise 1, 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 where it should be!

Modeling Exercise 6 The periodic current in the circuit, $I_p(t)$, is given by $I_p(t) = q'_p(t)$. 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. Let $H(\omega)$ denote the amplitude of this sinusoidal voltage. Compute $H(\omega)$ as a function of ω , and plot on the range $0 \leq \omega \leq 10^7$.

Modeling Exercise 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 $V(t) = \sin(\omega_{rest})$, 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 kHz). 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 8 One way to sharpen up the response of this RLC filter circuit (so it screens out unwanted frequencies even more effectively) is to decrease R . Redo the above analysis of Modeling Exercises 1 to 7 with $R = 1$ ohm, and compare the value of $H(\omega_{res} \pm (2\pi)(1.0 \times 10^4))$ to that from Modeling Exercise 7.

4.6.3 Project: Parameter Estimation with Second Order ODE's

This modeling project is based on the SIMIODE Modeling Scenario “Models Motivating Second Order,” [106].

One of the fundamental tasks of mathematical modeling is *validation*, in which the predictions of a model are compared to the “real-world.” The goal is to determine if 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 instead 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 4.30, in which a mass was suspended on a spring attached to a rod on a stand. Data on the vertical motion 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 total number of 1,500 observations. An initial portion of “dirty” 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, to some approximation, 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 [6].

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. Computations are facilitated by the Maple, Mathematica, and Matlab notebooks posted on the book website.

Modeling Exercise 1 As mentioned above, the data consists of the distance from the detector head to the base of the mass hanging from the spring. The mean distance is not zero, of course. 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 this time period. Compute this mean value and then form a “centered” data set by subtracting u^* from each position measurement. Plot this centered data to make sure it looks reasonable.

This 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



Figure 4.30: The apparatus for collecting data on the bouncing spring mass system with the VernierGo!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.

equilibrium, that is, $y(t) = u(t) - u^*$.

The function $y(t)$ should satisfy $my''(t) + cy'(t) + ky(t) = 0$, with $y(0) = y_0$ for some y_0 and $y'(0) = 0$, since the data was edited to that the mass was at a point of maximum excursion from equilibrium.

One way to estimate c and k is from the period of the spring-mass oscillations and the decay rate of the oscillation amplitude. One way to approach this is to note that the solution to (4.4) is given by (4.39) for some choice of d_1 and d_2 . We reproduce that equation here for convenience,

$$y(t) = d_1 e^{-\alpha t} \cos(\omega t) + d_2 e^{-\alpha t} \sin(\omega t) \quad (4.129)$$

where from (4.35) $\alpha = \frac{c}{2m}$ and $\omega = \frac{\sqrt{4mk - c^2}}{2m}$. The value of m is known.

Modeling Exercise 2 What is the initial position $y(0) = y_0$ of the mass in the centered data set? What choice for d_1 in (4.129) is dictated by this y_0 value?

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 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 4 The value of α in (4.129) 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 (4.129) with $d_2 = 0$ and the values you obtained for d_1 and ω ; then adjust α .

Modeling Exercise 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 determining 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 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?

4.6.4 Project: Bike Shock Absorber

Reread Examples 4.1 and 4.59 in which we modeled and did some analysis of a mountain bike front shock. You may also find it helpful to do Reading Exercise 93 and Exercise 4.3.5.

The goal in this project is to design a front shock absorber that, under the conditions of the examples above

- Has as compliant (least stiff) a spring as possible that makes the shock displacement no more than 140 mm when the rider rides off a 1.5 meter drop.
- Is not excessively overdamped (this makes riding on rugged terrain feel harsh) or under-damped (this makes the bike feels too “bouncy.”)

A computer algebra system is highly recommended for this project.

Modeling Exercise 1 Suppose that as modeled in Example 4.1, the ODE that governs the shock displacement $y(t)$ is given by

$$my''(t) + cy'(t) + ky(t) = -mg \quad (4.130)$$

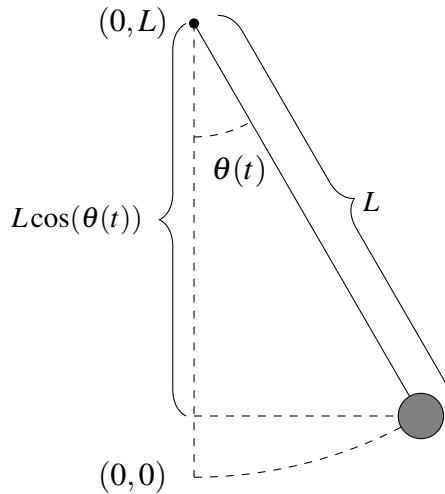


Figure 4.31: Pendulum of length L at angle $\theta(t)$ with respect to vertical.

where $m = 46$ kg and $g = 9.8$. However, let us leave c and k undefined for the moment (as 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 (4.130)? It should depend on k . Use (4.45) to write out a general solution to the homogeneous ODE $my'' + c^*y'(t) + ky(t) = 0$.

Modeling Exercise 2 Use the method of undetermined coefficients to find a particular solution $y_p(t)$ to (4.130); this solution will depend on k . Then write out a general solution to (4.130). The general solution should also depend on k .

Modeling Exercise 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 (4.130) 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 4 Determine, either graphically or analytically, the smallest value $k = k^*$ for k that results in the shock deflecting no more than -0.14 meters. What is the corresponding value for c^* ?

Modeling Exercise 5 With these values for k^* and c^* , 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 fact that rider's legs act as shock absorber's, nor the tires.)

Modeling Exercise 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

Consider a pendulum of length L as depicted in Figure 4.31. 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 ODE's with a simpler linear second order ODE, and compare the solutions.

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 given by the equations

$$x(t) = L \sin(\theta(t)) \quad \text{and} \quad y(t) = L - L \cos(\theta(t)). \quad (4.131)$$

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 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 “zero potential energy” position. 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 2 Equations (4.131) give the position of the bob at any time. Show that the speed v at which the bob is moving is given by

$$v(t) = L|\theta'(t)|.$$

Use this to show that the kinetic energy of the pendulum at any time is given by

$$K(t) = \frac{1}{2}mL^2(\theta'(t))^2.$$

Modeling Exercise 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 1 and 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

$$U + K = U_0 \quad (4.132)$$

where U and K from Modeling Exercises 1 and 2 are functions of time. Use (4.132) to show that $\theta(t)$ obeys the first order differential condition

$$(\theta'(t))^2 = \frac{2g(\cos(\theta(t)) - \cos(\theta_0))}{L} \quad (4.133)$$

with initial condition $\theta(0) = \theta_0$.

Modeling Exercise 4 We can solve (4.133) 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 swinging clockwise ($\theta'(t) < 0$), and so obtain separable ODE’s for these two phases of the pendulum’s motion. However, the resulting ODE’s have no simple analytical solution.

Instead, let’s do this: Differentiate both sides of (4.133) 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

$$\theta''(t) + \frac{g}{L} \sin(\theta(t)) = 0. \quad (4.134)$$

With $\theta(0) = \theta_0$ and $\theta'(0) = 0$ this is equivalent to equation (4.133), in that the same function $\theta(t)$ satisfies both.

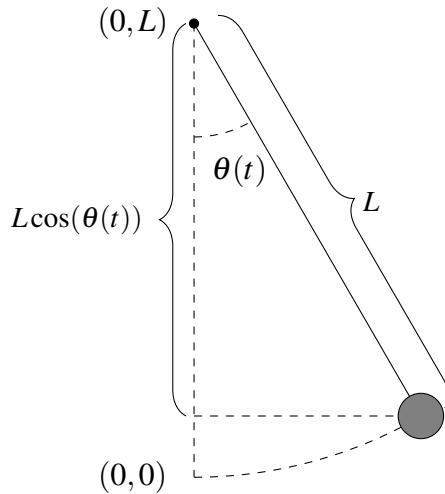


Figure 4.32: Pendulum of length L at angle $\theta(t)$ with respect to vertical.

A Linearized Approximation

Modeling Exercise 5 Unfortunately, (4.134) 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 (4.134) may be approximated with the simpler “linearized” second order differential equation

$$\theta''(t) + \frac{g}{L}\theta(t) = 0. \quad (4.135)$$

Hint: Look at the Taylor’s series for $\sin(\theta)$ about the point $\theta = 0$.

Modeling Exercise 6 Solve (4.134) with parameters $g = 9.8, L = 1$ and initial conditions $\theta(0) = 0.1, \theta'(0) = 0$; you’ll need to use a numerical ODE solver. Plot the solution on the interval $0 \leq t \leq 5$. Repeat with equation (4.135) and compare plots. Is the approximation accurate?

Modeling Exercise 7 Repeat Modeling Exercise 6 with initial conditions $y(0) = 1.5, y'(0) = 0$. Comment.

4.6.6 Project: The Pendulum 2

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 given by

$$\mathbf{r}(t) = L\langle \sin(\theta(t)), -\cos(\theta(t)) \rangle = L\mathbf{u}(t)$$

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.

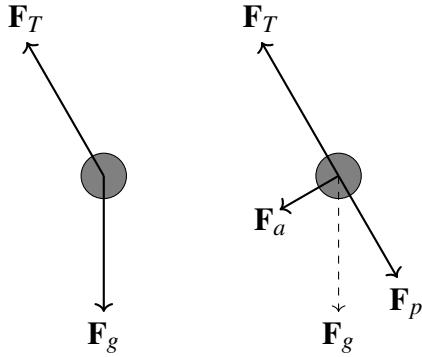


Figure 4.33: Left panel: Free body diagram of pendulum bob with gravitational force \mathbf{F}_g and rod force \mathbf{F}_T . Right panel: Same free-body diagram but with force \mathbf{F}_g decomposed into component \mathbf{F}_p orthogonal to path of motion and \mathbf{F}_a parallel to path of motion.

Modeling Exercise 1 Differentiate $\mathbf{r}(t)$ with respect to t and show that the velocity of the pendulum is given by

$$\mathbf{v}(t) = L\theta'(t)\mathbf{u}^\perp(t) \quad (4.136)$$

where $\mathbf{u}^\perp(t) = \langle \cos(\theta(t)), \sin(\theta(t)) \rangle$. Use the dot product to verify that $\mathbf{u}(t)$ and $\mathbf{u}^\perp(t)$ (and so $\mathbf{r}(t)$ and $\mathbf{v}(t)$) are orthogonal at all times.

Modeling Exercise 2 Differentiate $\mathbf{v}(t)$ in (4.136) with respect to t and show that the acceleration $\mathbf{a}(t)$ of the pendulum is given by

$$\mathbf{a}(t) = L\theta''(t)\mathbf{u}^\perp(t) - L(\theta'(t))^2\mathbf{u}(t). \quad (4.137)$$

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 it to 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,

$$\mathbf{F}_g = \mathbf{F}_p + \mathbf{F}_a \quad (4.138)$$

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 given by

$$\begin{aligned} \mathbf{F}_p &= mg \cos(\theta(t))\mathbf{u}(t), \\ \mathbf{F}_a &= -mg \sin(\theta(t))\mathbf{u}^\perp(t). \end{aligned} \quad (4.139)$$

Modeling Exercise 3 Verify that (4.138) holds if \mathbf{F}_a and \mathbf{F}_p are as given in (4.139).

Based on (4.138) and (4.139), the total force acting on the bob is

$$\mathbf{F} = -mg \sin(\theta(t))\mathbf{u}^\perp(t) + (mg \cos(\theta(t)) - T)\mathbf{u}(t) \quad (4.140)$$

where T denotes the tension in the rod, which is the magnitude of the force that the rod exerts on the bob.

Modeling Exercise 4 Use Newton's Second Law of Motion, $\mathbf{F} = m\mathbf{a}$, in conjunction with (4.140) and (4.137), to conclude that

$$\theta''(t) + \frac{g}{L} \sin(\theta(t)) = 0. \quad (4.141)$$

Analysis and Approximation of the Pendulum Equation

Equation (4.141) 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 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 (4.141) may be approximated with the simpler “linearized” second order differential equation

$$\theta''(t) + \frac{g}{L} \theta(t) = 0. \quad (4.142)$$

Hint: Look at the Taylor's series for $\sin(\theta)$ about the point $\theta = 0$.

Modeling Exercise 6 Solve (4.141) with parameters $g = 9.8, L = 1$ and initial conditions $\theta(0) = 0.1, \theta'(0) = 0$; you'll need to use a numerical ODE solver. Plot the solution for $0 \leq t \leq 5$. Repeat with equation (4.142) and compare plots. Is the approximation accurate?

Modeling Exercise 7 Repeat Modeling Exercise 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. This force is therefore of the form

$$\mathbf{F}_{fric} = -c\mathbf{v} = cL\theta'(t)\mathbf{u}^\perp(t) \quad (4.143)$$

for some constant $c > 0$, where we have used (4.136). The larger the value of c , the greater the frictional force on the pendulum.

Modeling Exercise 8 Argue that (4.141) then becomes

$$\theta''(t) + c\theta'(t) + \frac{g}{L} \sin(\theta(t)) = 0. \quad (4.144)$$

This is the equation of the *damped pendulum*.

Modeling Exercise 9 Repeat Modeling Exercises 5 to 7 with (4.144) in place of (4.142), with the choice $c = 0.5$. Compare the behavior of the solution to the nonlinear ODE (4.141) with that of the linearized ODE (4.142).

5. The Laplace Transform

5.1 Discontinuous Forcing Functions

5.1.1 Motivation: Pharmacokinetics

After major surgery patients typically have significant pain. One standard drug for providing analgesic relief for 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 blood stream 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 range of concentration in the body. Precisely controlling the amount or concentration present in the body can be challenging, and mathematical models have proven useful in this regard. This is the field of *pharmacokinetics*, which is concerned with modeling how the body metabolizes and excretes drugs.

The rate r_{out} at which many drugs are eliminated from the body is often modeled simply as [81, 69]

$$r_{out} = -ku(t) \quad (5.1)$$

where $u(t)$ denotes the amount (usually mass) of drug in the body at time t hours and $k > 0$ is some 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 just 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 [69], and this dictates the value $k \approx 0.173$ (see Exercise 2.1.12). The constant k here has units of reciprocal hours.

An ODE Model

In addition to bolus dosing, morphine can be administered gradually, via drip or an infusion pump, at some rate $r(t)$. 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

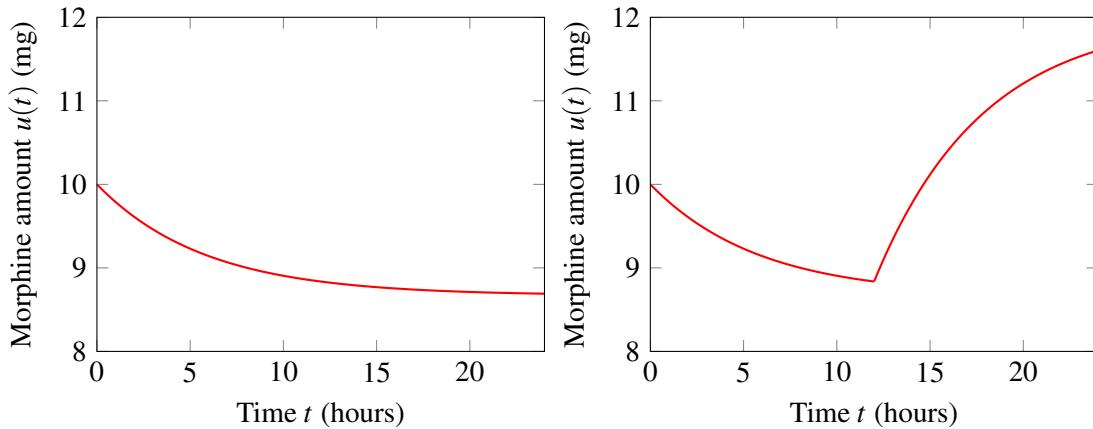


Figure 5.1: Left panel: Solution to (5.2) with $u(0) = 10$ mg, $r(t) = 1.5$ mg per hour. Right panel: Solution to (5.2) with $u(0) = 10$ mg, $r(t) = 1.5$ mg per hour for $t \leq 12$, $r(t) = 2.08$ mg per hour for $t > 12$.

1.2 for intracochlear drug delivery. By making use of (5.1) we are led to

$$\underbrace{u'(t)}_{\text{rate of change of } u(t)} = \underbrace{r(t)}_{\text{rate in}} - \underbrace{ku(t)}_{\text{rate out}}. \quad (5.2)$$

In what follows time t is measured in hours and the amount $u(t)$ of morphine in the body in milligrams (mg).

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, although another common approach is to first administer an initial bolus to immediately raise the blood concentration up to a therapeutic value and then start additional medication using an IV infusion pump. 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 107 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 via 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 $u(t) \approx 8.67 + 1.33e^{-kt}$, where recall $k \approx 0.173$. 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. ■

5.1.2 Complication: Discontinuous Forcing

But 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. However, 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 as noted above, yields $u_1(t) \approx 8.67 + 1.33e^{-kt}$, 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 initial 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 $u_2(12) = 8.84$. 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 108 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 this approach be accounted for 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 often termed to be *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 $t = 12.0005$ hours, or something like this. 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 109 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 given by $u_1(t) = 8.67 + 1.33e^{-kt}$ 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 ODE's 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 these situations.

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 ODE's so obtained. The Laplace Transform also finds uses beyond application to ODE's; it plays an important role in *control theory* in which one wishes to steer a physical system to a desired goal, e.g., 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 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 dropped back 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 $18 < t < 24$.

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 (2.15) 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 $k = 0.05$. 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 time $t = 15$ the force drops to zero. Find the position of the mass for $t > 0$. Hint: Split the problem into three pieces, according to whether $0 \leq t \leq 10$, $10 < t < 15$, or $t \geq 15$. 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, then illustrate how the Laplace transform can be used to solve ODE's. We'll momentarily put aside modeling and applications in this section to concentrate on the essential mathematics, but with the pharmacokinetic model of the last section for occasional inspiration. Our focus will be on homogeneous first and second-order linear equations. In this section you'll see how the Laplace transform turns ODE's into algebra problems. It may seem that we're merely solving ODE's 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 ODE's 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 further 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 answer 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 110 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

$$\begin{aligned} \int_0^T Ce^{-st} dt &= C \int_0^T e^{-st} dt \\ &= -\frac{e^{-st}}{s} \Big|_{t=0}^{t=T} \\ &= -C \frac{e^{-sT}}{s} + C \frac{e^{(0)s}}{s} \\ &= \frac{C}{s} - C \frac{e^{-sT}}{s}. \end{aligned} \quad (\text{Use } e^0 = 1 \text{ above.}) \quad (5.10)$$

Now use (5.10) in (5.9) to find that

$$\begin{aligned} \int_0^\infty Ce^{-st} dt &= \lim_{T \rightarrow \infty} \left(\frac{C}{s} - C \frac{e^{-sT}}{s} \right) \\ &= \lim_{T \rightarrow \infty} \frac{C}{s} - C \lim_{T \rightarrow \infty} \frac{e^{-sT}}{s} \end{aligned} \quad (5.11)$$

$$= \frac{C}{s}. \quad (5.12)$$

Above 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 (5.11) doesn't even depend on T , so that limit is just C/s . Equation (5.12) 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 $s > 0$. Note that if $s \leq 0$ then the second limit on the right in (5.11) doesn't exist, since in that case the exponential on the right in (5.11) grows without limit 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 $f(t) = e^{3t}$. The Laplace transform of this function is

$$F(s) = \int_0^\infty e^{3t} e^{-st} dt. \quad (5.13)$$

Again, this is an improper integral and we will do it with some care, 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 (5.13) is computed as

$$\int_0^\infty e^{(3-s)t} dt = \lim_{T \rightarrow \infty} \left(\int_0^T e^{(3-s)t} dt \right) \quad (5.14)$$

assuming the limit on the right above exists. A suitable antiderivative for the integrand on the right in (5.14) is $e^{(3-s)t}/(3-s)$; make sure you believe this. Then

$$\begin{aligned} \int_0^T e^{(3-s)t} dt &= \frac{e^{(3-s)t}}{3-s} \Big|_{t=0}^{t=T} \\ &= \frac{e^{(3-s)T}}{3-s} - \frac{e^{(3-s)(0)}}{3-s} \\ &= \frac{1}{s-3} + \frac{e^{(3-s)T}}{3-s}. \end{aligned} \quad (\text{Use } e^0 = 1 \text{ above.}) \quad (5.15)$$

Use (5.15) in (5.14) to find that

$$\begin{aligned} \int_0^\infty e^{(3-s)t} dt &= \lim_{T \rightarrow \infty} \left(\frac{1}{s-3} + \frac{e^{(3-s)T}}{3-s} \right) \\ &= \lim_{T \rightarrow \infty} \left(\frac{1}{s-3} \right) + \lim_{T \rightarrow \infty} \left(\frac{e^{(3-s)T}}{3-s} \right). \end{aligned} \quad (5.16)$$

Here's where we have to be a little bit careful. The first term on the right in (5.16) is an easy limit with value $1/(s-3)$. In the second term on the right in (5.16) 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 (5.14) 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 (5.16) limits to $1/(s-3)$, if $s > 3$. We conclude from (5.14) 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 easily 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 Calculus 2 integration techniques. But first, let's briefly consider what general types of functions $f(t)$ can be Laplace transformed, for not every function has a Laplace transform. 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 only converged 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 constants a and $M > 0$ the inequality

$$|f(t)| \leq M e^{at} \quad (5.17)$$

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 it outgrows any function $M e^{at}$, for any choice of M and a , so the inequality (5.17) can never hold for any a and M and all $t \geq 0$; see Reading Exercise 111.

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 111 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 contradicts (5.17).

A second technical condition on the functions $f(t)$ to be Laplace transformed is necessary. Specifically, f should be *piecewise continuous*. 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

$$L^- = \lim_{t \rightarrow t_0^-} f(t) \quad \text{and} \quad L^+ = \lim_{t \rightarrow t_0^+} f(t)$$

both exist but are not equal. If these limits are equal then f is continuous as $t = t_0$. At a jump discontinuity f must be bounded since these one-sided limits must 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 $0 \leq t \leq T$.

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 generally only makes sense when applied to continuous or piecewise continuous functions.

We summarize our above observations in the following theorem. Although the work above is not a rigorous proof of this theorem, such a proof can be found in [27].

Theorem 5.2.1 If $f(t)$ is defined for $t \geq 0$, piecewise continuous, and of exponential order with some constant a in (5.17) 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

$$\mathcal{L}(f+g)(s) = F(s) + G(s), \mathcal{L}(cf)(s) = cF(s).$$

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
$\sin(bt)$	$b/(s^2 + b^2)$	
$\cos(bt)$	$s/(s^2 + b^2)$	
$e^{at} \sin(bt)$	$b/((s-a)^2 + b^2)$	$s > a$
$e^{at} \cos(bt)$	$(s-a)/((s-a)^2 + b^2)$	$s > a$

Table 5.1: Laplace transforms of elementary functions.

Both of these facts follow from the linearity of integration and limits, for

$$\begin{aligned}\int_0^\infty e^{-st}(f(t) + g(t)) dt &= \int_0^\infty e^{-st} f(t) dt + \int_0^\infty e^{-st} g(t) dt, \\ \int_0^\infty e^{-st}(cf(t)) dt &= c \int_0^\infty e^{-st} f(t) dt,\end{aligned}$$

if the various integrals above converge. 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. Note that 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 Calculus 2.

Reading Exercise 112 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 ODE's.

■ **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

$$\begin{aligned}\mathcal{L}(f) &= 7\mathcal{L}(e^{-2t}) - 3\mathcal{L}(\cos(2t)) + \mathcal{L}(t^3) \\ &= \frac{7}{s+2} - \frac{3s}{s^2+4} + \frac{12}{s^4}\end{aligned}$$

where the s was left off of the \mathcal{L} expressions. ■

Reading Exercise 113 Use Table 5.1 to compute the Laplace transform of $f(t) = t^2 - 3\sin(3t) + 5$.

Remark 9 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 for our table of Laplace transforms.

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 (5.17). 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 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 that $\mathcal{L}(f') = -9/(s^2 + 9)$ 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 very definition of the Laplace transform of $f'(t)$ according to (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 work the integral on the right in (5.19) using what is arguably the single most important technique in applied mathematics, integration by parts:¹

$$\int_c^d u(t)v'(t) dt = u(t)v(t) \Big|_{t=c}^{t=d} - \int_c^d u'(t)v(t) dt.$$

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); 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 the truth of (5.18).

¹According to Bill Symes, Rice University.

5.2.4 Solving Differential Equations Using Laplace Transforms

Let's consider how the Laplace transform can be used to solve ODE's. We 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 ODE's. 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 inverse Laplace transform $U(s)$ to find $u(t)$. Up to this point the solution process has been fairly mechanical, but now a bit of creativity may be required. 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 Table 5.1 makes this easier, in particular, the left column of that table. 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 ODE's. 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. However, we will later see situations in which the Laplace transform is used to analyze ODE's that would be more difficult to handle using earlier methods.

Reading Exercise 114 Emulate the computation of Example 5.6 to solve the ODE $u'(t) = -2u(t)$ with $u(0) = 3$.

The General Flow for Solving ODE's Using Laplace Transforms

As in Example 5.6, the solution process for ODE's 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 we follow the somewhat circuitous route obtained by Laplace transforming (moving us to the upper right corner), then performing routine algebra (moves us to the lower right corner) and then inverse Laplace transforming to arrive at the solution in the lower left corner.

The idea behind Figure 5.2 is in fact 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

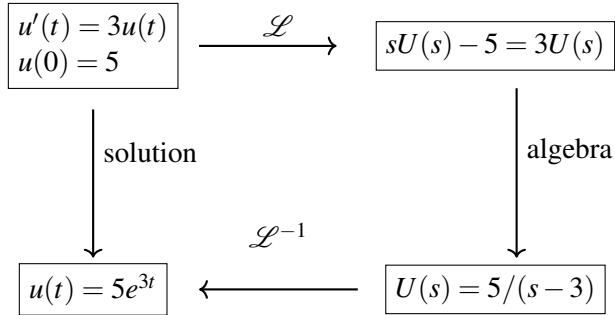


Figure 5.2: Commutative diagram for Laplace transform solution flow.

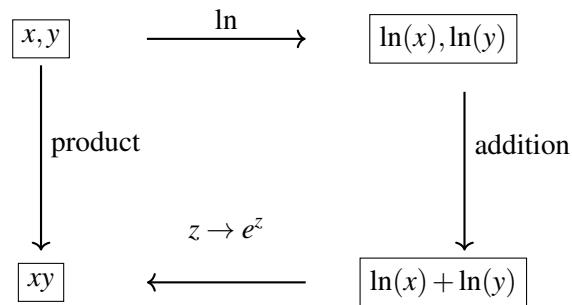


Figure 5.3: Commutative diagram for multiplication with logarithms/exponentials.

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. 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.

Reading Exercise 115 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 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 $F(s) = 1/(s+2)^2$. From (5.25) and the fact that $f(0) = 0$ and $f'(t) = -2te^{-2t} + e^{-2t}$ (so $f'(0) = 1$) this means 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 116 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 ODE's. In many cases some creative algebra is needed in the solution process, and this algebra strongly resembles that involved in certain integration techniques from Calculus 2. From Section 4.2 we know that there are three possibilities for linear, constant coefficient, homogeneous, second-order ODE's: overdamped, critically damped, and underdamped (if we lump undamped with underdamped). Let's consider a two examples, and overdamped and an underdamped system.

We begin with the overdamped example, done rather thoroughly.

■ **Example 5.8** Consider the ODE

$$u''(t) + 4u'(t) + 3u(t) = 0$$

with initial conditions $u(0) = 2, u'(0) = 4$. 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, find $U(S)$ using elementary algebra, 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 $u(0) = 2, u'(0) = 4$ yields

$$s^2U(s) - 2s - 4 + 4(sU(s) - 2) + 3U(s) = 0. \quad (5.26)$$

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 above is exactly the characteristic polynomial for this ODE. Divide both sides above by $s^2 + 4s + 3$ to find

$$U(s) = \frac{2s+12}{s^2+4s+3}. \quad (5.27)$$

We now know the Laplace transform of the solution to the ODE is $U(s)$ as given by (5.27). The last step is to inverse Laplace transform to obtain the solution. 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 Calculus 2 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 inverse Laplace transforming.

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, which 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) then

$$U(s) = \frac{5}{s + 1} - \frac{3}{s + 3}. \quad (5.30)$$

It's now easy to inverse Laplace transform to obtain $u(t)$. 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 familiarize ourselves with the process in anticipation of eventually tackling more challenging problems with discontinuous and impulsive forces. ■

Reading Exercise 117 Emulate the computations of Example 5.8 to solve the ODE $u''(t) + 3u'(t) + 2u(t) = 0$ with initial data $u(0) = 1, u'(0) = -3$.

Let's consider an underdamped system, with an emphasize on some issues that arise here.

■ **Example 5.9** Consider the ODE

$$u''(t) + 4u'(t) + 20u(t) = 0$$

with initial conditions $u(0) = 2, u'(0) = 4$. 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 above 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*, that is, it has no real roots and so does not factor.

As with the underdamped case, $U(s)$ needs a bit of algebraic preprocessing before we can employ Table 5.1. As in that case the correct course of action is dictated by Calculus 2 integration techniques; 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 have pairings

$$\begin{aligned} e^{-2t} \sin(4t) &\iff \frac{4}{(s + 2)^2 + 4^2} \\ e^{-2t} \cos(4t) &\iff \frac{s + 2}{(s + 2)^2 + 4^2}. \end{aligned} \quad (5.33)$$

In order to complete the inverse transform we must 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 (or (5.33)) 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 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 118 Use the First Shifting Theorem 5.2.2 to compute $\mathcal{L}(te^{-t})$.

5.2.6 The Inverse Laplace Transform

Formula (5.6) gives a cut-and-dried method for computing Laplace transforms. Table 5.1 facilitates this process by tabulating the transforms of commonly encountered functions, but we can always appeal to the definition (5.6) to transform something that's not in the table, if we can work the integral. The reader might wonder if there's any similar formula for computing the inverse Laplace transform. If $U(s)$ is known, is there a simple way to compute $u(t)$, perhaps by doing an integral?

But one 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 [41]. 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, something we'll explore 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 in fact 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 do find use for the approximate numerical inversion of Laplace transforms. 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 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 analyzing ODE's we appeal to technology and software like Maple, Mathematica, and Sage, much as we do for evaluating complicated integrals. In the following sections many of the computations that come up are simply not reasonable to tackle by hand. Tutorials have been posted on the book website [6] illustrating how to use these software packages to handle Laplace and inverse Laplace transforms, and apply them to ODE's.

The Time Domain and the s-Domain

Here's a bit of terminology that we'll start using, very common in science, engineering, and mathematics when using the Laplace transform. The *time domain* or *t-domain* encompasses all those functions and objects (e.g., ODE's, 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 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 *s*-domain back to the time domain.

Reading Exercise 119 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 glean some important information about a function $f(t)$ in the time domain without fully inverting its transform $F(s)$. In most cases the transform $F(s)$ will be a rational function (a ratio of two polynomials)

$$F(s) = \frac{p(s)}{q(s)}$$

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 $f(t)$.

■ **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. 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, it has no real roots. 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 120 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 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, 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 each other (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. To illustrate, 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 instead 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. From (5.37) this means that $f(t)$ must be of the form

$$f(t) = A_1 e^{-2t} + A_2 e^{(-1+4i)t} + A_3 e^{(-1-4i)t}.$$

But via Euler's identity, $e^{(-1+4i)t} = e^{-t} \cos(4t) + ie^{-t} \sin(4t)$ 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 (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 4th degree polynomial $p(s)$. The quadratic $s^2 + 4s + 8$ is irreducible and has roots $s = -2 \pm 2i$, so the roots of $q(s)$ are $s = 4$ with multiplicity 3, and $s = -2 + 2i$ and $s = -2 - 2i$, each with multiplicity 1. The partial fraction expansion for $F(s)$ is of the form

$$F(s) = \frac{A_1}{s-4} + \frac{A_2}{(s-4)^2} + \frac{A_3}{(s-4)^3} + \frac{A_4}{s-(-2+2i)} + \frac{A_5}{s-(-2-2i)}.$$

Based on Table 5.1 the function $f(t) = \mathcal{L}^{-1}(F(s))$ will thus be 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 given term might turn out to be zero. ■

Reading Exercise 121 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 what can be deduced 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

$$e^{\alpha_j t} \sin(\beta_j t), \quad e^{\alpha_j t} \cos(\beta_j t), \quad \dots, \quad t^{m_j-1}e^{\alpha_j t} \sin(\beta_j t), \quad t^{m_j-1}e^{\alpha_j t} \cos(\beta_j t).$$

Note that 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, the behavior of $F(s)$ as $s \rightarrow \infty$ and $f(t)$ 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 $f(t)/F(s)$ is correct.

The *Initial Value Theorem* for the Laplace transforms relates the behavior of $f(t)$ as $t \rightarrow 0$ and $F(s)$ 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 $\lim_{t \rightarrow 0^+} f(t)$ 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 [41].

■ **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. If $f(t)$ is piecewise continuous and of exponential order and $F(s) = \mathcal{L}(f(t))$. Suppose also that every pole $s = a + bi$ of F satisfies either $a < 0$ or, if F has a pole at $s = 0$ then s 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 [41].

■ **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

$$\lim_{t \rightarrow \infty} f(t) = \lim_{t \rightarrow \infty} (3 + t)e^{-t} = 3$$

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 the Final Value 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 122 Let $f(t) = 2 + 3e^{-t}$. Compute $F(s)$ and verify that the Initial Value Theorem 5.2.4 and Final Value 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 ODE's, 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 ODE's.

5.2.9 Exercises

Exercise 5.2.1 Use Table 5.1 to Laplace transform the following functions. There may be multiple ways to do any given problem.

- (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 do any given problem.

- (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 $F(s) = p(s)/q(s)$ for the various possibilities of $q(s)$ listed below, with $p(s)$ a polynomial of degree less than $q(s)$. 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.

- (a) $u''(t) + 3u'(t) + 2u(t) = 0$ with $u(0) = 6, u'(0) = 4$
- (b) $4u''(t) + 8u'(t) + 4u(t) = 0$ with $u(0) = 5, u'(0) = 3$
- (c) $u''(t) + 2u'(t) + 10u(t) = 0$ with $u(0) = 1, u'(0) = 2$
- (d) $2u''(t) + 22u'(t) + 36u(t) = 0$ with $u(0) = 1, u'(0) = 12$
- (e) $3u''(t) + 6u'(t) + 6u(t) = 0$ with $u(0) = 1, u'(0) = -2$
- (f) $3u''(t) + 18u'(t) + 27u(t) = 0$ with $u(0) = 1, u'(0) = -2$

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 sinh and cosh functions. Compare to the transforms of the sine and cosine functions.

Exercise 5.2.7 Compute the Laplace transform of e^{it} (where $i = \sqrt{-1}$), just treating i as a constant, e.g., take $a = i$ 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

$$H(t) = \begin{cases} 0, & t < 0 \\ 1, & t \geq 0 \end{cases}$$

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 $t = c$, 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 (slip 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 Initial Value 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 whether the hypotheses of the Final Value Theorem 5.2.4 are satisfied. If so, check 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, $x'(0) = 0$ 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 Laplace transform. ■

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 $t \geq 0$ which is three times (continuously) differentiable, and assume f, f', f'', 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

$$\mathcal{L}(f''') = s^3 F(s) - s^2 f(0) - sf'(0) - f''(0).$$

Hint: Apply 5.18) to f'' , noting that $(f'')' = f'''$.

- (b) Perform a similar computation for $f^{(4)}$ (the fourth derivative of f .)
- (c) Show that

$$\mathcal{L}(f^{(n)}) = s^n F(s) - s^{n-1} f(0) - s^{n-2} f'(0) - \cdots - s^2 f^{(n-3)}(0) - s f^{(n-2)}(0) - f^{(n-1)}(0).$$

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 that $g(0) = 0$. Solve the differential equation $g'(t) = f(t)$ using the method of 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 becomes messier and messier as they are differentiated repeatedly. 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), what do you conjecture that the expression inside the limit on the right side of (5.38) equals for $F(s) = 1/(1+s)$ and a general choice of 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 as $k \rightarrow \infty$? ■

5.3 Nonhomogeneous Problems and Discontinuous Forcing Functions

In this section we consider linear, constant coefficient ODE's 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 ODE's with constant coefficients. To illustrate, here are a couple of brief examples, one first order, 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 expansion 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

$$u(t) = -2 + 3e^{3t}.$$

Second order equations are similar. ■

■ **Example 5.17** Let us solve the ODE $u''(t) + 3u'(t) + 2u(t) = 2e^{-3t}$ with initial data $u(0) = 1, u'(0) = 2$. 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 and collecting all $U(s)$ terms on the left, 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 expansion 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

$$u(t) = 5^{-t} - 5e^{-2t} + e^{-3t}.$$

5.3.2 Discontinuous Forcing

The Laplace transform gives a unified framework for solving nonhomogeneous problems in which the forcing function in an ODE may have 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 patient's inadequate morphine dose, 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 intervals, $0 \leq t \leq 12$ and $t > 12$, and then solve on each separately, stitching 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*.²

Definition 5.3.1 The *Heaviside function* $H(t)$ (also known as the *unit step function*) is defined as

$$H(t) = \begin{cases} 0, & t < 0, \\ 1, & t \geq 0 \end{cases}$$

Notice that $H(0)$ was left undefined. Why? Because it doesn't matter, at least for any task we will undertake. The Heaviside function is supposed to model a switch being flipped from off to on. Does it really matter whether the switch was fully off, fully on, or anywhere in between at $t = 0$? Some textbooks define $H(0) = 0$, others $H(0) = 1$, yet others $H(0) = 1/2$. 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 given 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.

■ **Example 5.18** Let's use the $H(t)$ to express the function

$$\phi(t) = \begin{cases} 0, & 0 \leq t < 2 \\ 3, & 2 \leq t < 5 \\ 0, & t \geq 5 \end{cases}$$

²Named for Oliver Heaviside, 1850-1925, an extremely influential electrical engineer, mathematician, and physicist.

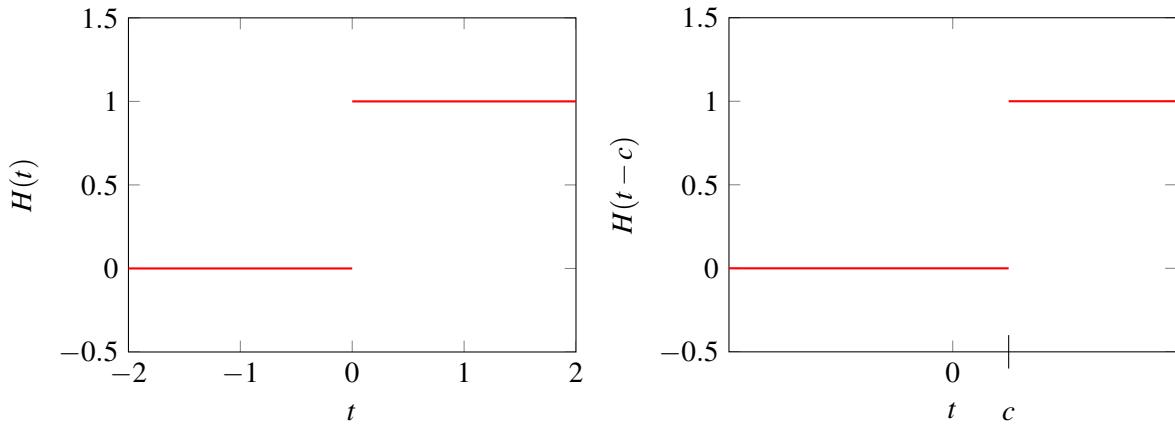


Figure 5.4: Left panel: the graph of the Heaviside function $H(t)$. Right panel: the graph of $H(t - c)$ for some $c > 0$.

The precise value of $\phi(t)$ at the discontinuities $t = 2$ and $t = 5$ is unimportant. The function $\phi(t)$ can be constructed as

$$\phi(t) = 3(H(t - 2) - H(t - 5)).$$

To see why this works, suppose $t < 2$. Then both $t - 2 < 0$ and $t - 5 < 0$, so both Heaviside functions are switched off. When $2 < t < 5$ 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. ■

■ 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

$$q(t) = \begin{cases} 0, & 0 \leq t < 2 \\ t^2, & 2 \leq t < 5 \\ e^t, & 5 \leq t < 7 \\ \cos(t), & t \geq 7 \end{cases}$$

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 $a < t < b$ and zero otherwise. ■

Reading Exercise 123 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 time.

Back to Morphine Administration

Let's formulate the ODE (5.39) with $r(t)$ as in (5.40), by using the Heaviside function. Following the flow of Example 5.19 we have

$$\begin{aligned} r(t) &= 1.5(H(t) - H(t - 12)) + 2.08H(t - 12) \\ &= 1.5 + 0.58H(t - 12). \end{aligned} \tag{5.41}$$

The last line (5.41) follows from the line above 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.

5.3.3 Laplace Transforming $H(t - c)$

We now compute the Laplace transform of $H(t - c)$ where $c \geq 0$. If you did Exercise 5.2.8 you've already done this computation. 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 124 What is the Laplace transform of $H(t - c)$ if $c < 0$? Hint: recall Remark 9 in Section 5.2.

5.3.4 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:

Theorem 5.3.1 — Second Shifting Theorem. If $f(t)$ is piecewise continuous for $t \geq 0$, of exponential order, has Laplace transform $F(s)$ defined for $s > a$, and $c \geq 0$ is any constant then

$$\mathcal{L}(H(t - c)f(t - c)) = e^{-cs}F(s),$$

defined for $s > a$.

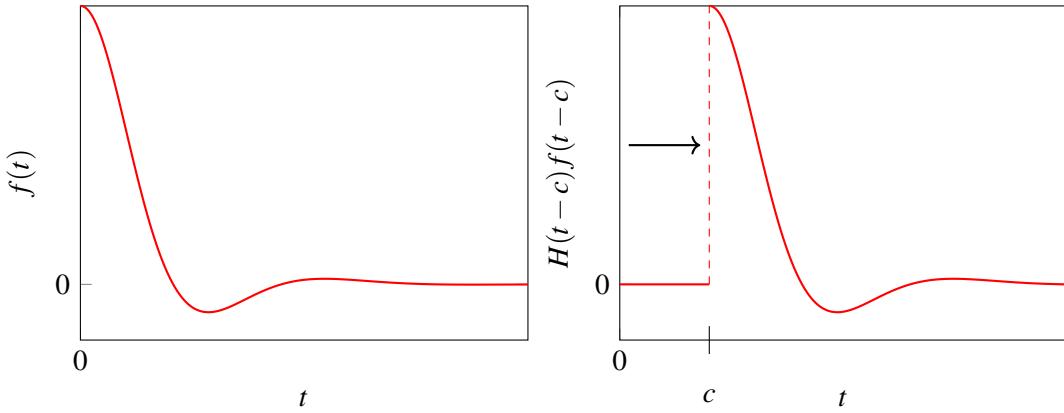


Figure 5.5: Left panel: Function $f(t)$ defined for $t \geq 0$. Right panel: $H(t - c)f(t - c)$ for some $c > 0$.

Compare the First Shifting Theorem 5.2.2 and the Second Shifting Theorem 5.3.1. There is an interesting parallel in the time and s -domains,

$$\begin{aligned} e^{ct}f(t) &\iff F(s - c) \\ f(t - c) &\iff e^{-cs}F(s). \end{aligned}$$

Multiplication by e^{ct} in the time domain corresponds to a right shift in the s -domain. Multiplication by e^{-cs} in the s -domain corresponds to a right shift by c units in the time domain. The proof of the Second Shifting Theorem is left as Exercise 5.3.11; it is essentially a Calculus 2 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

$$\phi(t) = \begin{cases} 0, & t < 2 \\ t - 2, & t \geq 2 \end{cases}$$

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.) This 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 125 Compute $\mathcal{L}(\phi)$ where

$$\phi(t) = \begin{cases} 0, & t < 3 \\ e^{t-3}, & t \geq 3 \end{cases}$$

■ **Example 5.21** Things can be a bit subtler than the last example. Here's a slight variation. Let $\phi(t)$ be defined as

$$\phi(t) = \begin{cases} 0, & t < 2 \\ t, & t \geq 2 \end{cases}$$

(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 $\phi(t) = H(t - 2)f(t - 2)$. 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 126 Compute $\mathcal{L}(\phi)$ where $\phi(t) = H(t - 7)t^2$.

Example 5.21 and Reading Exercise 126 illustrate a variation on the Second Shifting Theorem that is often useful.

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 noting 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 another slightly more involved example.

■ **Example 5.22** Let $\phi(t)$ be defined as

$$\phi(t) = \begin{cases} 3, & t < 2 \\ t, & 2 \leq t < 6 \\ 0, & t \geq 6 \end{cases}$$

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 $t \geq 0$. 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 above 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

$$\mathcal{L}(\phi) = \frac{3}{s} + \frac{e^{-2s}}{s^2} - \frac{e^{-2s}}{s} - \frac{e^{-6s}}{s} - \frac{6e^{-6s}}{s}.$$

■

Computing Inverse Transforms Using the Second Shifting Theorem

The Second Shifting Theorem is even more 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 line 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 $4e^{3t} \sin(t)$. From the Second Shifting Theorem 5.3.1, the time domain partner for $P(s)$ above is the function

$$p(t) = 4H(t - 2)e^{3(t-2)} \sin(t - 2).$$

■

Reading Exercise 127 Use the Second Shifting Theorem 5.3.2 to redo Reading Exercise 126.

Finish 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 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 find solution

$$\begin{aligned} u(t) &= 10e^{-kt} + 1.5\phi(t) + 0.58H(t-12)\phi(t-12) \\ &= 10e^{-kt} + \frac{1.5}{k} - \frac{1.5e^{-kt}}{k} + 0.58H(t-12)\left(\frac{1}{k} - \frac{e^{-k(t-12)}}{k}\right) \end{aligned}$$

This is, after simplifying separately on the intervals $0 < t < 12$ and $t > 12$, the same piecewise function defined in (5.4).

Reading Exercise 128 Solve the ODE $u'(t) = -u(t) + H(t-1)$ with initial condition $u(0) = 1$.

5.3.5 Some More Models and Examples

Let's look at a two more models from start to finish, that involve the use of the Heaviside function and are facilitated by the use of Laplace transforms.

■ **Example 5.24** Suppose an object has temperature $u(t)$ at time t with $u(0) = 50$ and is in an environment with ambient temperature $A = 80$ degrees. The object obeys Newton's Law of Cooling $u'(t) = -k(u(t) - A)$ with cooling constant $k = 0.1$. At time $t = 20$ 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

$$A = 80 - 50H(t-20)$$

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 expansion 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 $B = -30$. Then

$$\frac{50s+8}{s(s+0.1)} = \frac{80}{s} - \frac{30}{s+0.1}.$$

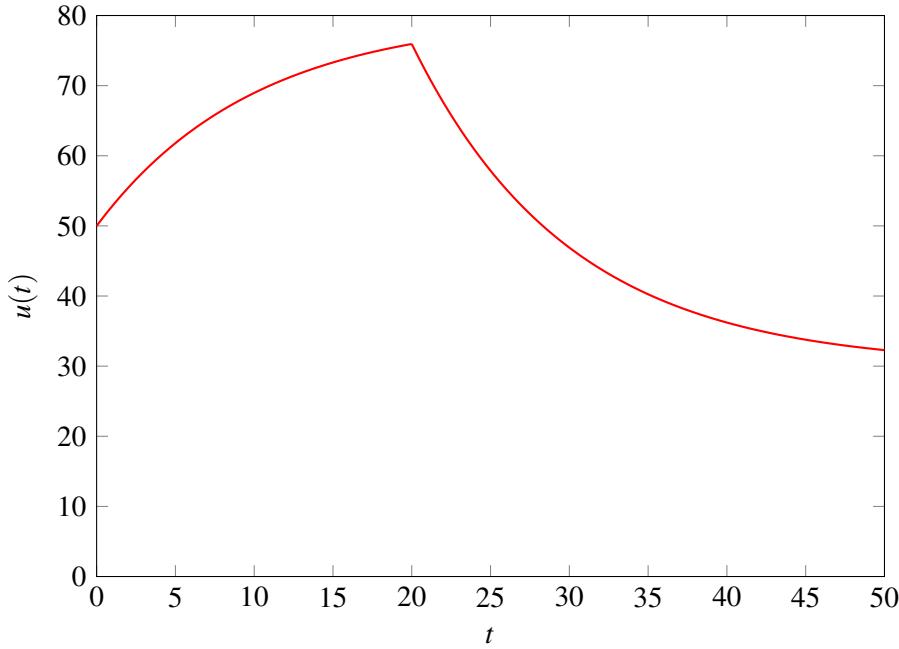


Figure 5.6: Solution to (5.48) with $u(0) = 50$.

The inverse Laplace transform of this term is $80 - 30e^{-0.1t}$. For the second term on the right in (5.49), first ignore the e^{-20s} factor and seek a partial fraction expansion 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 $B = -50$. Then

$$\frac{5}{s(s+0.1)} = \frac{50}{s} - \frac{50}{s+0.1}.$$

The inverse transform of this quantity is the function $\phi(t) = 50 - 50e^{-0.1t}$, and so by the Second Shifting Theorem 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

$$\begin{aligned} u(t) &= 80 - 30e^{-0.1t} + H(t-20)\phi(t-20) \\ &= 80 - 30e^{-0.1t} + 50H(t-20)(1 - e^{-0.1(t-20)}) \end{aligned}$$

and this provides the solution to the ODE (5.48). The function $u(t)$ is graphed in Figure 5.6. ■

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 $c = 2$ newtons per meter per second, $k = 10$ 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.

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

$$U(s) = \frac{10e^{-3s}}{s(s^2 + 2s + 10)}. \quad (5.52)$$

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 second term on the right 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 that can be read off of the bottom two lines in Table 5.1. If we let $\phi(t)$ denote the inverse transform of $\frac{1}{(s + 1)^2 + 3^2}$ 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 the Second Shifting 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 129 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 $t \rightarrow \infty$.

5.3.6 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 comes time to actually solve ODE's with discontinuous forcing functions manually the Laplace transform computations are probably just as tedious and time consuming as breaking the

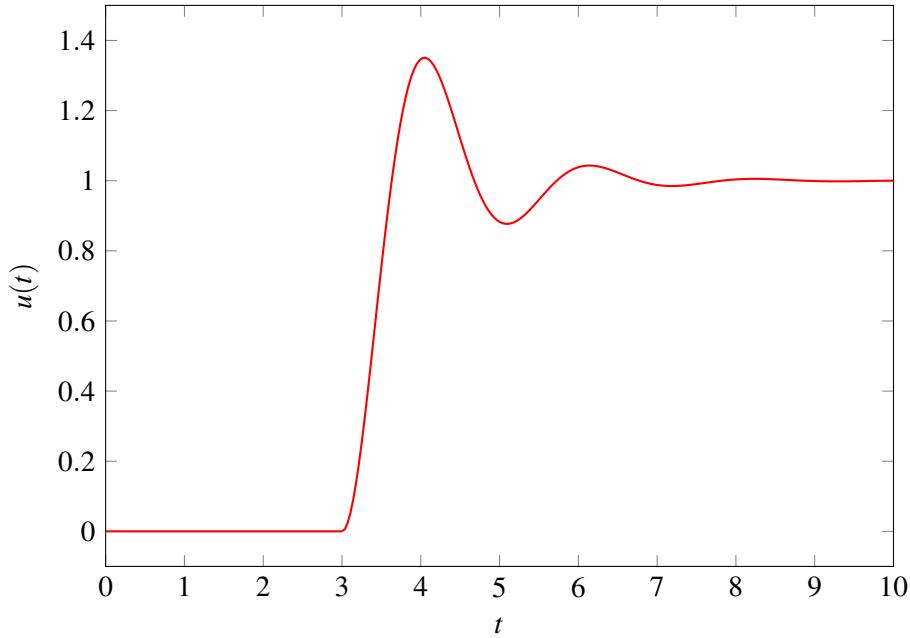


Figure 5.7: Solution to (5.51) with $u(0) = u'(0) = 0$.

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 ODE's. This 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 this says about 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.

5.3.7 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)

$$f(t) = \begin{cases} 0, & 0 \leq t < 5 \\ 7, & t \geq 5 \end{cases}$$

(b)

$$f(t) = \begin{cases} 7, & 0 \leq t < 5 \\ 0, & t \geq 5 \end{cases}$$

(c)

$$f(t) = \begin{cases} 2, & 0 \leq t < 3 \\ 5, & 3 \leq t \leq 6 \\ -3, & t > 6 \end{cases}$$

(d)

$$f(t) = \begin{cases} 1, & 0 \leq t < 3 \\ t, & 3 \leq t \leq 6 \\ e^{-t}, & t > 6 \end{cases}$$

(e)

$$f(t) = \begin{cases} 0, & 0 \leq t < 3 \\ e^t, & 3 \leq t \leq 6 \\ e^{2t}, & 6 < t \leq 10 \\ 4, & t > 10 \end{cases}$$

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 ODE's using the method of Laplace transforms and plot the solution on the interval $0 \leq t \leq 10$.

(a) $u'(t) = -2u(t) + 4H(t-5)$ with $u(0) = 1$.

(b) $u'(t) = -3u(t) + 3H(t-3) - 6H(t-5)$ with $u(0) = 1$.

(c) $u'(t) = -u(t) + tH(t-1) - H(t-2)$ with $u(0) = 2$.

Exercise 5.3.5 Solve the following second order ODE's 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 $u(0) = 1, u'(0) = 2$.

(d) $u''(t) + 16u(t) = H(t-\pi)\cos(4t)$ with $u(0) = 0, u'(0) = -1$. Hint: the result of Exercise 5.2.10 may be useful, with $F(s) = 1/(s^2 + 16)$.

The following 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 infusion 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 $t = 0$. 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 the drops to $r(t) = 200$ dollars per year for time $t \geq 2$. Formulate and appropriate ODE with 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 (2.15) 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 and solving using 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. A time $t = 15$ the force drops to zero. Find the position of the mass for $t > 0$ by formulating an appropriate ODE with 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 transforms. Plot the solution on the interval $0 \leq t \leq 0.01$.

Exercise 5.3.11 The proof of the Second Shifting Theorem 5.3.1 is a straightforward Calculus 2 computation.

- (a) From the definition of the Laplace transform

$$\mathcal{L}(H(t-c)f(t-c)) = \int_0^\infty e^{-st} H(t-c)f(t-c) dt.$$

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 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 the Second Shifting 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 ODE's 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 all those rules! But we'll do this with some care, and for a good reason: it will facilitate modeling impulsive phenomena in the same unified 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*³, which 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 130 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 phenomena 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.

■ **Example 5.26** 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.

³Popularized by physicist Paul Dirac in his work on quantum mechanics, though the ideas go back much further.

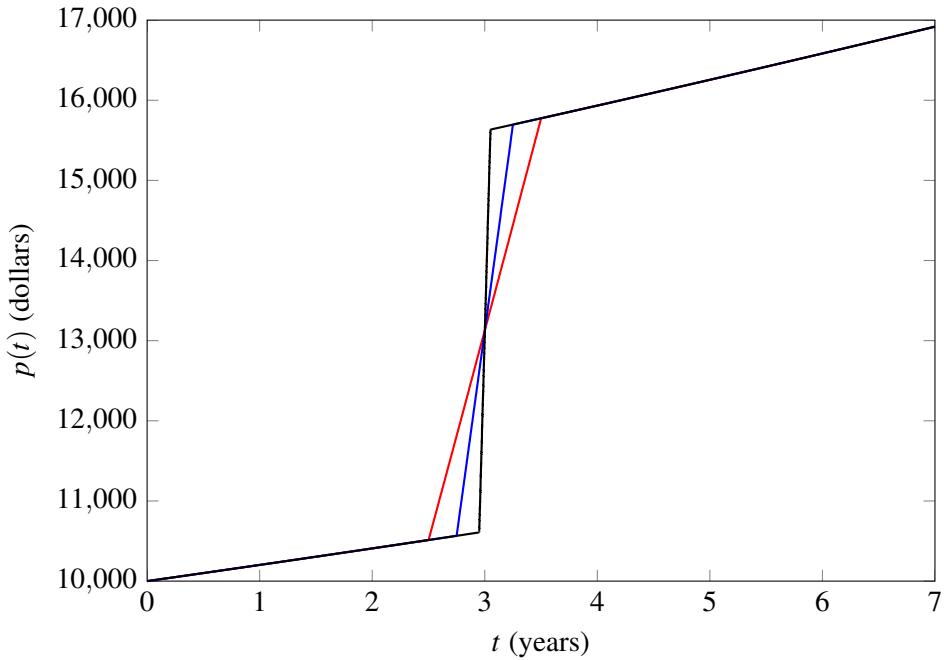


Figure 5.8: Graph of solution to (5.59) with $\varepsilon = 0.5$ (red), $\varepsilon = 0.25$ (blue), $\varepsilon = 0.05$ (black).

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 $r(t) = 5000/(2\varepsilon)$ dollars per year for 2ε years, from $t = 3 - \varepsilon$ to $t = 3 + \varepsilon$ with $\varepsilon > 0$. This corresponds to a total deposit of $\frac{5000}{2\varepsilon} \frac{\text{dollars}}{\text{year}} \times 2\varepsilon \text{ years} = \$5,000$, and the smaller we take ε , the more closely this models an instantaneous deposit. In this case the deposit rate is

$$r(t) = \frac{5000}{2\varepsilon}(H(t - 3 + \varepsilon) - H(t - 3 - \varepsilon)). \quad (5.57)$$

For future reference, the fact that $\frac{5000}{2\varepsilon} \frac{\text{dollars}}{\text{year}} \times 2\varepsilon \text{ years} = \$5,000$ can also be expressed by noting that

$$\int_0^\infty r(t) dt = 5000 \quad (5.58)$$

for any $\varepsilon > 0$, since this integral computes the area under the graph of $r(t)$, a rectangle with height $5,000/(2\varepsilon)$ and base width 2ε . With this $r(t)$ as in (5.57) the ODE 5.56) becomes

$$p'(t) = 0.02p(t) + \frac{5000}{2\varepsilon}(H(t - 3 + \varepsilon) - H(t - 3 - \varepsilon)) \quad (5.59)$$

with initial condition $p(0) = 10000$. The solution to this ODE, obtained via Laplace transforms, is plotted in Figure 5.8 for $\varepsilon = 0.5$, $\varepsilon = 0.25$, and $\varepsilon = 0.05$.

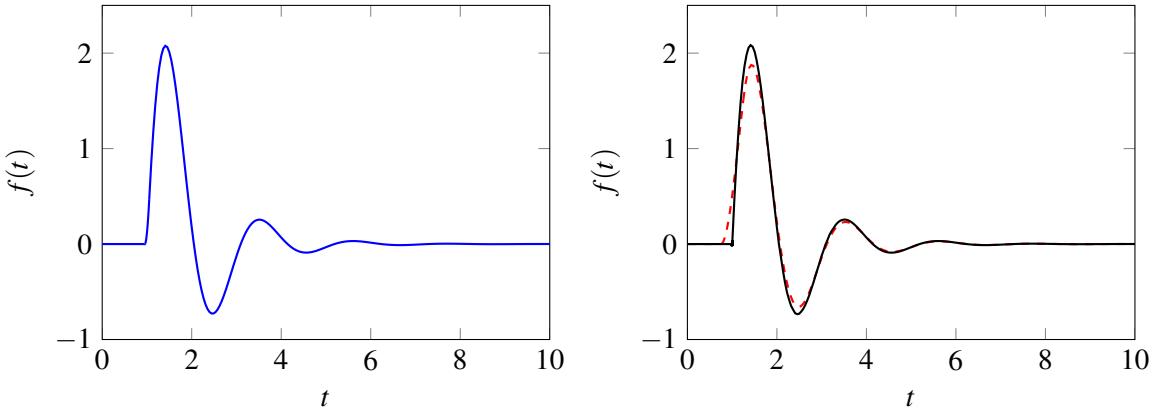


Figure 5.9: Graph of solution to (5.59) with $\varepsilon = 0.25$ (red), $\varepsilon = 0.05$ (blue), $\varepsilon = 0.005$ (black).

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. ■

Example 5.27 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

$$u''(t) + 2u'(t) + 10u(t) = 100(H(t - 0.95) - H(t - 1.05))$$

with $u(0) = u'(0) = 0$. The solution is easily obtained via Laplace transforms and is shown as the blue curve 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 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, total impulse of 10 newton-seconds), shown as the red/dashed curve in the right panel, 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 $t = t_0 - \varepsilon$ to $t = t_0 + \varepsilon$ as a constant force of $10/(2\varepsilon)$ newtons over the time interval of length 2ε seconds, where $\varepsilon > 0$. This kind of force can be written in terms of Heaviside functions as

$$f(t) = \frac{10}{2\varepsilon}(H(t - t_0 + \varepsilon) - H(t - t_0 - \varepsilon)). \quad (5.60)$$

The area under the graph of $f(t)$ is a rectangle of base width 2ε and height $\frac{10}{2\varepsilon}$, so the fact that the

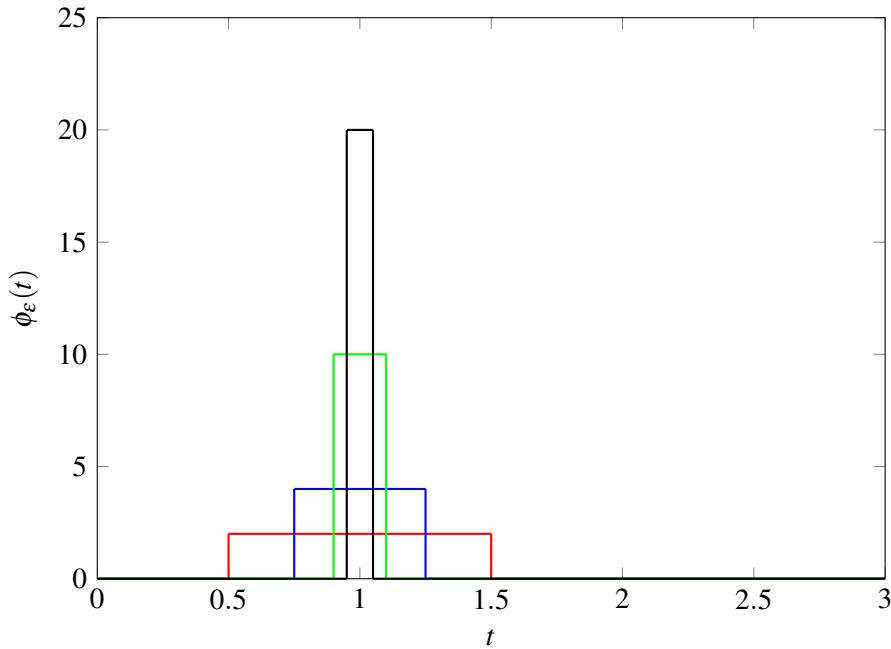


Figure 5.10: Graph of $\phi_\varepsilon(t)$ with $t_0 = 1$ and $A = 2$ for $\varepsilon = 0.5$ (red), $\varepsilon = 0.25$ (blue), $\varepsilon = 0.1$ (green), $\varepsilon = 0.05$ (black).

total impulse is 10 newton-seconds can also be expressed as

$$\int_0^\infty f(t) dt = 10 \quad (5.61)$$

for any $\varepsilon > 0$, in exactly the same way that the integral (5.58) expressed a deposit of \$5,000. A truly instantaneous hammer blow would correspond to $\varepsilon = 0$, 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 integral 10 newton-seconds! This all sounds fairly suspect, but the Dirac delta function let's 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

$$\phi_\varepsilon(t) = \frac{A}{2\varepsilon} (H(t - t_0 + \varepsilon) - H(t - t_0 - \varepsilon)) \quad (5.62)$$

for some constant A and time t_0 , where the subscript ε on ϕ_ε explicitly indicates the dependence of the right side of (5.62) on the parameter $\varepsilon > 0$. The parameter A in (5.62) 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 given by

$$\int_a^b \phi_\varepsilon(t) dt = \int_{t_0-\varepsilon}^{t_0+\varepsilon} \frac{A}{2\varepsilon} dt = (2\varepsilon) \times \left(\frac{A}{2\varepsilon} \right) = A, \quad (5.63)$$

as long as the limits of integration a and b on the left in (5.63) 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 (5.63) are common.

Reading Exercise 131 Argue that if $a < b < t_0 - \varepsilon$ or $t_0 + \varepsilon < a < b$ in (5.63) then the integral in (5.63) equals zero.

The Dirac Delta Function Definition 1

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

$$\phi_\varepsilon(t) = \frac{H(t + \varepsilon) - H(t - \varepsilon)}{2\varepsilon}.$$

The right side above is a difference quotient one would use in Calculus 1 to compute $H'(t)$ by letting $\varepsilon \rightarrow 0$. Moreover, $\varepsilon \rightarrow 0$ is exactly how we are modeling instantaneous impulses. This leads to the Dirac delta function, which might be defined as

$$\begin{aligned}\delta(t) &= \lim_{\varepsilon \rightarrow 0^+} \phi_\varepsilon(t) \\ &= \lim_{\varepsilon \rightarrow 0^+} \left(\frac{H(t + \varepsilon) - H(t - \varepsilon)}{2\varepsilon} \right) \\ &= H'(t).\end{aligned}\tag{5.64}$$

The observant reader will notice one drawback to this definition: it's absolute nonsense.

The limit in (5.64) 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 doesn't pass the sniff test of mathematical rigor, though, since for any $t \neq 0$ you can check that the limit in (5.64) is 0, while if $t = 0$ the limit does not exist. That is, $\delta(t) = 0$ for $t \neq 0$, while $\delta(0)$ has no value. The function $\delta(t)$ defined by (5.64) makes no sense, and yet the process that led to it, as in Examples 5.26 and 5.27, is clearly meaningful and of value.

How do we reconcile the need for $\delta(t)$ with its slippery definition?

A More Careful Approach

The integration in (5.63) 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 in the manner of (5.63) in the limit $\varepsilon \rightarrow 0$. Based on (5.63) with $A = 1$ and $t_0 = 0$ we find that if $a < 0$ and $b > 0$ then

$$\int_a^b \delta(t) dt = \lim_{\varepsilon \rightarrow 0^+} \int_{-\varepsilon}^{\varepsilon} \frac{1}{2\varepsilon} dt = \lim_{\varepsilon \rightarrow 0^+} 1 = 1.$$

Reading Exercise 132 What is $\int_a^b \delta(t) dt$ if $0 < a < b$ or $a < b < 0$? Hint: Consult Reading Exercise 131.

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, via integration, but in addition the Delta function will 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

$$\int_a^b A\delta(t - t_0)g(t) dt. \quad (5.65)$$

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 just $\delta(t)$, which motivates the following definition.

Definition 5.4.1 The value of the integral in (5.65) is defined to be

$$\int_a^b A\delta(t - t_0)g(t) dt = \lim_{\varepsilon \rightarrow 0} \int_a^b \frac{A}{2\varepsilon} (H(t - t_0 + \varepsilon) - H(t - t_0 - \varepsilon))g(t) dt \quad (5.66)$$

if this limit exists.

We can compute the limit of the integral on the right in (5.66) quite explicitly. It may be helpful to refer to Figure 5.11, where we assume that $a < t_0$ and $b > t_0$.

Theorem 5.4.1 If $g(t)$ is continuous and $a < t_0 < b$ then

$$\lim_{\varepsilon \rightarrow 0} \int_a^b \frac{A}{2\varepsilon} (H(t - t_0 + \varepsilon) - H(t - t_0 - \varepsilon))g(t) dt = Ag(t_0). \quad (5.67)$$

As a result of Theorem 5.4.1, we set

$$\int_a^b A\delta(t - t_0)g(t) dt = Ag(t_0) \quad (5.68)$$

whenever we encounter a Dirac delta function under an integral, provided $a < t_0 < b$ and g is continuous at $t = t_0$. Equation (5.68) 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 a couple of examples.

■ **Example 5.28** Let's use (5.68) to compute

$$\int_1^3 \delta(t - 2)e^t \sin(t) dt.$$

From (5.71) the value of this integral (take $t_0 = 2, A = 1, a = 1, b = 3$) is $e^2 \sin(2)$. ■

Reading Exercise 133 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

$$\int_0^5 \frac{1}{2\varepsilon} (H(t - 2 + \varepsilon) - H(t - 2 - \varepsilon))g(t) dt$$

(we're using $A = 1, a = 0, b = 5$) as a function of ε . Then take the limit as $\varepsilon \rightarrow 0^+$. Compare to $g(2)$.

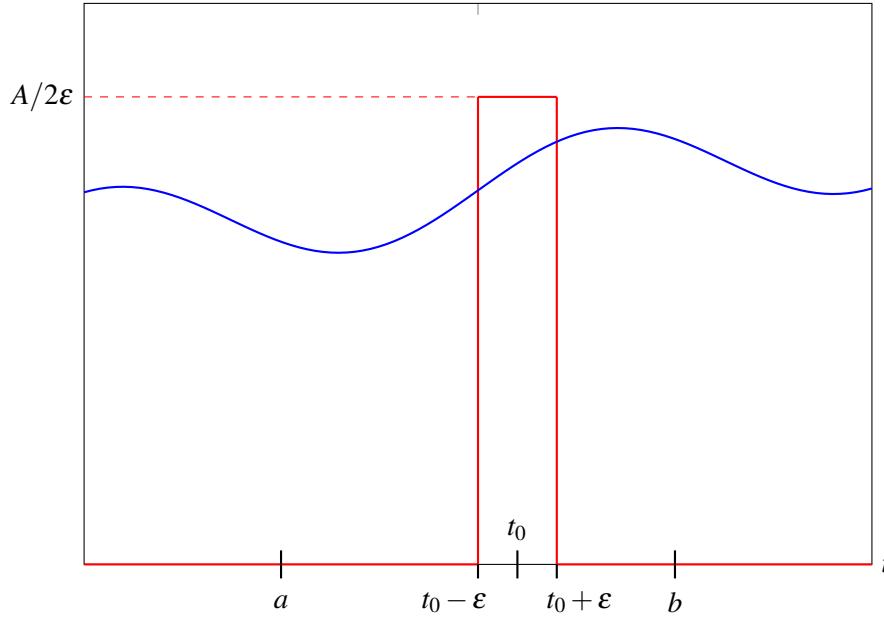


Figure 5.11: Interpreting the Dirac delta function. Graph of $\frac{A}{2\epsilon}(H(t-t_0+\epsilon)-H(t-t_0-\epsilon))$ (red) and continuous function $g(t)$ (blue).

Proof of Lemma 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 (5.67) 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 (5.66) is

$$\lim_{\epsilon \rightarrow 0^+} \int_a^b \frac{A}{2\epsilon}(H(t-t_0+\epsilon)-H(t-t_0-\epsilon))g(t) dt = \lim_{\epsilon \rightarrow 0^+} \frac{A}{2\epsilon} \int_{t_0-\epsilon}^{t_0+\epsilon} g(t) dt \quad (5.69)$$

after moving the constant $A/2\epsilon$ out in front of the integral. To work the integral on the right in (5.69), let $G(t)$ be an antiderivative for $g(t)$, so $G'(t) = g(t)$. Then by the Fundamental Theorem of Calculus

$$\int_{t_0-\epsilon}^{t_0+\epsilon} g(t) dt = G(t) \Big|_{t=t_0-\epsilon}^{t=t_0+\epsilon} = G(t_0+\epsilon) - G(t_0). \quad (5.70)$$

The use of (5.69) and (5.70) in (5.66) 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 above 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 134 Argue that if $a < b < t_0$ or $t_0 < a < b$ then

$$\int_a^b A\delta(t-t_0)g(t) dt = 0. \quad (5.71)$$

Hint: Refer to Figure 5.11 and note that the integrand is zero, at least once ε is close enough to zero.

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 $t = t_0$; such a function is sometimes called a *Dirac mass*. When we integrate a product $\delta(t - t_0)g(t)$ over an interval $a \leq t \leq b$ and the Dirac mass is strictly inside the interval, (so $a < t_0 < b$) then (5.68) holds. When the mass is strictly outside the interval ($t_0 < a$ or $t_0 > b$) then (5.71) holds.

But what if $t_0 = a$ or $t_0 = b$? This question is a bit delicate. Using the definition (5.66) 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 135). In some sense this considers half of the Dirac mass to lie in the interval of integration. But there are variations on the definition (5.66) 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

$$\begin{aligned}\int_a^b A\delta(t - a)g(t) dt &= \lim_{t_0 \rightarrow a^+} \int_a^b A\delta(t - t_0)g(t) dt \\ &= \lim_{t_0 \rightarrow a^+} Ag(t_0).\end{aligned}\tag{5.72}$$

If g is continuous at $t = a$ the right side of (5.72) is exactly $Ag(a)$.

Reading Exercise 135 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

$$\int_2^5 \frac{1}{2\varepsilon} (H(t - 2 + \varepsilon) - H(t - 2 - \varepsilon))g(t) dt$$

and take the limit as $\varepsilon \rightarrow 0^+$. Compare to $g(2)$. This would be the definition of $\int_2^5 \delta(t - 2)g(t) dt$ according to (5.66). Hint: the limits on the integral above are effectively $t = 2$ to $t = 2 + \varepsilon$.

- (b) Assume $2 < t_0 < 5$. Compute $\int_2^5 \delta(t - t_0)g(t) dt$ according to (5.68).
- (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 (5.72), which we will adopt. Compare to $g(2)$ and the answer from part (a).

5.4.3 Three Models

Let's return to three models with impulsive forcing that we've already encountered, and reinterpret them using the Dirac delta function.

■ **Example 5.29** Recall Example 5.26 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 $r(t) = 5000/(2\epsilon)$ for a very short window of time, $t = 3 - \epsilon$ to $t = 3 + \epsilon$. 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 5.30** Recall Example 5.27, 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, $f(t) = 10/(2\epsilon)$ for a very short time window $t = 1 - \epsilon$ to $t = 1 + \epsilon$. 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 5.31** 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

$$u'(t) = -ku(t) + r(t) \quad (5.73)$$

with initial condition $u(0) = 10$, where $k \approx 0.173$. 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 attempt to provide better pain control. This was equation (5.45), in which $r(t) = 1.5 + 0.58H(t - 12)$ in (5.73). The solution to the resulting ODE was shown in the right Figure 5.1. That solution indicates that the amount of morphine in 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, $r(t) = 5/(2\epsilon)$ for a short time window $t = 12 - \epsilon$ to $t = 12 + \epsilon$, or more directly, by adding $5\delta(t - 12)$ to $r(t)$. All in all the morphine amount is modeled using (5.73) 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 5.29-5.31 contain ODE's that model impulsive phenomena using Dirac delta functions on their right sides, a nice conceptual and notational simplification compared to expressions like $\frac{1}{2\epsilon}(H(t + \epsilon) - H(t - \epsilon))$. But how do we actually solve these types of ODE's? 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 given by the integral

$$\mathcal{L}(\delta(t - t_0)) = \int_0^\infty e^{-st} \delta(t - t_0) dt. \quad (5.74)$$

When $t_0 > 0$ the value assigned to this integral is easy to deduce from (5.68) with the choice $A = 1$ and $g(t) = e^{-st}$, and we obtain

$$\mathcal{L}(\delta(t - t_0)) = e^{-st_0}. \quad (5.75)$$

When $t_0 = 0$ the Laplace transform is given by

$$\mathcal{L}(\delta(t)) = \int_0^\infty e^{-st} \delta(t) dt$$

in which the Dirac mass lies at the left end of the interval of integration. This puts us precisely in the situation of equation (5.72) 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 (5.74), the third from (5.75), 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 ODE's with impulsive forcing functions.

5.4.5 Solving ODE's with Dirac Delta Functions

Let's return to Examples 5.29 - 5.31 and solve each relevant ODE.

■ **Example 5.32** In Example 5.29 the ODE of interest is

$$p'(t) = 0.02p(t) + 5000\delta(t - 3). \quad (5.76)$$

with $p(0) = 10000$. Laplace transforming both sides of (5.76) and making use of (5.75) 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 above 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 the Second Shifting Theorem 5.3.1. The inverse transform of this second term is $5000H(t - 3)e^{0.02(t-3)}$. All in all then

$$p(t) = 10000e^{0.02t} + 5000H(t - 3)e^{0.02(t-3)}.$$

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 interest on the \$5,000 deposit at $t = 3$.

■

■ **Example 5.33** In Example 5.30 the ODE of interest is

$$u''(t) + 2u'(t) + 10u(t) = 10\delta(t - 1) \quad (5.77)$$

with initial conditions $u(0) = u'(0) = 0$. Laplace transforming both sides of (5.77) and making use of (5.75) 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 the Second Shifting Theorem 5.3.1 it follows that the inverse Laplace transform of $U(s)$ is then

$$u(t) = \frac{10}{3}H(t-1)e^{-(t-1)} \sin(3(t-1)). \quad (5.78)$$

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 136 The momentum of the mass in Example 5.33 just after the hammer blow lands is $\lim_{t \rightarrow 1^+} mu'(t)$ (mass times velocity) where $m = 1$. Use (5.78) 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. Comment. See also Exercise 5.4.10.

■ **Example 5.34** In Example 5.31 the ODE of interest was

$$u'(t) = -ku(t) + 1.5 + 0.58H(t-12) + 5\delta(t-12) \quad (5.79)$$

with initial condition $u(0) = 10$. Laplace transforming both sides of (5.79) 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

$$U(s) = \frac{10}{s+k} + \frac{1.5}{s(s+k)} + \frac{0.58e^{-12s}}{s(s+k)} + \frac{5e^{-12s}}{s+k}. \quad (5.80)$$

To find $u(t)$ we inverse transform each term on the right in (5.80).

The inverse transform of $1/(s+k)$ is e^{-kt} , so that the inverse transform of the first on the right in (5.80) is $10e^{-kt}$. The inverse transform of the fourth term, $5e^{-12s}/(s+k)$ can also be computed from this result and using the Second Shifting 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 expansion

$$\frac{1}{s(s+k)} = \frac{1}{ks} - \frac{1}{k(s+k)}$$

in 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 (5.80) is $1.5\phi(t)$. The inverse transform of the third term, $\frac{0.58e^{-12s}}{s(s+k)}$, follows from the Second Shifting Theorem 5.3.1 and is $0.58H(t-12)\phi(t-12)$. All in all the solution to (5.79) 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. ■

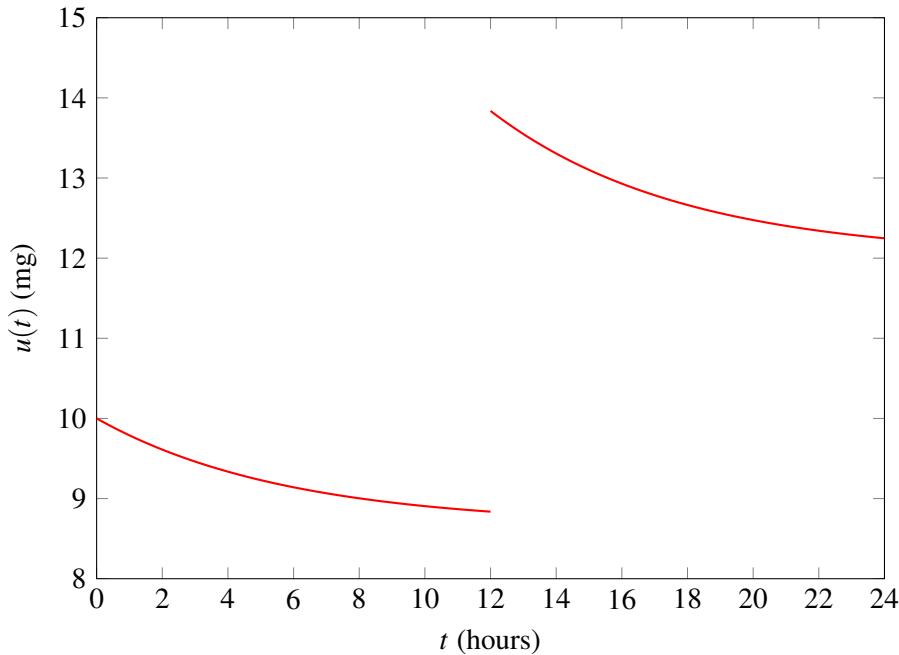


Figure 5.12: Graph of solution to (5.79) with $u(0) = 10$.

5.4.6 Summary and a Few Remarks

The Dirac delta function is a mathematical idealization of impulsive phenomena in the physical world and provides a reasonable approximation, while greatly simplifying the framework and computation for the ODE's 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 $\delta(t)$, or $H(t)$. Consider the simple ODE

$$u'(t) = -1 + 2H(t - 1)$$

with $u(0) = 1$. The solution can be obtained via 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 [109] 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.

5.4.8 Exercises

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)
$-t f(t)$	$F'(s)$	(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)
$H(t-c)g(t)$	$e^{-cs}F(s)$	where $f(t) = g(t+c)$ (Second Shifting Theorem II)
$\delta(t-c)$	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/ s -domain Laplace transform pairs.

Exercise 5.4.1 Solve the following first order ODE's 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 ODE's 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 $u(0) = 1, u'(0) = 2$.
- (d) $u''(t) + 4u(t) = \cos(2t) - 20\delta(t-3)$ with $u(0) = 1, u'(0) = 0$.

Exercise 5.4.3 Define a function $\phi(t)$ via 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 $t > 0$. How does the answer compare to $H(t)$? Why is this answer consistent with the nonsensical statement (5.64)?

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 $t = 5$, at time

a hammer blow strikes the mass with total impulse 20 newton-seconds.

- Model the situation using an appropriate ODE and initial conditions.
- 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.

- Model the situation with a first order DE, with $x(t)$ as the mass of salt in the tank at time t . Solve this ODE using the Laplace transform.
- 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 $t = 4, 8, 12$ hours. Formulate an appropriate ODE to model this situation using (5.2) with $k = 0.173$. Solve the ODE and plot the solution for $0 \leq t \leq 24$. Comment.

Exercise 5.4.7 A 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 $t = 8$, 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 $2u''(t) + 4u'(t) + 52u(t) = 0$ with initial conditions $u(0) = 1, u'(0) = -1$.

- 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 $u(0) = 1, u'(0) = -1$, 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 continually, 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 should be take for A ?

Exercise 5.4.10 Suppose a particle of mass m is at rest somewhere on the x -axis, for all times $t < t_0$ for some $t_0 > 0$. 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. 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

$$mv'(t) = A\delta(t - t_0). \quad (5.81)$$

- Solve (5.81) for $v(t)$, using initial condition $v(0) = 0$.
- Compute the momentum of the particle just after the hammer blow, by computing

$$\lim_{t \rightarrow t_0^+} mv(t).$$

Show that the result is 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 the process's *transfer function*. 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 3.5 we considered methods for estimating unknown parameters in ODE's 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 measuring the displacement of the mass periodically. In previous cases the estimation problem was solved 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 will help set up a different approach to the problem, in which we'll view things in the s -domain. Consider a driven spring-mass-damper system governed by

$$mu''(t) + cu'(t) + ku(t) = f(t) \quad (5.82)$$

with some initial conditions, which we will take to be $u(0) = u'(0) = 0$. The parameters m, c , and k

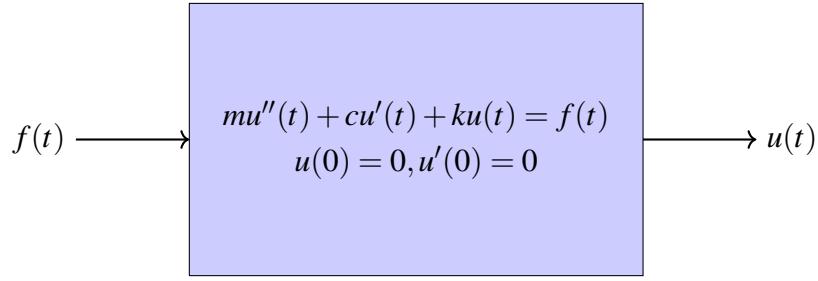


Figure 5.13: Spring-mass as an input-output system in the time domain.

are considered as unknowns, and the goal is to estimate these parameters. To do this we apply a stimulus $f(t)$ to the system, measure the response $u(t)$, and from this information determine m, c , and k .

As a specific example, suppose the forcing function is $f(t) = H(t - 1) \cos(t - 1)$ in (5.82), and suppose the mass responds according to

$$u(t) = H(t - 1) \left(\frac{\cos(t - 1)}{20} + \frac{\sin(t - 1)}{10} - \frac{e^{-(t-1)}}{8} + \frac{3e^{-3(t-1)}}{40} \right). \quad (5.83)$$

What information does this give 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 5.35.

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. As a relevant and concrete realization of such a system we'll use the driven spring-mass-damper example above, which encompasses many of the real situations we've looked at, e.g., the earthquake model, bike shock absorber, vibration isolation table, or the RLC circuit of Section 4.1. The corresponding models were all of the general form of (5.82). 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, k , and turned into the output response $u(t)$, and 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 $u(0) = u'(0) = 0$.

It would be nice if the process by which $f(t)$ is turned into $u(t)$ was simple, for example, $u(t) = f'(t)$, but this is rarely the case. To obtain $u(t)$ from $f(t)$ we have to solve the ODE (5.82), and the resulting dependence of $u(t)$ on $f(t)$ isn't very explicit. But if we use the Laplace transform to map the ODE into the s -domain, things become much easier, both conceptually and computationally. Laplace transforming both sides of the ODE (5.82) and substituting in the initial conditions yields

$$(ms^2 + cs + k)U(s) = F(s) \quad (5.84)$$

where $U = \mathcal{L}(u(t))$ and $F = \mathcal{L}(f(t))$. The s -domain version of Figure 5.13 is as shown in Figure 5.14.

The input-output process in the s -domain is governed by (5.84), which makes the manner by which $F(s)$ is turned into $U(s)$ completely transparent. We can easily see that

$$U(s) = G(s)F(s) \quad (5.85)$$

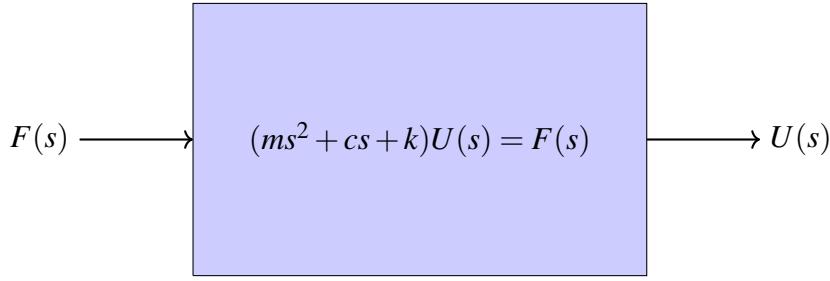


Figure 5.14: Spring-mass as an input-output system in the s -domain.

where

$$G(s) = \frac{1}{ms^2 + cs + k}. \quad (5.86)$$

The function $G(s)$ is called the *transfer function* for this spring-mass-damper system.

Reading Exercise 137 Show that $G = \mathcal{L}(g(t))$ where $g(t)$ satisfies

$$mg''(t) + cg'(t) + kg(t) = \delta(t)$$

with $g(0) = g'(0) = 0$.

We will revisit the idea of transfer functions in a more general context later, 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 5.35** 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 responded with $u(t)$ as given in (5.83); think of $u(t)$ as 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 (5.85) 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 (5.83). We also know $F(s) = \mathcal{L}(f(t)) = \frac{se^{-s}}{s^2 + 1}$. Substitute this information into (5.85) 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 above show that

$$\frac{1}{2(s+1)(s+3)} = \frac{1}{2s^2 + 8s + 6} = G(s) \quad (5.87)$$

and we conclude that the system transfer function is $G(s) = 1/(2s^2 + 8s + 6)$. But of course we also know from equation (5.86) 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 so the system parameters. ■

5.5.3 Convolution

Equation (5.85) that quantifies how an input $F(s)$ is processed into an output $U(s)$ 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 $F(s) \rightarrow U(s) = G(s)F(s)$. 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 (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

$$\begin{aligned} 237 &= 2 \cdot 10^2 + 3 \cdot 10^1 + 7 \cdot 10^0 \\ 461 &= 4 \cdot 10^2 + 6 \cdot 10^1 + 1 \cdot 10^0 \end{aligned}$$

and then multiplying every term in the expansion of 237 times 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

$$\begin{aligned} 237 \cdot 461 &= (2 \cdot 4)10^4 + (2 \cdot 6 + 4 \cdot 3)10^3 + (2 \cdot 1 + 3 \cdot 6 + 7 \cdot 1)10^2 \\ &\quad + (3 \cdot 1 + 7 \cdot 6)10^1 + (7 \cdot 1)10^0 \\ &= (8) \cdot 10^4 + (24) \cdot 10^3 + (27) \cdot 10^2 + (45) \cdot 10^1 + (7) \cdot 10^0. \end{aligned} \tag{5.88}$$

That's the heart of long multiplication, although at this point it would be traditional to perform a carry in (5.88), 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 (5.88) 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

$$237 \cdot 461 = 8|24|27|45|7$$

to express the right side of (5.88), where the symbol “|” delineates the powers of 10, with the understanding that the far right term on the right above (in this case, 7) is the 10^0 term.

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 $a = a_210^2 + a_110^1 + a_010^0$ and $b = b_210^2 + b_110^1 + b_010^0$) and find that

$$ab = (a_2b_2) \Big| (a_2b_1 + a_2b_2) \Big| (a_2b_0 + a_1b_1 + a_0b_2) \Big| (a_1b_0 + a_0b_1) \Big| (a_0b_0).$$

More generally, you can easily convince yourself that if we have integers a and b with

$$\begin{aligned} a &= a_n|a_{n-1}| \cdots |a_1|a_0, \\ b &= b_m|b_{m-1}| \cdots |b_1|b_0 \end{aligned} \tag{5.89}$$

then the product $c = ab$ can be written as

$$c = c_{m+n}|c_{m+n-1}| \cdots |c_1|c_0 \tag{5.90}$$

where

$$c_k = \sum_{j=0}^k a_j b_{k-j}, \quad (5.91)$$

though the c_k are not necessarily digits in the traditional 0 to 9 range. The integer sequence with entries c_k in (5.90) where c_k is given by (5.91) is the *convolution* of the integer sequences that defined a and b in (5.89). Convolution is usually denoted with a “*”. A common notation would be to write

$$c_{m+n}|c_{m+n-1}|\cdots|c_1|c_0 = (a_n|a_{n-1}|\cdots|a_1|a_0) * (b_m|b_{m-1}|\cdots|b_1|b_0).$$

We might even write simply $c = a * b$. This is a form of discrete convolution that's common for integer sequences.

The above reminder from 5th grade arithmetic might start to convince you that convolution is in fact an operation that arises rather naturally in many areas of mathematics, science, and engineering. The corresponding convolution operation appropriate to the study of ODE's has many parallels to this discrete version.

Reading Exercise 138 Suppose $a = 2|3|0|4$ and $b = 1|3|7$ in the notation above. Use (5.91) 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 right answer 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

$$(f * g)(t) = \int_0^t f(\tau)g(t - \tau) d\tau. \quad (5.92)$$

It's somewhat conventional to parenthesize the convolution, i.e., 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 (5.92) 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.

Reading Exercise 140 has you explore similarities between (5.91) and (5.92), but first let's look at an example.

■ **Example 5.36** Let's compute the convolution of $f(t) = t$ and $g(t) = t^2$. From (5.92) we have

$$\begin{aligned} (f * g)(t) &= \int_0^t \tau(t - \tau)^2 d\tau \\ &= \int_0^t (t^2\tau - 2t\tau^2 + \tau^3) d\tau \\ &= \left(\frac{t^2\tau^2}{2} - \frac{2t\tau^3}{3} + \frac{\tau^4}{4} \right) \Big|_{t=0}^{t=\tau} \\ &= \frac{t^4}{12} \end{aligned}$$

after simplifying. ■

Reading Exercise 139 Compute the convolution of $f(t) = e^{-t}$ and $g(t) = e^{-2t}$.

Reading Exercise 140 Compare (5.91) with (5.92) using the correspondences $t \leftrightarrow k$, $\tau \leftrightarrow j$, $f \leftrightarrow a$, $g \leftrightarrow b$, and $q \leftrightarrow c$, and with the sum in (5.91) replaced by the integral in (5.92). Convince yourself that they have precisely the same general structure.

Properties of Convolution

Convolution has a few important algebraic properties. If f_1 , f_2 , and g are piecewise continuous on $0 \leq t < \infty$ and a, b are any scalars then convolution satisfies the properties listed below.

- **Commutativity:** $f_1 * g = g * f_1$.
- **Distributivity:** $(af_1 + bf_2) * g = af_1 * g + bf_2 * g$.
- **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 by using the Laplace transform; see Exercise 5.5.12. It can also be shown that if the functions f_1 and f_2 are of exponential order then $f_1 * f_2$ is of exponential order; see Exercise 5.5.13.

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

$$Q(s) = F_1(s)F_2(s). \quad (5.93)$$

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 (5.93) might also be written as $\mathcal{L}(f_1 * f_2) = \mathcal{L}(f_1)\mathcal{L}(f_2)$. The proof of the Convolution Theorem 5.5.1 is a straightforward change-of-variable in a double integral and is given in Section 5.5.6. Let's look at some examples.

■ **Example 5.37** In Example 5.36 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 various Laplace transforms are given by $F_1(s) = 1/s^2$, $F_2(s) = 2/s^3$, and $Q(s) = 2/s^5$. It is easy to see that $Q(s) = F_1(s)F_2(s)$ in accord with the Convolution Theorem 5.5.1. ■

■ **Example 5.38** The Convolution 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 5.1 or 5.2 we have $f_1(t) = te^{-t}$ and $f_2(t) = H(t)$, and so by the Convolution Theorem $q = f * g$, or

$$\begin{aligned} q(t) &= \int_0^t f_1(\tau)f_2(t-\tau)d\tau \\ &= \int_0^t \tau e^{-\tau} H(t-\tau)d\tau \\ &= \int_0^t \tau e^{-\tau} d\tau \text{ (since } H(t-\tau) = 1 \text{ for } 0 \leq \tau \leq t) \\ &= -(\tau+1)e^{-\tau} \Big|_{\tau=0}^{\tau=t} \\ &= 1 - e^{-t}(1+t). \end{aligned}$$

Convolution with Heaviside Functions

In Example 5.38 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

$$\begin{aligned}(f * H)(t) &= \int_0^t f(\tau)H(t - \tau) d\tau \\ &= \int_0^t f(\tau) d\tau.\end{aligned}\tag{5.94}$$

Note that (5.94) 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 $\mathcal{L}(H(t)) = 1/s$ the Convolution Theorem shows that

$$\mathcal{L}(f * H) = \mathcal{L}(f(t))\mathcal{L}(H(t)) = \frac{F(s)}{s}\tag{5.95}$$

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 (5.68) (with $A = 1$, $g = f$, and variable of integration τ instead of t),

$$\begin{aligned}(\delta_{t_0} * f)(t) &= \int_0^t \delta_{t_0}(\tau)f(t - \tau) d\tau \\ &= \int_0^t \delta(\tau - t_0)f(t - \tau) d\tau.\end{aligned}\tag{5.96}$$

The integral in (5.96) can be evaluated as follows: If $0 < t_0 < t$ then the mass of the Dirac delta function is in the range of integration and from (5.68) 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 (5.96) is zero. This can be expressed as

$$(\delta_{t_0} * f)(t) = H(t - t_0)f(t - t_0).\tag{5.97}$$

In the special case that $t_0 = 0$ the convolution of $\delta(t)$ with f can be computed using (5.72) with $A = 1, a = 0$, and $b = \infty$ to find

$$(\delta * f)(t) = \int_0^\infty \delta(\tau)f(t - \tau) d\tau = f(t).\tag{5.98}$$

Equation (5.98) is consistent with the Convolution Theorem 5.5.1, for if we Laplace transform the left side of (5.98) 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 (5.98).

Reading Exercise 141 Use the Second Shifting 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 accord with the Convolution Theorem 5.5.1.

5.5.4 The Impulse Response of a System

Consider a forced harmonic oscillator governed $mu'' + cu' + ku = f$, though the system need not be spring-mass. Suppose $f(t)$ is a Dirac delta function $f(t) = \delta(t)$ and the system starts at equilibrium and at rest. The motion of the mass is given as the solution $u_\delta(t)$ to

$$mu_\delta''(t) + cu_\delta'(t) + ku_\delta(t) = \delta(t)\tag{5.99}$$

with $u_\delta(0) = u'_\delta(0) = 0$. The function u_δ is called the *unit impulse response* or just *impulse response* of the system. For a spring-mass-damper oscillator, it would be the motion of the mass in response of a hammer blow of total impulse 1 at time $t = 0$. The same process applies to first order systems: the impulse response of a system governed by $au'(t) + bu(t) = f(t)$ is the solution $u_\delta(t)$ to

$$au'_\delta(t) + bu_\delta(t) = \delta(t) \quad (5.100)$$

with $u_\delta(0) = 0$.

If we Laplace transform both sides of (5.99) and use $u_\delta(0) = u'_\delta(0) = 0$ the result is

$$(ms^2 + cs + k)\mathcal{L}(u_\delta(t)) = 1$$

which leads to

$$\mathcal{L}(u_\delta(t)) = \frac{1}{ms^2 + cs + k}.$$

That is, $\mathcal{L}(u_\delta(t))$ is precisely the transfer function $G(s)$ defined in (5.86). For a first order system, Laplace transforming (5.100), substituting in $u_\delta(0) = 0$, and then solving for $\mathcal{L}(u_\delta(t))$ yields

$$\mathcal{L}(u_\delta(t)) = \frac{1}{as + b}. \quad (5.101)$$

The function of s on the right side of (5.101) is the transfer function of the system governed by (5.100).

An analogous procedure can be used to define impulse responses and transfer functions for higher order ODE's. In fact these concepts have utility beyond the study of ODE's; they are useful in studying any type of *linear translation invariant* system. See [53].

Reading Exercise 142 Compute the transfer function for the third order system governed by $ax'''(t) + bx''(t) + cx'(t) + dx(t) = f(t)$ (that is, solve the ODE with $f(t) = \delta(t)$ and initial data $u(0) = u'(0) = u''(0) = 0$; use Laplace transforms).

5.5.5 Using Transfer Functions and Impulse Responses

The transfer function gives a conceptually simple way to analyze an ODE like $mu'' + cu' + ku = f$ with initial conditions $u(0) = u'(0) = 0$ in the time domain. This was illustrated by the input-output model in Figure 5.14, but in the s -domain. In the s -domain the Laplace transform $U(s) = \mathcal{L}(u(t))$ is given by $U(s) = G(s)F(s)$, where $G(s) = \mathcal{L}(u_\delta(t))$ is the transfer function for the system. In the time domain the relation between the solution $u(t)$ and driving function $f(t)$ is given by the convolution

$$u(t) = (u_\delta * f)(t). \quad (5.102)$$

This is easily verified with the Convolution Theorem 5.5.1 which shows that (5.102) is the time domain counterpart to equation (5.85). Similar reasoning holds for a first order system.

■ **Example 5.39** Consider the ODE $u'(t) + 3u(t) = \sin(t)$ with $u(0) = 0$. The transfer function for this system follows from (5.101) and is

$$G(s) = \frac{1}{s+3}.$$

Inverse Laplace transforming $G(s)$ yields impulse response

$$u_\delta(t) = e^{-3t}.$$

The ODE $u'(t) + 3u(t) = \sin(t)$ can be solved in the s -domain using (5.85) with $f(t) = \sin(t)$ so that $F(s) = 1/(s^2 + 1)$. This yields

$$U(s) = G(s)F(s) = \frac{1}{(s+3)(s^2+1)}.$$

This is the s -domain version of the solution process, and an inverse transform shows that $u(t)$. Alternatively we can compute $u(t)$ via (5.102), the time domain counterpart to (5.85), to find

$$\begin{aligned} u(t) &= u_\delta(t) * \sin(t) \\ &= \int_0^t e^{-3\tau} \sin(t-\tau) d\tau \\ &= \frac{e^{-3t} - \cos(t) + 3\sin(t)}{10}. \end{aligned}$$

■

Example 5.40 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 $u_\delta(t) = \mathcal{L}^{-1}(G(s))$ or

$$u_\delta(t) = \frac{1}{2}e^{-t} \sin(t).$$

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 can be found by inverse transforming $U(s)$ above, or as a convolution

$$\begin{aligned} u(t) &= u_\delta(t) * t \\ &= \int_0^t \frac{1}{2}e^{-\tau} \sin(\tau)(t-\tau) d\tau \\ &= \frac{1}{4}(t-1 + e^{-t} \cos(t)). \end{aligned}$$

■

5.5.6 System Identification with Impulsive Input

It's quite common to use an impulsive input to determine the parameters in a system.

Example 5.41 A system is governed by an ODE $au'(t) + bu(t) = f(t)$. Suppose $f(t) = 6\delta(t-2)$ and the initial data is $u(0) = 0$. The response of the system is $u(t) = 2H(t-2)e^{-(t-2)/3}$. We can use this to find a and b . First compute $F(s) = \mathcal{L}(f(t)) = 6e^{-2s}$ and $U(s) = \mathcal{L}(u(t)) = 6e^{-2s}/(3s+1)$. From $U(s) = G(s)F(s)$ we conclude that

$$\frac{6e^{-2s}}{3s+1} = G(s)6e^{-2s}. \quad (5.103)$$

Divide both sides above by $6e^{-2s}$ to conclude that $G(s) = 1/(3s+1)$. From (5.101) it is apparent that $a = 3$ and $b = 1$.

■

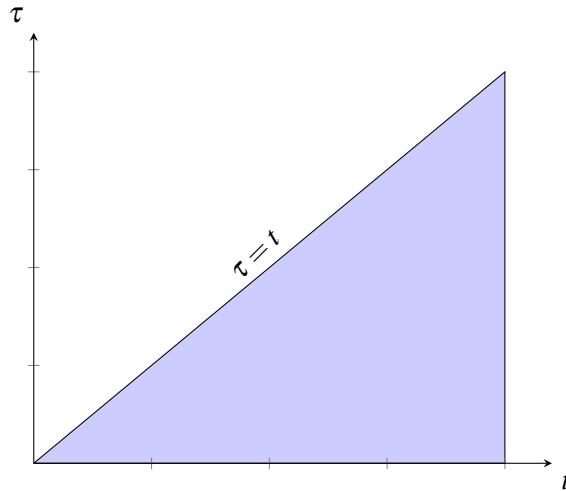


Figure 5.15: Region of integration for double integral in (5.104).

Example 5.41 and the second order Example 5.35 were noise-free and we assumed that we had a functional form for both the input $f(t)$ and output response $u(t)$. For more realistic estimation problems see Exercises 5.5.7 or 5.5.8, and the projects in Section 5.7.

Remark 10 Note that in both parameter estimation problems above we ultimately worked entirely in the s -domain, using (5.87) or (5.103). This is not uncommon. In many applications the Laplace transform is used to take the problem into the s -domain and obtain what is needed there, without ever inverse transforming back to the time domain. The next section contains further examples.

Proof of the Convolution Theorem

The proof of the Convolution Theorem 5.5.1 is 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 $p(t)$ (see Exercise 5.5.13.) The Laplace transform $P(s) = \mathcal{L}(p(t))$ is given by the integral

$$\begin{aligned} P(s) &= \int_0^\infty e^{-st} p(t) dt \\ &= \int_0^\infty e^{-st} \left(\int_0^t f_1(\tau) f_2(t - \tau) d\tau \right) dt \\ &= \int_0^\infty \int_0^t e^{-st} f_1(\tau) f_2(t - \tau) d\tau dt. \end{aligned} \tag{5.104}$$

Thus $P(s)$ is defined by the double integral on the right in (5.104), 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/τ plane for the integral in (5.104) is defined by the inequalities $0 \leq \tau \leq t$, $0 \leq t < \infty$ and is depicted as the shaded region in Figure 5.15. The region consists of all points below the line $\tau = t$. This region can also be defined by the inequalities $\tau \leq t < \infty$ and $0 \leq \tau < \infty$, and so the iterated integral (5.104) can be evaluated in the order $dt d\tau$ to

find

$$\begin{aligned} P(s) &= \int_0^\infty \int_0^t e^{-st} f_1(\tau) f_2(t-\tau) d\tau dt \\ &= \int_0^\infty \int_\tau^\infty e^{-st} f_1(\tau) f_2(t-\tau) dt d\tau. \end{aligned} \quad (5.105)$$

Let us now make a substitution $t = t - \tau$ in the inside dt integral in (5.105), so $t = w + \tau$ and $dt = dw$. The limits of integration for the dt integral are $w = 0$ to $w = \infty$ and we have

$$\begin{aligned} P(s) &= \int_0^\infty \int_0^\infty e^{-s(w+\tau)} f_1(\tau) f_2(w) dw d\tau \\ &= \int_0^\infty \int_0^\infty e^{-sw} e^{-s\tau} f_1(\tau) f_2(w) dw d\tau \\ &= \underbrace{\left(\int_0^\infty e^{-s\tau} f_1(\tau) d\tau \right)}_{F_1(s)} \underbrace{\left(\int_0^\infty e^{-sw} f_2(w) dw \right)}_{F_2(s)} \end{aligned} \quad (5.106)$$

where in the last line we use the fact that for a separable integrand $f(x)g(y)$

$$\int_a^b \int_c^d f(x)g(y) dx dy = \left(\int_c^d f(x) dx \right) \left(\int_a^b g(y) dy \right).$$

Since $p = f_1 * f_2$, (5.106) is exactly the statement that $\mathcal{L}(f_1 * f_2) = \mathcal{L}(f_1)\mathcal{L}(f_2)$, which completes the proof of the Convolution Theorem 5.5.1.

5.5.7 Exercises

Exercise 5.5.1 For each pair of functions $f_1(t)$ and $f_2(t)$ defined for $t \geq 0$ below

- 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 Laplace transform $P(s) = \mathcal{L}(p(t))$. Verify that $P(s) = F_1(s)F_2(s)$.
 - $f_1(t) = t, f_2(t) = t$
 - $f_1(t) = t, f_2(t) = 1$
 - $f_1(t) = t, f_2(t) = e^t$.
 - $f_1(t) = e^{at}, f_2(t) = e^{bt}$, a and b constants
 - $f_1(t) = \sin(t), f_2(t) = \sin(t)$
 - $f_1(t) = t^2, f_2(t) = \delta(t)$. Take note of (5.98).
 - $f_1(t) = t+3, f_2(t) = \delta(t-2)$

Exercise 5.5.2 Solve each ODE below with zero initial conditions to find the unit impulse response of the system.

- $u'(t) + 4u(t) = \delta(t)$
- $u'(t) - 2u(t) = \delta(t)$
- $u'(t) = \delta(t)$
- $u''(t) + 3u'(t) + 2u(t) = \delta(t)$
- $u''(t) + u(t) = \delta(t)$
- $2u''(t) + 4u'(t) + 10u(t) = \delta(t)$

- (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 to be found. This is an example of a *deconvolution* problem. Hint: Use the Convolution Theorem 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 $u(0) = 0$ 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, damping constant, and 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 . No choice may be perfect, due to the rounding to three significant figures. Report your estimate 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 , C , and R .

and C . Note, the measured voltage may not be perfectly consistent with any choice for L and C , due to the round-off 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 $t = 0$. At this time water with 0.1 kg per liter of salt 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 $a, b, c > 0$. 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, 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

$$U(s) = mv_0G(s)$$

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

$$u(t) = mv_0u_\delta(t)$$

where u_δ is the impulse response. ■

Exercise 5.5.12 Prove the following properties of convolution:

- **Commutativity:** $f_1 * g = g * f_1$.
- **Distributivity:** $(af_1 + bf_2) * g = af_1 * g + bf_2 * g$.
- **Associativity:** $(f_1 * f_2) * g = f_1 * (f_2 * g)$.

Hint: In each case you can just appeal to the Convolution Theorem 5.5.1. ■

Exercise 5.5.13 Suppose $f(t)$ is defined for $t \geq 0$ and of exponential order, $|f(t)| \leq M_1 e^{at}$ for some constants M_1 and a , and all $t \geq 0$. Suppose similarly $g(t)$ is defined for $t \geq 0$ and of exponential order, $|g(t)| \leq M_2 e^{bt}$ for some constants M_2 and b , and 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 $|f * g)(t)| \leq Ke^{dt}$, by following these steps.

- (a) Start with the definition of convolution (5.92) 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 Calculus 2 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 $t \geq 0$, 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 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 the table top should remain motionless. Situations like these abound in engineering and science, and more generally life. 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 similarly for 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 will be considered in the projects at the end of this section.

The appropriate control to achieve the desired end is not always obvious. This 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 143 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 in which biological materials like bacteria, plants, or even animals are grown or maintained. Incubators are used to maintain specimens at carefully controlled temperatures in ambient environments which may vary in temperature, possibly in unpredictable ways.

Consider a typical table-top 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. 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 certainly 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 in accordance with Newton's Law of Cooling,

$$y'(t) = -k(y(t) - a(t)) \quad (5.107)$$

for some constant $k > 0$, which has units of reciprocal hours. The lower the value of k , the better insulated the incubator is from the outside environment. The most desirable situation would be $k = 0$, though this is unattainable.

However, the incubator has a heating/cooling element that can be used to control the incubator 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 constant $K_1 > 0$, or removing heat if $u(t) < 0$. The constant K_1 depends on how $u(t)$ is normalized, e.g., perhaps $u(t) = 1$ means $J(t) = 1000$ 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 constant $K_2 > 0$, 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 $-k(y(t) - a(t))$ from (5.107) then we arrive at a model

$$y'(t) = -k(y(t) - a(t)) + Ku(t) \quad (5.108)$$

where $K = K_1 K_2$ is some positive constant. Equation (5.108) is a common variation of Newton's law of cooling that incorporates a heat source/sink. With initial condition $y(0) = y_0$ (5.108) has a unique solution.

For any given incubator the constants k and K could be estimated from data using techniques from Section 3.5. Thus consider 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/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 which, as we'll see later, leave something to be desired. Assume we have accurate values for the constants k and K in the model (5.108), 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. The ODE (5.108) that governs the incubator temperature becomes

$$y'(t) = -ky(t) + Ku(t) \quad (5.109)$$

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 $r(t)$, although 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 simple 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 factor into control theory, let's approach the problem in the s -domain.

Laplace transforming both sides of (5.109) 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

$$Y(s) = G_p(s)U(s) \quad (5.110)$$

where

$$G_p(s) = \frac{K}{s+k}. \quad (5.111)$$

Here $G_p(s)$ is the transfer function in the s -domain model (5.110) that governs how the input control $u(t)$ is turned into the output process variable $y(t)$, the incubator temperature.

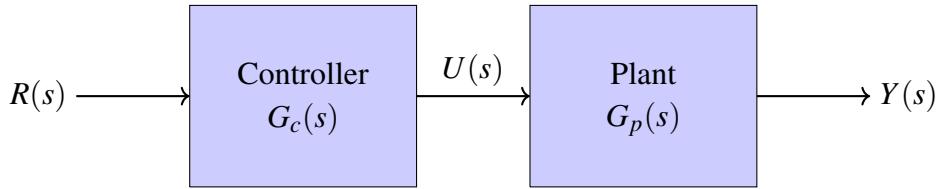
To obtain $y(t) = r(t)$ for all $t \geq 0$ we need $Y(s) = R(s)$, where $R(s) = \mathcal{L}(r(t))$. Substituting $Y(s) = R(s)$ into (5.110) shows that $U(s)$ must satisfy

$$U(s) = G_c(s)R(s) \quad (5.112)$$

where

$$G_c(s) = \frac{1}{G_p(s)} = \frac{s+k}{K}. \quad (5.113)$$

Equation (5.112) 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 (5.112) to compute $U(s)$, then inverse transform to find $u(t)$. Some examples are given below.

Figure 5.16: s -domain block diagram for open-loop control.

Control Block Diagram

This entire computation (5.110)-(5.113) can be neatly summarized with a block diagram as shown in Figure 5.16, 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 (5.110)-(5.113), is the product $G_c(s)G_p(s)$. Since $G_c(s)G_p(s) = 1$ here (this follows from (5.113) 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.

Open Loop Control Examples

■ **Example 5.42** Suppose that for a particular incubator $k = 0.05$ (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 want a setpoint temperature of $r(t) = 10(1 - e^{-t})$ degrees Celsius in the incubator, relative to the lab ambient temperature. From (5.113)

$$G_c(s) = \frac{s + 0.05}{0.5} = 2s + 0.1.$$

A Laplace transform shows that $R(s) = \frac{10}{s(s+1)}$, so that from (5.112) $U(s) = \frac{10(2s+0.1)}{s(s+1)}$. An inverse transform shows that the appropriate control function is

$$u(t) = 1 + 19e^{-t}.$$

Under these conditions this control function will asymptotically raise the temperature $y(t)$ from 0 to 10 degrees above ambient, as $y(t) = 10(1 - e^{-t})$. ■

Reading Exercise 144 Solve (5.109) with $u(t) = 1 + 19e^{-t}$ and constants as in Example 5.42, with $y(0) = 0$, and verify that $y(t) = 10(1 - e^{-t})$.

Reading Exercise 145 Solve (5.109) with $u(t) = 1 + 19e^{-t}$ and constants as in Example 5.42, but with general initial condition $y(0) = y_0$. Show that the solution approaches 10 degrees for any value of y_0 .

■ **Example 5.43** Suppose that in the setting of Example 5.42 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 (5.110) and (5.113) it follows that

$$U(s) = \frac{60\pi(2s + 0.1)}{\pi^2 + 144s^2}.$$

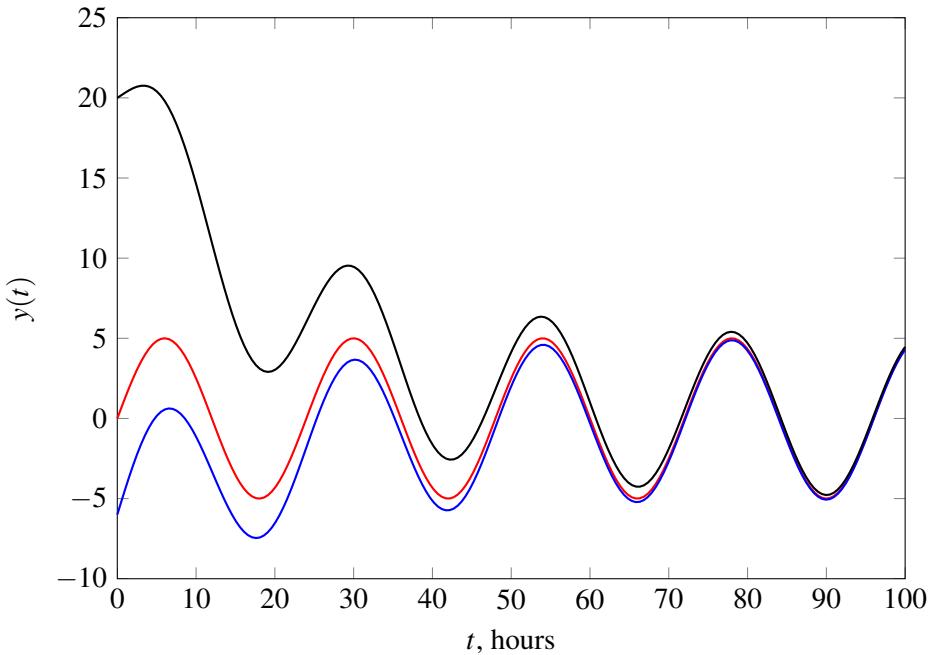


Figure 5.17: Solutions to controlled ODE (5.109) with initial conditions $y(0) = 0$ (red), $y(0) = -6$ (blue), $y(0) = 20$ (black).

An inverse transform yields 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 (5.109) 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 146 and Figure 5.17, in which we show the solution $y(t)$ with $y(0) = r(0) = 0$ in red (so $y(t) = r(t)$), the solution $y(t)$ with $y(0) = -6$ in blue, and the solution $y(t)$ with $y(0) = 20$ in black. No matter what initial condition we start with, if we wait long enough the controlled temperature profile approaches the target curve, although here it takes the better part of 100 hours. ■

Reading Exercise 146 Solve (5.109) with $u(t) = \frac{5\pi}{6} \cos(\pi t/12) + \frac{1}{2} \sin(\pi t/12)$ and constants as in Example 5.43 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, for as (5.113) 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, as Reading Exercise 146 illustrates, the system will still asymptotically approach the desired temperature profile $r(t)$. 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 graduate student leaves the window in the lab open and $a(t)$ 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/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, so K_2 will increase. 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 above cannot adapt to these disturbances, so some poor individual will be endlessly fiddling with a knob to maintain the proper temperature.

Reading Exercise 147 Consider the situation of Example 5.42 with the same constants k , K_1 , K_2 , and ambient temperature $a(t) = 0$, and desired temperature setpoint $r(t) = 10 - 10e^{-t}$. The open loop control $u(t) = 1 + 19e^{-t}$ for all t as in that example, designed with $K_2 = 0.0005$ in mind. But now the incubator has been filled with specimens and $K_2 = 0.0004$, so $K = K_1 K_2 = 0.4$. 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 far off does the incubator stray from the desired 10 degree temperature?

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 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 would 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, *proportional control*. If the incubator temperature is too cold then we should take $u(t) > 0$ to add heat energy, and 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

$$u(t) = K_p e(t) \tag{5.114}$$

in (5.109), where $e(t) = r(t) - y(t)$ is the error between the setpoint and current temperature and $K_p > 0$ is some constant, called the *gain* for the controller. 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. Note also that the control function $u(t)$ depends only on the desired temperature $r(t)$ and current temperature $y(t)$ but nothing else, which is desirable. It might continue to work well even if the parameters k , K , or $a(t)$ change or are unknown.

Reading Exercise 148 Consider the case in which $a(t) = 0$ so $y(t)$ satisfies (5.109), and suppose $u(t)$ is given by (5.114). Show that

$$y'(t) = -(k + KK_p)y(t) + KK_p r(t). \tag{5.115}$$

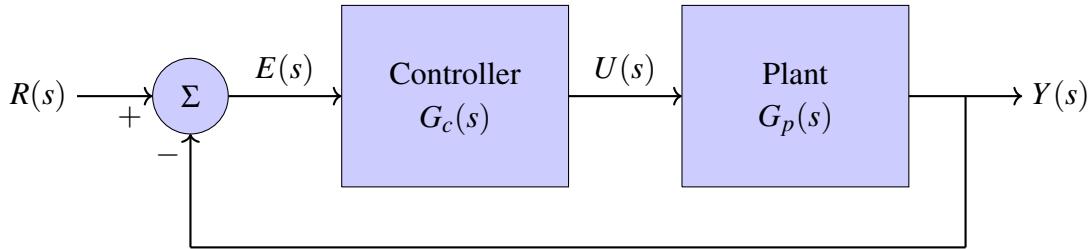


Figure 5.18: s -domain block diagram for feedback or closed-loop control.

Then take $k = 0.05$, $K = 0.5$, and $K_p = 1$ with setpoint $r(t) = 10 - 10e^{-t}$. Solve the ODE (5.115) 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 does this affect the performance of this control?

Reading Exercise 149 Repeat Reading Exercise 148 with $k = 0.05$, $K = 0.5$, and $K_p = 1$, but with $r(t) = 5 \sin(2\pi t/24)$. Plot and compare $r(t)$ and $y(t)$ for $0 \leq t \leq 200$.

Proportional Control Analysis in the s -Domain

The constant K_p in (5.114) 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 for 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 multiplications, rather than the time domain where these operations are convolutions. The reader may find it helpful to refer to Figure 5.18 for a view of the general flow of the feedback control process in the s -domain.

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 as 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 (5.111) still holds, assuming $y(0) = 0$. This is embodied by the rectangle labeled “Plant” on the right in Figure 5.18, with $U(s)$ as the input to the plant and $Y(s)$ as the output.

For the controller transfer function use (5.114) to compute

$$U(s) = G_c(s)E(s) \quad (5.116)$$

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 (5.116) relates the controller input $e(t)$ to the controller output $u(t)$ but in the s -domain. This is embodied by the left rectangle labelled “Controller” in Figure 5.18.

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 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 5.18 should make it clear why this is called feedback or closed-loop control, as the path from $Y(s)$ back to Σ closes the control loop. Contrast this to Figure 5.16.

To find the transfer function that maps $R(s)$ to $Y(s)$ note that

$$\begin{aligned} Y(s) &= G_p(s)U(s) \\ &= G_p(s)G_c(s)E(s) \\ &= G_p(s)G_c(s)(R(s) - Y(s)). \end{aligned} \quad (5.117)$$

Solving (5.117) for $Y(s)$ yields

$$Y(s) = \underbrace{\frac{G_p(s)G_c(s)}{1 + G_p(s)G_c(s)}}_{G(s)} R(s). \quad (5.118)$$

In (5.118) the function $G(s)$ is the *closed-loop transfer function* for the whole system that specifies how an input setpoint $r(t)$ is transformed into temperature output $y(t)$, but in the s -domain.

■ **Example 5.44** Consider the situation of Reading Exercise 148 in which $y(t)$ satisfies (5.109) with $k = 0.05$, $K = 0.5$, $u(t)$ given by (5.114) 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 (5.118) it follows that $Y(s) = G(s)R(s)$ where

$$G(s) = \frac{10}{20s + 11}$$

after simplifying. Since $R(s) = \mathcal{L}(r(t)) = \frac{10}{s(s+1)}$

$$Y(s) = \frac{100}{s(s+1)(20s+11)}.$$

From this it follows that in the time domain the incubator temperature is

$$y(t) = \frac{100}{11} + \frac{100}{9}e^{-t} - \frac{2000}{99}e^{-11t/20}. \quad (5.119)$$

The setpoint $r(t)$ and temperature $y(t)$ are plotted in Figure 5.19, in red and blue, respectively. For comparison we also plot the controlled temperature obtain by taking $K_p = 5$ and $K_p = 0.1$, in black and green, respectively. ■

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 (5.118) to compute that the closed-loop transfer function in general is given by

$$G(s) = \frac{KK_p}{KK_p + s + k}. \quad (5.120)$$

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. This is supported by the graphs in Figure 5.19. The main issue with large values for K_p is that since the control is given by $u(t) = K_p(r(t) - y(t))$, 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

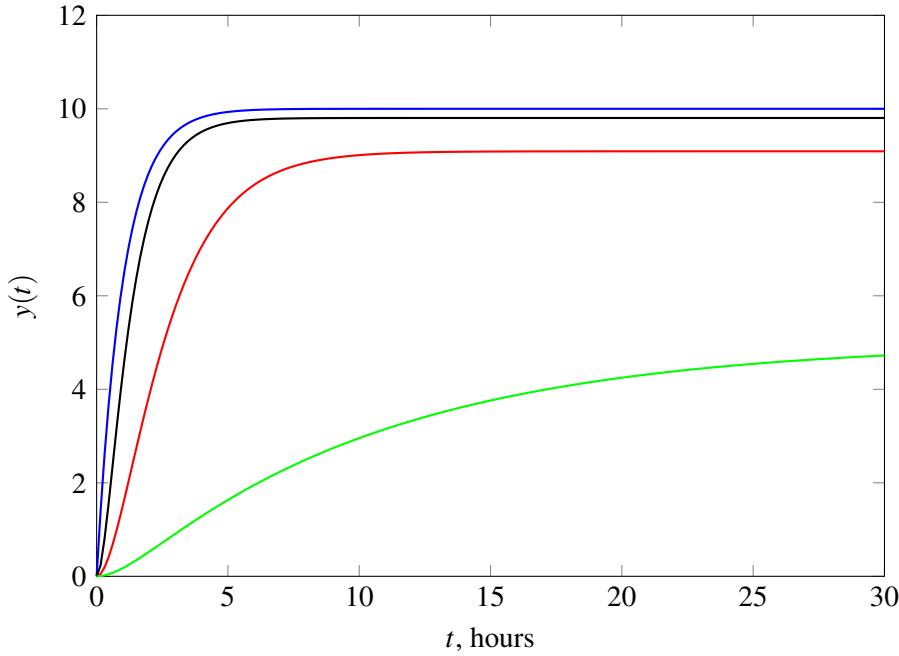


Figure 5.19: Solution to ODE (5.109) with proportional control for temperature, with $K_p = 1$ (red) and desired temperature profile (blue). Controlled temperature with $K_p = 5$ (black) and $K_p = 0.1$ (green) also shown,

a constant. Such an $r(t)$ would correspond to the case in which the incubator specimens are to be held at constant temperature. Not even this can be achieved with proportional control.

To see why suppose that

$$\lim_{t \rightarrow \infty} r(t) = r_0$$

for some constant r_0 . From the Final Value Theorem 5.2.5 (with appropriate hypotheses checked) this yields

$$\lim_{s \rightarrow 0^+} sR(s) = r_0.$$

From (5.120) and $Y(s) = G(s)R(s)$ we find

$$\begin{aligned} \lim_{s \rightarrow 0^+} sY(s) &= \lim_{s \rightarrow 0^+} sG(s)R(s) \\ &= \left(\lim_{s \rightarrow 0^+} G(s) \right) \left(\lim_{s \rightarrow 0^+} sR(s) \right) \\ &= \left(\frac{KK_p}{KK_p + k} \right) r_0 \end{aligned} \tag{5.121}$$

and then another application of the Final Value Theorem gives

$$\begin{aligned} \lim_{t \rightarrow \infty} y(t) &= \left(\frac{KK_p}{KK_p + k} \right) r_0 \\ &= r_0 - \underbrace{\left(\frac{k}{KK_p + k} \right)}_{\delta} r_0. \end{aligned} \tag{5.122}$$

Since $k > 0$ we see from (5.122) that $\delta > 0$ and so proportional control (5.114) will not give us what we want, 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. This is what was seen in Example 5.44, especially in Figure 5.19, and what you may have noticed in Reading Exercise 148.

The proportional control (5.114) can be modified by adding a carefully chosen offset δ_0 in the form $u(t) = K_p e(t) + \delta_0$ to make the bias δ in (5.122) equal to 0 so that $y(t)$ approaches the correct value when $r(t)$ is constant, but the resulting scheme ends up being a form of open loop control. In particular, δ_0 and so $u(t)$ ends up depending on k, K , and $a(t)$, which is undesirable.

Reading Exercise 150 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 with 0.1 degree of r_0 ?

Reading Exercise 151 Redo Example 5.44 but with proportional gain $K_p = -1$. What is $y(t)$? In plain English, what 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*) (5.114) is modified as

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau \quad (5.123)$$

where as before, $e(t) = r(t) - y(t)$; K_p and K_i are specified constants, called the *proportional gain* and the *integral gain*. This control strategy 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 $a(t) = 0$ and $r(t) \rightarrow r_0$ above we had an error δr_0 given by (5.122), 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 same block diagram of Figure 5.18 is applicable here. The only change is to $G_c(s)$ and precisely the same computation of (5.117) is applicable to again show that $Y(s) = G(s)R(s)$ with closed-loop transfer function $G(s)$ given as in (5.118). Let's work out $G(s)$ for this specific PI controller.

Laplace transforming (5.123) yields (recall (5.94)-(5.95) or see Exercise 5.2.18)

$$G_c(s) = K_p + \frac{K_i}{s} \quad (5.124)$$

as the transfer function for the PI controller. Use (5.124) and $G_p(s) = K/(s+k)$ in $G(s)$ from (5.118) to find that

$$G(s) = \frac{K(K_p s + K_i)}{s^2 + (KK_p + k)s + K_i} \quad (5.125)$$

is the closed-loop transfer function from $R(s)$ to $Y(s)$.

■ **Example 5.45** Let's revisit the situation of Example 5.44 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 (5.109) and examine how well $y(t)$ tracks the setpoint $r(t)$.

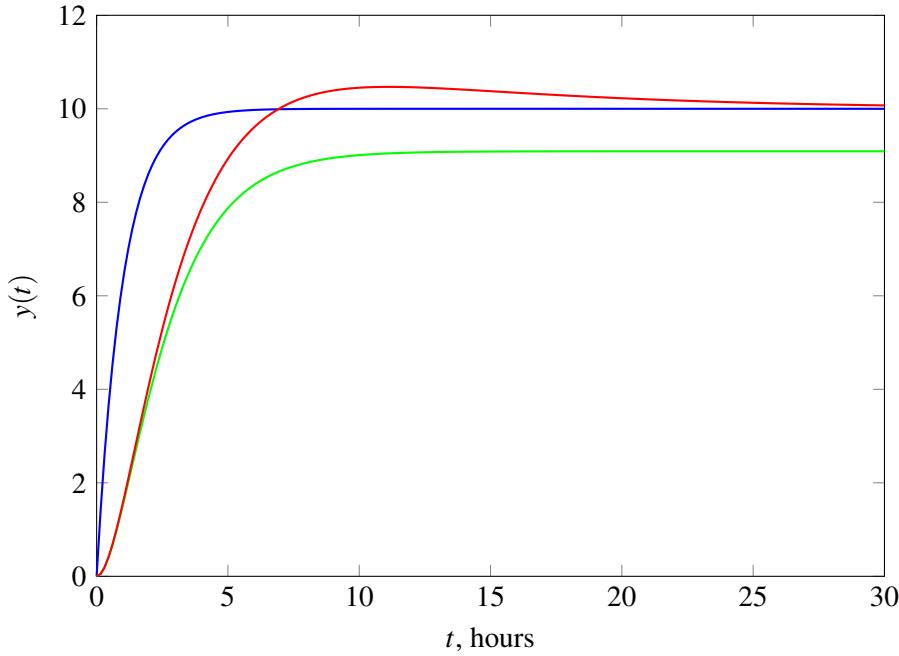


Figure 5.20: Solution to ODE (5.109) with PI control (red) and desired temperature profile (blue). Solution (5.119) with proportional control ($K_p = 1$) from Example 5.44 shown in green.

This is most easily done with Laplace transforms. We have $Y(s) = G(s)R(s)$ and with the chosen constants in (5.125)

$$G(s) = \frac{0.5s + 0.05}{s^2 + 0.55s + 0.05}$$

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

$$y(t) \approx 10 + 9e^{-t} + 2.29e^{-0.115t} - 21.29e^{-0.435t}.$$

This function is plotted in Figure 5.20 as the red curve, with the desired setpoint $r(t)$ in blue and the solution (5.119) with proportional control ($K_p = 1$) from Example 5.44 in green, 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 and/or K_i give better results? This is known as *tuning* the controller, something we'll discuss shortly. ■

Observations on PI Control

Since the numerator for G in (5.125) 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 (5.125) we have

$$\lim_{s \rightarrow 0^+} G(s) = \lim_{s \rightarrow 0^+} \left(\frac{K(K_p s + K_i)}{s^2 + (KK_p + k)s + K_i} \right) = 1 \quad (5.126)$$

if $K_i \neq 0$. A similar computation to (5.121)-(5.122) then shows that

$$\begin{aligned}\lim_{s \rightarrow 0^+} sY(s) &= \lim_{s \rightarrow 0^+} sG(s)R(s) \\ &= \left(\lim_{s \rightarrow 0^+} sR(s) \right) \left(\lim_{s \rightarrow 0^+} G(s) \right) \\ &= \left(\lim_{s \rightarrow 0^+} sR(s) \right).\end{aligned}\quad (5.127)$$

If $R(s)$ satisfies the hypotheses of the Final Value Theorem 5.2.5 we can conclude that if $\lim_{t \rightarrow \infty} r(t) = r_0$ then

$$r_0 = \lim_{s \rightarrow 0^+} sR(s) = \lim_{s \rightarrow 0^+} sY(s) = \lim_{t \rightarrow \infty} y(t)$$

This means 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 5.20.

Reading Exercise 152 Redo Example 5.45 but with integral control gain $K_i = 1$. In particular, find $y(t)$ and plot 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

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d e'(t) \quad (5.128)$$

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 (5.128) to (5.114) or (5.123). By taking any of the constants in (5.128) 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 5.18 and (5.118) still holds. For the incubator model we still have $G_p(s) = K/(s+k)$. Based on (5.128) we find that $U(s) = G_c(s)E(s)$ (again, recall (5.94)-(5.95) or Exercise 5.2.18) where

$$G_c(s) = K_p + \frac{K_i}{s} + K_d s \quad (5.129)$$

is the transfer function for the PID controller. Based on (5.118) the closed-loop transfer function is

$$G(s) = \frac{K(K_d s^2 + K_p s + K_i)}{(1 + K K_d)s^2 + (K K_p + k)s + K K_i}. \quad (5.130)$$

With PID control there are three parameters, K_p , K_i , and K_d at our disposal. For any choice of these parameters, however,

$$\lim_{s \rightarrow 0^+} G(s) = 1 \quad (5.131)$$

provided $K K_i \neq 0$, so that the same computations of (5.126)-(5.127) show that if the setpoint $r(t)$ limits to a value r_0 , the temperature will do the same.

Reading Exercise 153 Verify (5.131).

■ **Example 5.46** Let's revisit the situation of Example 5.45 (recall also Example 5.44) in which $y(t)$ satisfies (5.109) with $k = 0.05$, $K = 0.5$ and setpoint $r(t) = 10 - 10e^{-t}$. In Example 5.45 we used PI control with $K_p = 1$ and $K_i = 0.1$. Let us now try PID control with the same choice for K_p and K_i and additionally $K_d = 1$. 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 (5.109).

As before, this is most easily done with Laplace transforms. From (5.130) and the given 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 $R(s) = \mathcal{L}(r(t)) = \frac{10}{s(s+1)}$ yields

$$Y(s) = \frac{100s^2 + 100s + 10}{s(s+1)(5s+1)(6s+1)}$$

after simplification. An inverse Laplace transform shows that

$$y(t) = 10 + 28e^{-t/6} - e^{-t}/2 - 75e^{-t/5}/2.$$

This solution is plotted in Figure 5.21 as the red curve, with the desired temperature $r(t)$ in blue and the temperature obtained from PI control from Figure 5.20 in black. As in that figure the controlled solution 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, which 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 disturbances, which are unexpected changes in the system or environment, or how the controller handles desired setpoint changes. In the next example we consider the response of proportional control, PI control, and PID control for the incubator to an abrupt change in the ambient temperature or setpoint temperature.

■ **Example 5.47** 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, so the controlled ODE is

$$y'(t) = -k(y(t) - a(t)) + Ku(t) \quad (5.132)$$

with possibly nonzero initial temperature $y(t) = y_0$. The PID control function $u(t)$ will take the form (5.129) where $e(t) = r(t) - y(t)$. For simplicity let us take $K = 1$. We will make the specific choices $r(t) = 3H(t - 20)$ for the desired setpoint, $y_0 = 3$, and ambient temperature $a(t) = -5H(t - 40)$. This means that we want the incubator temperature to be 0 up to time $t = 20$ and then increase to 3, an abrupt (but here planned) change in the setpoint. Also note that the incubator starts at the wrong initial temperature. The lab ambient temperature is 0 up to time $t = 40$ at which point the ambient temperature suddenly drops 5 degrees. 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 (5.132), a task suited to Laplace transforms. Begin by computing the Laplace transform

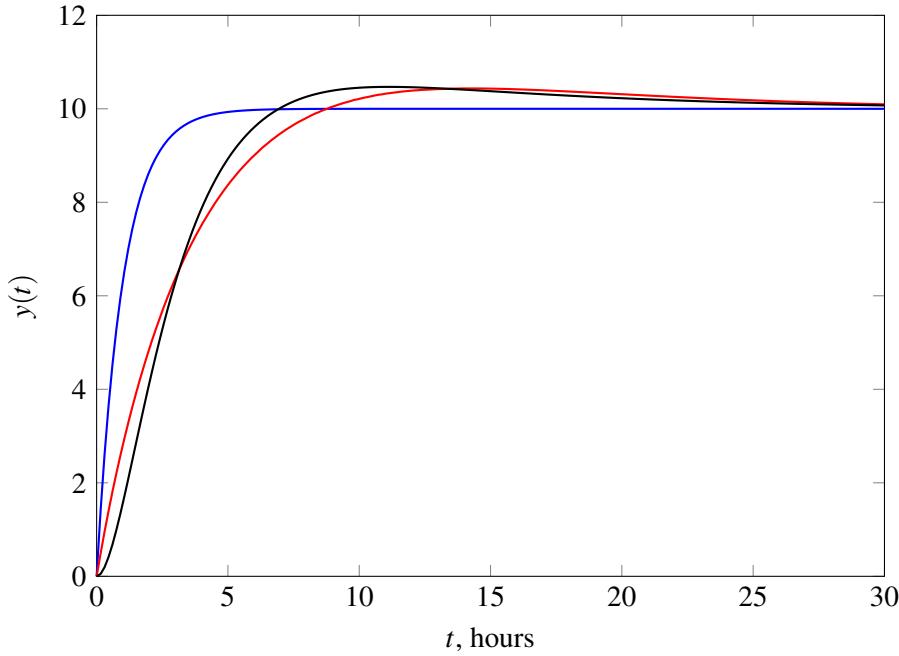


Figure 5.21: Solution to ODE (5.109) with PID control (red), PI control (black) and desired temperature profile (blue).

of both sides of (5.132) 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 $U(s) = G_c(s)(R(s) - Y(s))$ where $G_c(s)$ is given by (5.129) 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 here $K = 1$) 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

$$Y(s) = G(s)R(s) + \frac{y_0G_p(s)}{1+G_p(s)G_c(s)} + \frac{kA(s)G_p(s)}{1+G_p(s)G_c(s)} \quad (5.133)$$

where $G(s) = G_p(s)G_c(s)/(1+G_p(s)G_c(s))$ is the closed-loop transfer function as given by (5.118). The time domain response $y(t)$ can be found by inverse transforming (5.133), after choosing and substituting in the chosen control gains K_p , K_i , and K_d that appear in $G_c(s)$. 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 solution $y(t)$ for each type of control is show in Figure 5.22. ■

Stability and Tuning the Controller Parameters

Finding good choices for the parameters K_p , K_i , and/or K_d in PID control is known as *tuning* the controller and is beyond the scope of this discussion. However, one important consideration is that the controller be stable. Roughly speaking, an input-output system is *stable* if any $r(t)$ that is

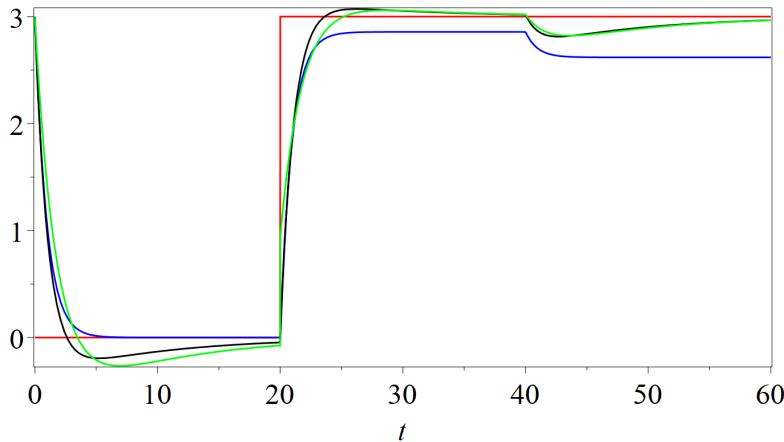


Figure 5.22: Setpoint (red). $y(t)$ for proportional control (blue), $y(t)$ for PI control (black), and $y(t)$ for PID control (green).

bounded for $t \geq 0$ yields an output $y(t)$ that is also bounded for all $t \geq 0$. Depending on the system, 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 (5.120) 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 151). 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 an exponentially growing term. 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 $e^{9t/20}$.

In general, stable control requires that the closed-loop system transfer function $G(s)$ defined in (5.118) has only poles with negative real part. This 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 [53] 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. The mathematical operations involved in PID control can be implemented electronically in digital or analog form (see [57]). Any piece of modern technology that has internal regulation, monitoring, and self-correction makes use of these techniques. This is why virtually all engineering curricula include course work 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 [70, 20]. By the appropriate application of control theory we get a plane that is both stealthy and that a human being can actually fly. See the project “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

$$u(t) = \frac{r'(t) + kr(t)}{K} \quad (5.134)$$

for an open-loop control in (5.109) with $y(0) = r(0)$ yields $y(t) = r(t)$. Then Laplace transform both sides of (5.134) and compare to (5.112) under the assumption $r(0) = 0$ (as was done there) noting that $G_c(s) = (s + k)/K$ from (5.113). Does the time domain choice for $u(t)$ in (5.134) 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

$$y'(t) = u(t). \quad (5.135)$$

We want to control $y(t)$ so that $y(t) = r(t)$ for some chosen 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 (5.135) 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 5.16. Specifically,
 - Use (5.135) 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 (5.135) from Exercise 5.6.2, with setpoint $r(t)$, and initial condition $y(0) = r(0) = 0$. Suppose we implement proportional control in (5.135), 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 (5.118) 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 and compare $y(t)$ and $r(t)$ on the range $0 \leq t \leq 5$. 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 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 (5.127). ■

Exercise 5.6.4 Consider again the ODE (5.135) from Exercise 5.6.2, with setpoint $r(t)$, and initial condition $y(0) = r(0) = 0$. Suppose we implement full PID control in (5.135), by taking

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d e'(t)$$

(this is just equation (5.128) 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 (5.118) 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 range $0 \leq t \leq 25$. Does the controlled solution stabilize? To what value? Experiment with other values for the control constants.
- (c) Suppose the controlled system in (5.135) is subject to an impulsive disturbance of total impulse 7 at time $t = 10$, so (5.135) becomes

$$y'(t) = 7\delta(t - 10) + u(t). \quad (5.136)$$

Solve (5.136) 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 (5.136) 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 (5.129). Use this and (5.118) 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 range $0 \leq t \leq 30$. Comment.
- (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 for $0 \leq t \leq 30$. Then compute the poles for $G(s)$ and explain.

■

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, due to various physiological and metabolic processes, the amount $u(t)$ (in mg) of this drug present in the patient's system diminishes according to

$$u'(t) = -ku(t) \quad (5.137)$$

for some constant $k > 0$, in the absence of any additional doses.

Modeling Exercise 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 administer the drug with the least frequency possible, e.g., every 4 hours is better than every 2 (nurses are busy people). Moreover, the interval between doses should be something 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 2 Suppose we begin with an initial bolus of 30 mg at time $t = 0$. Solve (5.137) with $u(0) = 30$ and plot the solution for $0 \leq t \leq 20$.

Modeling Exercise 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 (5.137) appropriately and solve. Plot the solution on the interval $0 \leq t \leq 20$.

Modeling Exercise 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 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 $76 \leq t \leq 144$ 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” [103], which is itself based on a modeling project in [43] (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 quantify the failure or removal of machines from service over time 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 1

- What is $F(0)$? Hint: All of the machines in operation at time t_0 are still in operation at time $t_0 + 0$.
- 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?
- What if 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?
- What if 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. To quantify this we introduce a function $R(t)$ for 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 in this case, differentiable.

Our goal is, given a target number of machines $N(t)$ and failure rate function $F(t)$, to choose an appropriate replacement function $R(t)$ that ensures there will always be $N(t)$ machines in operation at time t .

Modeling Exercise 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 limit to 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

$$R(\tau_{k+1}) - R(\tau_k) \approx R'(\tau_k)\Delta\tau_k. \quad (5.138)$$

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 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 consisting of the τ_k 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 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 .

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 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 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 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 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 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 any given machine has a 50/50 probability of failing in any interval $[t, t+a]$, but we won't pursue that at this time. As in part (d) of Modeling Exercise 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 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. You should 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 $\mathcal{L}(R)$ in Reading Exercise 6 is a bit nonstandard. We first use the geometric series identity

$$\frac{1}{1-x} = 1 + x + x^2 + \dots = \sum_{k=0}^{\infty} x^k,$$

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 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

- (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 much be replaced at time t . What is $R'(t)$ in part (a)? Hint: Recall equation (5.64). 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 actions would it advise them to do?

5.7.3 Project: Vibration 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 [5]. Vibration at any frequency is a problem, although those less than 30 Hz are most problematic (see [91]) and difficult to control. The base of the table is subjected to a vertical displacement $d(t)$ and, through the spring-mass-damper leg, the table top itself experiences motion a $y(t)$ that satisfies (4.12), 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 table top, 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 table top 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 $z(t) = y(t) - A$ or $y(t) = z(t) + A$ for some offset A , so that the position of the table top when $d(t) = 0$ is given by $z(t) = 0$, or $y(t) = A$.

Modeling Exercise 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 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 3 With A and $z(t)$ as in Modeling Exercise 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 table top to move), so that $e(t) = -z(t)$.

This implementation of control, like the others we've seen in this text, 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 $d(0) = 0$, so any disturbance has not started at this time. If we Laplace transform both sides of (5.151) and use plant transfer function $G_p(s) = 1/(ms^2 + cs + k)$ from Modeling Exercise 2 and 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 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/ground motion $d(t)$ to the table top motion $z(t)$ in the s -domain, with the effect of the PID controller included.

Modeling Exercise 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 table top 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 table top motion (displacement from equilibrium) for $0 \leq t \leq 10$.
- (b) Compare the motion of the table top with no active control; you can do this by taking $K_p = K_i = K_d = 0$ and redoing part (a). Plot and compare the effect of the PID controller. Does it help?
- (c) Redo parts (a)-(c) at frequencies 0.1 Hz and 10 Hz.

Controlling Acceleration

The feedback signal for this situation would not likely be the table top position itself, but rather the table top acceleration $z''(t)$ (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 table top to measure vertical (and other) acceleration. Moreover, for a sensitive experiment it's not the table top position $z(t)$ or velocity $z'(t)$ that is the problem (within reason), but rather the acceleration of the table top, $z''(t)$. 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 5.18. There $r(t)$ is the table top 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 table top 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 table top acceleration, and this is also what is fed back to the controller.

Modeling Exercise 6 It may be helpful to refer to Figure 5.18. 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 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 . Can you make the amplitude of the table top response smaller than you obtained above while keeping the control force $u(t)$ in a similar range?

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, although things didn't work out as well as he had hoped. See [63] for an account of the history and development of this device, and [73] 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 pendulum's equation of motion and then add friction. The focus shifts 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. This was equation (4.144), reproduced here:

$$\theta''(t) + c\theta'(t) + \frac{g}{L} \sin(\theta(t)) = 0. \quad (5.155)$$

Refer to Figure 4.31. Here $\theta(t)$ is the angle the pendulum makes with vertical, $g > 0$ is gravitational acceleration, $L > 0$ is the length of the pendulum, and $c > 0$ is a frictional coefficient.

Modeling Exercise 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 range $0 \leq t \leq 40$. Interpret—what does the plot say about the pendulum's motion?

The Inverted Pendulum

Our interest here is the inverted pendulum, when $\theta \approx \pi$, and in particular, how to use active feedback control to balance an inverted pendulum. Let's make a change-of-variable in (5.155) by setting $\alpha(t) = \pi - \theta(t)$. 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 $\alpha(0) = 0$, $\alpha'(0) = 0$ is $\alpha(t) = 0$, which is a perfectly balanced pendulum. There it will stay, so long as nothing disturbs it.

Modeling Exercise 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 $\alpha(0) = 0.01$, $\alpha'(0) = 0$, so this pendulum starts almost upright. Plot the solution on the range $0 \leq t \leq 40$. 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 was done for the standard pendulum in Section 4.6) by using the approximation $\sin(\alpha) \approx \alpha$ for $\alpha \approx 0$; this is appropriate, since 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, but disturbances may be introduced.

Modeling Exercise 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 $\alpha(0) = 0.01$, $\alpha'(0) = 0$. Plot both solutions on the range $0 \leq t \leq 5$ and compare. Is the linearized ODE (5.158) a reasonable approximation to the full nonlinear ODE (5.157)? How large is α before the solutions differ significantly?

Control

Modeling Exercise 4 Show that the transfer function that takes $u(t)$ to $\alpha(t)$ in (5.158) in the s -domain is given by

$$G_p(s) = \frac{1}{s^2 + cs - g/L}.$$

Modeling Exercise 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 $U(s) = G_c(s)E(s)$ is exactly as in (5.129). Use $G_p(s)$ and $G_c(s)$ to write out the closed-loop transfer function $G(s)$ as in (5.118) explicitly in terms of K_p, K_i, K_d and c, g , and L .

Modeling Exercise 6 Suppose the controlled system in (5.158) is subject to an impulsive disturbance of total impulse 0.1 at time $t = 3$, 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 $\alpha(0) = \alpha'(0) = 0$, 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,

$$\frac{A(s)}{G_p(s)} = 0.1e^{-3s} + G_c(s)(0 - A(s)).$$

Then solve for $A(s)$, and inverse transform. Experiment with other choices for K_p, K_i , and K_d .

Modeling Exercise 7 If you experiment with other choices for K_p, K_i , and K_d in Modeling Exercise 6 you may find the controller is unstable. Try the choices $K_p = 5, K_i = 1, K_d = 1$, solve $\alpha(t)$ as in Modeling Exercise 5.159, and plot for $0 \leq t \leq 25$. What does this control do to the pendulum? Then compute the poles of $G(s)$ and explain what these poles have to do with the observed behavior.

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, solving with the Laplace transform is of no use, since 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 $e(t) = r(t) - \alpha(t) = -\alpha(t)$ (if $r(t) = 0$) and $f(t)$ is some type of disturbance, e.g., $f(t) = 0.1\delta(t - 3)$ as in Modeling Exercise 6. Equation (5.160) is a nonlinear *integrodifferential equation*, due to the presence of the integral on the right side.

Modeling Exercise 8 But 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

$$\alpha''(t) + c\alpha'(t) - g \sin(\alpha(t))/L = f(t) - K_p \alpha(t) - K_d \alpha'(t). \quad (5.161)$$

This is just a second order nonlinear ODE.

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 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?

6. Linear Systems of Differential Equations

6.1 Systems of Differential Equations

In this section we begin 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, devoted to the study of linear systems of differential equations. This material is based upon the SIMIODE project [102], and is expanded upon in the Project Section 6.5 at the end of this chapter.

6.1.1 Motivation: More Pharmacokinetics

The drug Lysergic acid diethylamide (known as “LSD” or “acid” in popular slang) gained attention in the 1950s and 1960s, by scientists [93], government agencies [1], and the general public [36], for its various psychoactive and hallucinogenic properties. A number of studies were conducted concerning LSD’s effect on humans. In the research study [17] 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 [17] 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 subject’s 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 *couldn’t do a single addition problem correctly*.

A Two-Compartment Model

Earlier in this text we considered a variety of one-compartment models, for example, in Section 1.2 for the intracochlear drug delivery model, in Section 5.1 for morphine metabolism, and more abstractly in Section 2.1.3 where we considered salt tank problems. In each case conservation modeling was used, by 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.

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.	3.1	0.8
	Perform Score (%)	73	60	35	50	48	73	97
Subject 2	Plasma Conc (ng/ml)	10.6	7.6	7.	4.8	2.8	2.5	2.
	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 [17, 74] on 5 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.

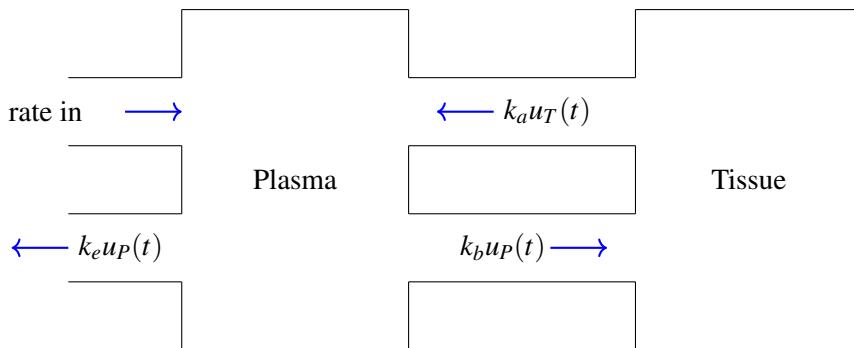


Figure 6.1: A two-compartment model with compartments corresponding to plasma and tissue.

The same approach can be used to model systems that consist of multiple compartments.

In the papers [74, 93] the authors offer a two-compartment model for the behavior of LSD in the body. The situation is illustrated schematically in Figure 6.1. In this model the authors divide the relevant areas 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, e.g., 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 the rate at which it 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. 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.

Let $u_P(t)$ denote the amount (mass) of the drug in the plasma at time t and $u_T(t)$ the amount

(mass) of the drug in the tissue at time t . In [74, 93] 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, and is assumed to be in proportion to u_T . The term $k_b u_P(t)$ in (6.2) is similar, but quantifies the flow of the drug from the tissue back to the plasma 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 introduced into the body via the plasma and $-k_e u_P(t)$ quantifies the rate at which the drug is removed from the plasma (or bloodstream) by the body. There are a number of physiological assumptions and some more detailed modeling that underly (6.1)-(6.2) that we will explore in Section 6.5.

The more immediate concern is this: equations (6.1)-(6.2) are a pair of coupled ODEs in which there are two unknowns functions, $u_P(t)$ and $u_T(t)$. The constants k_a , k_b , and k_e are considered known for now (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 tissue. 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 dependence of f_1 on t is 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 , $1 \leq j \leq n$; equivalently, we may write $\mathbf{x}(0) = \mathbf{q}$.

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 $b_j(t)$.

Reading Exercise 154 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 11 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 a parameterized curve $x_1 = x_1(t), \dots, x_n = x_n(t)$ 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 Systems

Second and higher scalar ODEs can usually be converted into equivalent system's 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 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, since this is merely $u'(t) = u'(t)$. 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 $x_2(0) = v_0$.

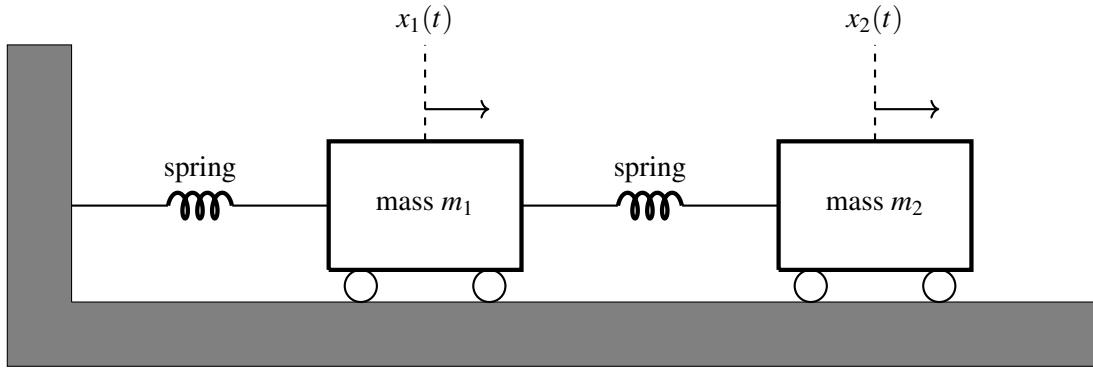


Figure 6.2: Double spring-mass system.

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 using (6.6) to substitute out \dot{x}_2 (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 155 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 will be nonlinear), by letting $x_1 = \theta$ and $x_2 = \dot{\theta}$.

■ **Example 6.2** Systems of second order (or higher) 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.

To model this system, let us assume that the masses have negligible widths (despite Figure 6.2) and that the rest length of the first spring (let us call it “spring 1”) connecting mass m_1 to the left wall is L_1 . Similarly the rest length of the spring connecting m_1 to m_2 (let us call it “spring 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 . Assume the springs each obey Hooke’s Law with springs 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, e.g., friction from the wheels in contact with the ground.

Reading Exercise 156

- Justify that the amount spring 1 is stretched/compressed is given by $x_1 - L_1$, and the amount spring 2 is stretched/compressed is given by $x_2 - x_1 - L_2$.
- 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 constant $c_1 \geq 0$. 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 constant $c_2 \geq 0$. Again, a free body diagram may help.

Reading Exercise 157 Based on Reading Exercise 156 and Newton's Second Law of Motion, show that $x_1(t)$ and $x_2(t)$ satisfy the coupled second order system

$$\begin{aligned} m_1\ddot{x}_1 &= -k_1(x_1 - L_1) + k_2(x_2 - x_1 - L_1) - c_1\dot{x}_1 \\ m_2\ddot{x}_2 &= -k_2(x_2 - x_1 - L_1) - c_2\dot{x}_2. \end{aligned} \quad (6.7)$$

The system (6.7) is a nonhomogeneous second order system. We can make a convenient change of variable, however, 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 natural length. With this change of dependent variables 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_1 , and the second ODE of the system (6.7) becomes

$$\ddot{u}_2 = \frac{k_2}{m_2}u_1 - \frac{k_2}{m_2}u_2 - \frac{c_2}{m_2}\dot{u}_2 \quad (6.9)$$

after dividing through by m_2 . 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 is removed from the equation. If we solve for u_1 and u_2 we can 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 equation

$$\dot{w}_1 = w_2 \quad (6.10)$$

is immediate. From $\ddot{u}_1 = \dot{w}_2$ and (6.8) we obtain

$$\dot{w}_2 = -\frac{k_1 + k_2}{m_1}w_1 - \frac{c_1}{m_1}w_2 + \frac{k_2}{m_1}w_3 \quad (6.11)$$

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

$$\dot{w}_3 = w_4 \quad (6.12)$$

and since $\ddot{u}_2 = \dot{w}_4$ we can use (6.9) to write

$$\dot{w}_4 = \frac{k_2}{m_2}w_1 - \frac{k_2}{m_2}w_3 - \frac{c_2}{m_2}w_4. \quad (6.13)$$

Equations (6.10)-(6.13) are a set of four first order equations that are equivalent to (6.8)-(6.9). ■

Reading Exercise 158 Show that (6.10)-(6.13) 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) with the initial conditions $x_1(t_0) = a_1, x_2(t_0) = a_2, \dots, x_n(t_0) = a_n$ for some time interval $t_0 - \varepsilon < t < t_0 + \varepsilon$ with $\varepsilon > 0$.

Briefly, if the right hand sides of (6.3) are continuous and have continuous partial derivatives near the initial point of interest, 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 below as linear or nonlinear.

- $\dot{x}_1 = x_1 x_2, \dot{x}_2 = x_1 - x_2$.
- $\dot{x}_1 = x_1 + x_2, \dot{x}_2 = x_1 - 2x_2$.
- $\dot{x}_1 = \sin(x_1 + x_2), \dot{x}_2 = \cos(x_1 - x_2)$.
- $\dot{x}_1 = tx_1 - x_2, \dot{x}_2 = x_1 + \sin(t)x_2$.
- $\dot{x}_1 = x_1/(t^2 + 1) + x_1/x_2, \dot{x}_2 = 3$.
- $\dot{x}_1 = e^{x_1+2x_2}, \dot{x}_2 = x_1 - 4x_2$.
- $\dot{x}_1 = x_1 + x_2 + x_3, \dot{x}_2 = x_1 + tx_2 - x_3, \dot{x}_3 = x_1 - 2x_2 + e^t x_3$.
- $\dot{x}_1 = x_1 + x_2 x_3, \dot{x}_2 = x_1 + tx_2 - x_3, \dot{x}_3 = x_1 - 2x_2 + e^t x_3$.

Exercise 6.1.2 Each system below is linear. Classify it as constant or variable coefficient.

- $\dot{x}_1 = x_1 + x_2, \dot{x}_2 = x_1 - x_2$.
- $\dot{x}_1 = x_1 + x_2, \dot{x}_2 = x_1 - 2x_2 + t^2$.
- $\dot{x}_1 = x_1 + x_2, \dot{x}_2 = \cos(t)x_1 + 5x_2$.
- $\dot{x}_1 = tx_1 - x_2, \dot{x}_2 = 4$.
- $\dot{x}_1 = x_2, \dot{x}_2 = 3 + t^3$.
- $\dot{x}_1 = x_1 + x_2 + 5e^t, \dot{x}_2 = x_1 - x_2$.
- $\dot{x}_1 = x_1 + x_2 + x_3, \dot{x}_2 = x_1 + tx_2 - x_3, \dot{x}_3 = x_1 - 2x_2 + 3x_3 - 2t^2$.
- $\dot{x}_1 = x_1 + x_2 + tx_3, \dot{x}_2 = x_1 + x_2 - x_3, \dot{x}_3 = x_1 - 2x_2 + x_3$.

Exercise 6.1.3 Convert each second or higher order equation below into an equivalent system of first order ODE's, with initial conditions.

- $3u''(t) + 5u'(t) + 4u(t) = 0, u(0) = 7, u'(0) = 5$.
- $u''(t) + u'(t) + u(t) = 0, u(0) = 1, u'(0) = 0$.
- $2u''(t) + 2\cos(u'(t)) + u(t) = 0, u(0) = 3, u'(0) = -1$.
- $u''(t) + u'(t)u(t) = 7, u(1) = 2, u'(1) = 4$.
- $u'''(t) + 2u''(t) + u'(t) + 5u(t) = 0, u(0) = 1, u'(0) = 0, u''(0) = -1$. Hint: Take $x_1 = u, x_2 = u',$ and $x_3 = u''$. Two of the ODE's are $\dot{x}_1 = x_2, \dot{x}_2 = x_3$.
- $u^{(4)}(t) + u'''(t) + 4u''(t) + 5u'(t) + 3u(t) = 0, u(0) = 1, u'(0) = 0, u''(0) = -1, u'''(0) = 4$. Hint: Take $x_1 = u, x_2 = u', x_3 = u'',$ and $x_4 = u'''$. Three of the ODE's are $\dot{x}_1 = x_2, \dot{x}_2 = x_3,$ and $\dot{x}_3 = x_4$.

Exercise 6.1.4 Convert each coupled first/second system below into an equivalent system of first order ODE's, with initial conditions.

(a) $u_1''(t) + u_1'(t) - u_2(t) = \sin(t)$, $u_2'(t) - u_2(t) + 3u_1(t) = 0$, $u_1(0) = 1$, $u_1'(0) = 3$, $u_2(0) = -2$. Hint: take $x_1 = u_1$, $x_2 = u_1'$, and $x_3 = u_2$.

(b) $3u_1''(t) + \sin(u_2'(t)) - u_2(t) = t$, $u_2'(t) - 2u_2(t) + u_1(t) = 0$, $u_1(0) = 1$, $u_1'(0) = 3$, $u_2(0) = -2$. Hint: see part (a). ■

Exercise 6.1.5 Consider equations (6.1)-(6.2) in the case that $k_a = 4$, $k_b = 1$, and $k_e = 3$ with $g(t) = 0$ the system becomes

$$\begin{aligned}\dot{u}_P(t) &= -4u_P(t) + 4u_T(t) \\ \dot{u}_T(t) &= u_P(t) - 4u_T(t).\end{aligned}$$

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/bloodstream at time $t = 0$, with none in the tissue.)

Verify that the functions

$$\begin{aligned}u_P(t) &= \frac{e^{-2t}}{2} + \frac{e^{-6t}}{2} \\ u_T(t) &= \frac{e^{-2t}}{4} - \frac{e^{-6t}}{4}\end{aligned}$$

provide a solution with the proper initial conditions. Plot 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 6.1.6 You already know how to solve linear constant coefficient systems using the Laplace transform. Consider the ODE pair

$$\begin{aligned}\dot{x}_1 &= 0x_1 + x_2 \\ \dot{x}_2 &= -x_1 + 0x_2\end{aligned}$$

with initial conditions $x_1(0) = 0$, $x_2(0) = 1$.

- (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 $sX_1(s) = X_2(s)$, $sX_2(s) - 1 = -X_1(s)$ to show that

$$X_1(s) = \frac{1}{s^2 + 1} \quad \text{and} \quad X_2(s) = \frac{s}{s^2 + 1}.$$

Inverse transform to show that $x_1(t) = \sin(t)$, $x_2(t) = \cos(t)$. Verify that these satisfy the ODE's of interest.

- (c) Use the Laplace transform to solve

$$\begin{aligned}\dot{x}_1 &= 2x_1 + 3x_2 + 12e^{-t} \\ \dot{x}_2 &= x_1 + 4x_2\end{aligned}$$

with initial data $x_1(0) = -8$, $x_2(0) = 2$. ■

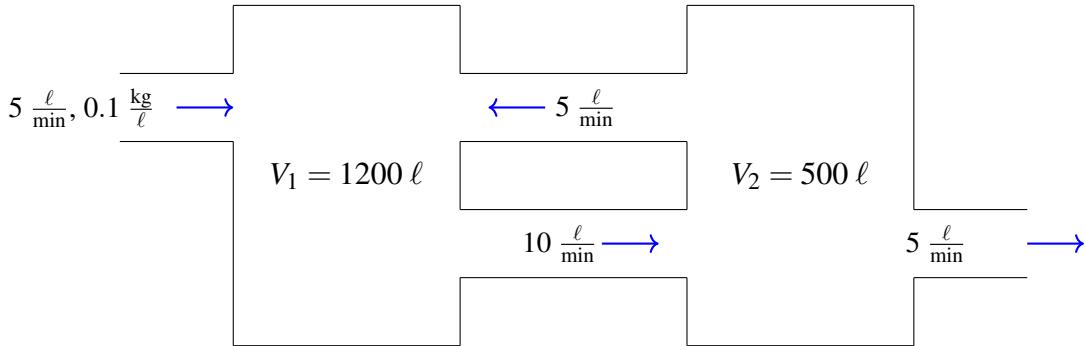


Figure 6.3: A salt tank problem with two tanks.

Exercise 6.1.7 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 at all times the concentration of salt in each tank is uniform over the tank volume (the tanks are well-stirred) and that both tanks start filled with pure water.

If salt is conserved then the rate at which the amount of salt in any given 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 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, both sides of this equation should 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

$$x_1(t) = 120 - \frac{60}{13}e^{-t/40} - \frac{1500}{13}e^{-t/300}$$

$$x_2(t) = 50 + \frac{100}{13}e^{-t/40} - \frac{750}{13}e^{-t/300}$$

satisfy the coupled ODE's for $x_1(t), x_2(t)$ from parts (c) and (d), with $x_1(0) = 0$ and $x_2(0) = 0$. Plot these functions on the range $0 \leq t \leq 2000$. 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?

6.2 Linear Constant Coefficient Homogeneous Systems of Differential Equations

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.

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

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}(t) \quad (6.14)$$

where $\mathbf{x}(t)$ and $\mathbf{b}(t)$ in (6.14) 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 various functions on t and write simply $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{b}$ for (6.14).

As with scalar linear equations, such a system may be homogeneous and nonhomogeneous.

Definition 6.2.1 The linear system of ODEs (6.14) 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 (6.14) 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 (6.14) in the homogeneous case $\mathbf{b}(t) = 0$. The system of interest is thus

$$\dot{\mathbf{x}} = \mathbf{Ax}. \quad (6.15)$$

Eigenvalues and Eigenvectors

To solve a system of the form (6.15) we take inspiration from the scalar case $x'(t) = Ax(t)$ in which A is constant. Solutions to this ODE are easily verified to be of the form $x(t) = e^{At}c$ for any constant c ; we put c on the right in $e^{At}c$ for a reason to be explained shortly. Since a solution to (6.15) is a vector-valued function of t , by analogy to the scalar case we might try an ansatz for (6.15) of the form

$$\mathbf{x}(t) = e^{\lambda t} \mathbf{v} \quad (6.16)$$

for some constant vector $\mathbf{v} = \langle v_1, \dots, v_n \rangle$ that plays the role of c in the scalar equation.

Remark 12 If we truly emulate the scalar ODE $x' = Ax$ it would make even more sense to try an ansatz of the form $\mathbf{x}(t) = e^{\Lambda t} \mathbf{v}$ in (6.15). This works, after we figure out what $e^{\Lambda t}$ means, that is, how to exponentiate a matrix. This is the topic of Section 6.4.

Based on (6.16) the components $x_j(t)$ of $\mathbf{x}(t)$ are given by $x_j(t) = e^{\lambda t} v_j$ 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 (6.15) along with (6.16) to replace $\mathbf{x}(t)$ and then pull the common $e^{\lambda t}$ out in front on both sides (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 (6.16) satisfies $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ then \mathbf{v} must 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 $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$.

■ **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 relevant 2×2 matrix \mathbf{A} here 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

$$\mathbf{w}_1(t) = e^{-2t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{w}_2(t) = e^{-6t} \begin{bmatrix} -3 \\ 1 \end{bmatrix}$$

satisfy the linear system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$. ■

Reading Exercise 159 The constants for the two-compartment LSD model of equations (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 this estimation problem in the projects section 6.5. 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 initial data can be obtain, by constructing 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 $x_1(0) = -1$, $x_2(0) = 3$. 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

$$\mathbf{x}(t) = c_1 \mathbf{w}_1(t) + c_2 \mathbf{w}_2(t) \tag{6.17}$$

where c_1 and c_2 are arbitrary scalars. We can show that the function $\mathbf{x}(t)$ satisfies $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ for any choice of c_1 and c_2 . This can be done by using the linearity of differentiation and matrix-vector

multiplication, so that

$$\begin{aligned}\dot{\mathbf{x}} &= c_1 \dot{\mathbf{w}}_1 + c_2 \dot{\mathbf{w}}_2 \\ &= c_1 \mathbf{A} \mathbf{w}_1 + c_2 \mathbf{A} \mathbf{w}_2 \\ &= \mathbf{A}(c_1 \mathbf{w}_1 + c_2 \mathbf{w}_2) \\ &= \mathbf{A}\mathbf{x}.\end{aligned}\tag{6.18}$$

for any choice of c_1 and c_2 . The function $\mathbf{x}(t)$ in (6.17) is the *general solution* to the ODE system in Example 6.4, for we can adjust c_1 and c_2 to obtain any initial conditions.

For the present example with $x_1(0) = -1$ and $x_2(0) = 3$ use (6.17) to compute that

$$\begin{aligned}\mathbf{x}(0) &= c_1 \mathbf{w}_1(0) + c_2 \mathbf{w}_2(0) \\ &= c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 \\ &= \begin{bmatrix} c_1 + c_2 \\ -3c_1 + c_2 \end{bmatrix}.\end{aligned}$$

To satisfy the initial data we thus need $c_1 - 3c_2 = -1$ and $c_1 + c_2 = 3$, two linear algebraic equations for c_1 and c_2 with solution $c_1 = 2, c_2 = 1$. 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 $x_1(t) = 2e^{-2t} - 3e^{-6t}, x_2(t) = 2e^{-2t} + e^{-6t}$ if written out in component form. ■

The General Procedure for Solving $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$

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 $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ 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 $\mathbf{v}_1, \dots, \mathbf{v}_n$. This is always the case if the eigenvalues for \mathbf{A} are all distinct (and may still be true if there are repeat eigenvalues).

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 $\dot{\mathbf{w}}_j = \mathbf{A} \mathbf{w}_j$ for each j with $1 \leq j \leq n$.
2. Since $\dot{\mathbf{w}}_j = \mathbf{A} \mathbf{w}_j$, the same linearity argument that led to (6.18), shows that any superposition of the \mathbf{w}_j satisfies the ODE $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$. Accordingly, define a vector-valued function as a superposition

$$\begin{aligned}\mathbf{x}(t) &= c_1 \mathbf{w}_1(t) + \cdots + c_n \mathbf{w}_n(t) \\ &= c_1 e^{\lambda_1 t} \mathbf{v}_1 + \cdots + c_n e^{\lambda_n t} \mathbf{v}_n\end{aligned}\tag{6.19}$$

where the c_j are arbitrary constants. The function $\mathbf{x}(t)$ in (6.19) is the *general solution* to the linear system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$, because the c_j can be adjusted to obtain any desired initial data, as shown in the next step.

3. To obtain $\mathbf{x}(0) = \mathbf{x}_0$, substitute $t = 0$ into (6.19) to see that we need

$$c_1 \mathbf{v}_1 + \cdots + c_n \mathbf{v}_n = \mathbf{x}_0.$$

This equation is entirely equivalent to the matrix-vector equation

$$\mathbf{P}\mathbf{c} = \mathbf{x}_0\tag{6.20}$$

where \mathbf{P} is the $n \times n$ matrix with j th column \mathbf{v}_j and \mathbf{c} is the vector with j th component 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. We can solve for $\mathbf{c} = \mathbf{P}^{-1} \mathbf{x}_0$ and use these c_j in (6.19) to obtain the solution $\mathbf{x}(t)$.

Examples

■ **Example 6.6** Exercise 6.1.7 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 that 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. (Using pure water for the inflow pipe makes the system homogeneous.) 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} , given by $\lambda_1 = -1/300$ 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 (6.19),

$$\mathbf{x}(t) = c_1 e^{-t/300} \begin{bmatrix} 2 \\ 1 \end{bmatrix} + c_2 e^{-t/40} \begin{bmatrix} -3 \\ 5 \end{bmatrix}.$$

The last step is to obtain the initial data by solving $\mathbf{x}(0) = c_1 \langle 3, 1 \rangle + c_2 \langle -3, 5 \rangle = \langle 1, 7 \rangle$, or in the matrix-vector form of (6.20),

$$\underbrace{\begin{bmatrix} 2 & -3 \\ 1 & 5 \end{bmatrix}}_{\mathbf{P}} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 7 \end{bmatrix}.$$

The matrix \mathbf{P} is invertible (equivalently, \mathbf{v}_1 and \mathbf{v}_2 are linearly independent) and we find a unique solution $c_1 = 2, c_2 = 1$. 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 160

- Use the functions $\mathbf{w}_1(t)$ and $\mathbf{w}_2(t)$ from Reading Exercise 159 to write out a general solution to the two compartment LSD model (6.1)-(6.2).
- In the study [17] 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 is introduced into the bloodstream/plasma, so $u_P(0) = 140 \mu\text{g}$, 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 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 $u(0) = 4, u'(0) = 0$. Although we already know how to solve this problem, let's approach it by converting 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 $x_1 = u, x_2 = u'$ to obtain

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -5 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (6.21)$$

with $x_1(0) = 4, x_2(0) = 0$. The eigenvalues for the matrix in (6.21) are $\lambda_1 = -1 + 2i$ and $\lambda_2 = -1 - 2i$, with eigenvectors $\mathbf{v}_1 = \langle -1 - 2i, 5 \rangle$ and $\mathbf{v}_2 = \langle -1 + 2i, 5 \rangle$. Note that the eigenvalues are conjugate, and the eigenvectors too, component by component. Based on (6.19) we conclude that

$$\mathbf{x}(t) = c_1 e^{(-1+2i)t} \begin{bmatrix} -1 - 2i \\ 5 \end{bmatrix} + c_2 e^{(-1-2i)t} \begin{bmatrix} -1 + 2i \\ 5 \end{bmatrix} \quad (6.22)$$

is a general solution for this system. We can obtain the desired initial data by setting $\mathbf{x}(0) = \langle 2, 0 \rangle$, which leads to

$$\underbrace{\begin{bmatrix} -1 - 2i & -1 + 2i \\ 5 & 5 \end{bmatrix}}_{\mathbf{P}} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

(this is (6.20)) with solution $c_1 = i, c_2 = -i$; note that c_1 and c_2 are conjugate. The solution with $x_1(0) = 4, x_2(0) = 0$ is then

$$\mathbf{x}(t) = ie^{(-1+2i)t} \begin{bmatrix} -1 - 2i \\ 5 \end{bmatrix} - ie^{(-1-2i)t} \begin{bmatrix} -1 + 2i \\ 5 \end{bmatrix}. \quad (6.23)$$

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 identity to write $e^{(-1+2i)t} = e^{-t}(\cos(2t) + i\sin(2t))$ and $e^{(-1-2i)t} = e^{-t}(\cos(2t) - i\sin(2t))$, then use this in (6.23) 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 just $x_1(t) = 2e^{-t} \sin(2t) + 4e^{-t} \cos(2t)$ and $x_2(t) = -10e^{-t} \sin(2t)$ in component form. The solution is real-valued, as it should be.

Note how the eigenvalues $-1 \pm 2i$ manifest themselves in the solution, especially the real-valued solution. The real part of the eigenvalues dictates the decay rate of the solution and the imaginary part indicates the radial frequency of vibration. ■

Real-Valued General Solutions

If the eigenvalues for \mathbf{A} are complex then the general solution (6.19) will be complex-valued. However if desired 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, 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 (6.19) can be expressed as

$$\mathbf{x}(t) = c_1 \mathbf{w}_1(t) + c_2 \mathbf{w}_2(t) \quad (6.24)$$

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 from Appendix A the vector-valued functions $\mathbf{w}_1(t)$ and $\mathbf{w}_2(t)$ must also be conjugate to each other, so that $\mathbf{w}_2(t) = \overline{\mathbf{w}_1(t)}$. If we break $\mathbf{w}_1(t)$ into its real and imaginary parts as

$$\mathbf{w}_1(t) = \mathbf{y}_r(t) + i\mathbf{y}_i(t)$$

then $\mathbf{w}_2(t) = \mathbf{y}_r(t) - i\mathbf{y}_i(t)$ and the general solution (6.24) can be expressed in the form

$$\begin{aligned} \mathbf{x}(t) &= c_1 \mathbf{w}_1(t) + c_2 \mathbf{w}_2(t) \\ &= c_1(\mathbf{y}_r(t) + i\mathbf{y}_i(t)) + c_2(\mathbf{y}_r(t) - i\mathbf{y}_i(t)) \\ &= \underbrace{(c_1 + c_2)}_{d_1} \mathbf{y}_r(t) + \underbrace{i(c_1 - c_2)}_{d_2} \mathbf{y}_i(t). \end{aligned} \quad (6.25)$$

Compare (6.25) to (4.38). Since c_1 and c_2 are arbitrary constants, so are d_1 and d_2 in (6.25), and (6.25) 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 (6.24) can be replaced by a linear combination of the real-valued quantities $\operatorname{Re}(\mathbf{w}_1(t))$ 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 given by

$$\mathbf{x}(t) = d_1 \mathbf{y}_r(t) + d_2 \mathbf{y}_i(t)$$

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 (4.39).

■ **Example 6.8** To illustrate Theorem 6.2.1, in the underdamped spring-mass system of Example 6.7 use eigenvalue $\lambda_1 = -1 + 2i$ and corresponding eigenvector $\mathbf{v}_1 = \langle -1 - 2i, 5 \rangle$, then find 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 (6.22) can be replaced by the real and imaginary parts on the right above. A real-valued general solution is then given by

$$\mathbf{x}(t) = \underbrace{d_1 e^{-2t} \begin{bmatrix} -\cos(2t) + 2\sin(2t) \\ 5\cos(2t) \end{bmatrix}}_{\operatorname{Re}(e^{\lambda_1 t} \mathbf{v}_1)} + \underbrace{d_2 e^{-2t} \begin{bmatrix} -2\cos(2t) - \sin(2t) \\ 5\sin(2t) \end{bmatrix}}_{\operatorname{Im}(e^{\lambda_1 t} \mathbf{v}_1)}.$$

■

Reading Exercise 161 Redo the computation of Example 6.8 but use 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.

The next example illustrates how to handle constructing 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, equations (6.10)-(6.13). We will focus on the undamped case $c_1 = c_2 = 0$, and choose specific values $k_1 = 2, k_2 = 1, m_1 = 2, m_2 = 1$. 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

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -3/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 \end{bmatrix}. \quad (6.26)$$

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

$$\mathbf{v}_1 = \begin{bmatrix} -i/\sqrt{2} \\ 1/2 \\ i\sqrt{2} \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} i/\sqrt{2} \\ 1/2 \\ -i\sqrt{2} \\ 1 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} i\sqrt{2} \\ -1 \\ -i\sqrt{2} \\ 1 \end{bmatrix}, \quad \mathbf{v}_4 = \begin{bmatrix} -i/\sqrt{2} \\ -1 \\ i\sqrt{2} \\ 1 \end{bmatrix}.$$

Note the eigenvalues come in complex-conjugate pairs, along with their corresponding eigenvectors.

A complex-valued general solution can be constructed using (6.19) 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 we used in the case that \mathbf{A} was 2×2 that led to Theorem 6.2.1. Specifically, replace the terms $e^{\lambda_1 t} \mathbf{v}_1$ and $e^{\lambda_2 t} \mathbf{v}_2$ in the general solution (that involve 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

$$\begin{aligned} \text{Re}(e^{\lambda_1 t} \mathbf{v}_1) &= \begin{bmatrix} \sin(t/\alpha)/\alpha \\ \cos(t/\alpha)/2 \\ \alpha \sin(t/\alpha) \\ \cos(t/\alpha) \end{bmatrix}, \quad \text{Im}(e^{\lambda_1 t} \mathbf{v}_1) = \begin{bmatrix} -\cos(t/\alpha)/\alpha \\ \sin(t/\alpha)/2 \\ -\alpha \cos(t/\alpha) \\ \sin(t/\alpha) \end{bmatrix} \\ \text{Re}(e^{\lambda_3 t} \mathbf{v}_3) &= \begin{bmatrix} -\sin(\alpha t)/\alpha \\ -\cos(\alpha t) \\ \sin(\alpha t)/\alpha \\ \cos(\alpha t) \end{bmatrix}, \quad \text{Im}(e^{\lambda_3 t} \mathbf{v}_3) = \begin{bmatrix} \cos(\alpha t)/\alpha \\ -\sin(\alpha t) \\ -\cos(\alpha t)/\alpha \\ \sin(\alpha) \end{bmatrix} \end{aligned}$$

These can be assembled into a real-valued general solution

$$\mathbf{w}(t) = c_1 \begin{bmatrix} \sin(t/\alpha)/\alpha \\ \cos(t/\alpha)/2 \\ \alpha \sin(t/\alpha) \\ \cos(t/\alpha) \end{bmatrix} + c_2 \begin{bmatrix} -\cos(t/\alpha)/\alpha \\ \sin(t/\alpha)/2 \\ -\alpha \cos(t/\alpha) \\ \sin(t/\alpha) \end{bmatrix} + c_3 \begin{bmatrix} -\sin(\alpha t)/\alpha \\ -\cos(\alpha t) \\ \sin(\alpha t)/\alpha \\ \cos(\alpha t) \end{bmatrix} + c_4 \begin{bmatrix} \cos(\alpha t)/\alpha \\ -\sin(\alpha t) \\ -\cos(\alpha t)/\alpha \\ \sin(\alpha) \end{bmatrix}.$$

Note that the solution is oscillatory and never decays, since 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

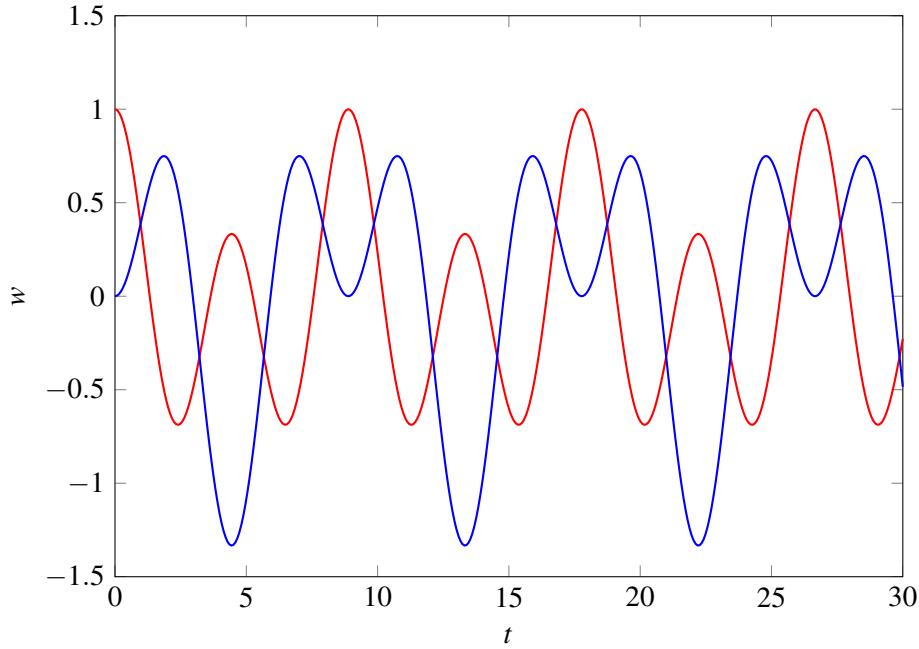


Figure 6.4: Solution to (6.10)-(6.13), position $w_1(t)$ of mass 1 in red, position $w_3(t)$ of mass 2 in blue.

in Figure 6.4, $w_1(t)$ in red, $w_3(t)$ in blue. 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. ■

6.2.4 Defective Matrices

The above analysis is predicated on the $n \times n$ matrix \mathbf{A} having n linearly independent eigenvectors, so that the system (6.20) can be solved to obtain any desired initial conditions, since the matrix \mathbf{P} will then 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

$$u''(t) + 4u'(t) + 4u(t) = 0, \quad (6.27)$$

which was analyzed in Example 4.10. If we let $x_1 = u, x_2 = u'$ then (6.27) is equivalent to the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ -4 & -4 \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (6.28)$$

The matrix \mathbf{A} has eigenvalues $\lambda_1 = \lambda_2 = -2$, that is to say 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 (4.45). In that case the characteristic equation had a double root at λ and we could produce a solution $x(t) = ce^{\lambda t}$ to the ODE, but not a second independent solution. The remedy was to consider the ansatz $x(t) = (c_1 + c_2t)e^{\lambda t}$, and this yielded another solution $x(t) = te^{\lambda t}$. If this approach worked once, it should work 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 one type of solution to this system, namely

$$\mathbf{w}_1(t) = e^{\lambda t} \mathbf{v} \quad (6.29)$$

or any multiple thereof. The goal now is to produce a second solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ that is not a multiple of $\mathbf{w}_1(t)$. To produce such solution we will try something of the form

$$\mathbf{w}_2(t) = e^{\alpha t}(\mathbf{v}_1 + t\mathbf{v}_2) \quad (6.30)$$

in $\dot{\mathbf{x}} = \mathbf{Ax}$, where the scalar α and vectors $\mathbf{v}_1, \mathbf{v}_2$ 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

$$\dot{\mathbf{w}}_2 = e^{\alpha t}(\alpha \mathbf{v}_1 + \alpha t \mathbf{v}_2 + \mathbf{v}_2).$$

Inserting this expression for $\dot{\mathbf{x}}$ into $\dot{\mathbf{x}} = \mathbf{Ax}$ along with using (6.30) for $\mathbf{x} = \mathbf{w}_2$ and dividing though by $e^{\alpha t}$ produces

$$\alpha \mathbf{v}_1 + t \alpha \mathbf{v}_2 + \mathbf{v}_2 = \mathbf{Av}_1 + t \mathbf{Av}_2. \quad (6.31)$$

The goal is to make both sides of (6.31) 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 (6.31) with the t coefficients will cancel. Of course $\mathbf{Av}_2 = \alpha \mathbf{v}_2$ means that \mathbf{v}_2 should be an eigenvector for \mathbf{A} with eigenvalue α . Since \mathbf{A} has only the eigenvalue λ and the corresponding eigenvector \mathbf{v} , we must take $\alpha = \lambda$ and $\mathbf{v}_2 = \mathbf{v}$ in (6.31).

With these choices (6.31) becomes $\lambda \mathbf{v}_1 + t \lambda \mathbf{v}_2 + \mathbf{v}_2 = \mathbf{Av}_1 + t \mathbf{Av}_2$, where $t \lambda \mathbf{v} = t \mathbf{Av}$. Cancelling these like terms produces

$$\lambda \mathbf{v}_1 + \mathbf{v} = \mathbf{Av}_1$$

or, after a bit of rearrangement,

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{v}_1 = \mathbf{v}. \quad (6.32)$$

It would be nice to assert that we can take $\mathbf{v}_1 = (\mathbf{A} - \lambda \mathbf{I})^{-1}\mathbf{v}$, but since 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 (6.32) 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 the equation (6.32) has a solution for \mathbf{v}_1 , and in fact there are infinitely many solutions. For a proof of this fact see [22]. We can use \mathbf{v}_1 in the ansatz (6.30) 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 $\mathbf{w}_1(t)$ defined by (6.29) to construct a general solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ as a linear combination $\mathbf{x}(t) = c_1\mathbf{w}_1(t) + c_2\mathbf{w}_2(t)$ or

$$\mathbf{x}(t) = c_1 e^{\lambda t} \mathbf{v} + c_2 e^{\lambda t} (\mathbf{v}_1 + t\mathbf{v}). \quad (6.33)$$

For the formula (6.33) 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} and \mathbf{v}_1 is any vector that satisfies (6.32), $(\mathbf{A} - \lambda \mathbf{I})\mathbf{v}_1 = \mathbf{v}$.

■ **Example 6.11** Let's return to the critically damped spring-mass system of Example 6.10 in which \mathbf{A} has defective double eigenvalue $\lambda = -2$ with eigenvector $\mathbf{v} = \langle -1, 2 \rangle$. To find \mathbf{v}_1 in the general solution (6.33) note that (6.32) becomes the equation

$$\begin{bmatrix} 2 & 1 \\ -4 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

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. According to (6.33) then

$$\mathbf{x}(t) = c_1 e^{-2t} \begin{bmatrix} -1 \\ 2 \end{bmatrix} + c_2 e^{-2t} \left(\begin{bmatrix} 1 \\ -3 \end{bmatrix} + t \begin{bmatrix} -1 \\ 2 \end{bmatrix} \right)$$

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 $-c_1 + c_2 = 1, 2c_1 - 3c_2 = -5$ 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 above for the 2×2 case can occur in larger systems of ODEs. 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 (6.19). The procedure above 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 as well as the technique of matrix exponentiation in the next section can handle these cases.

6.2.5 Exercises

Exercise 6.2.1 For each system of ODE's below:

- 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 this to write out a general solution using (6.19).
- If any eigenvalues for \mathbf{A} are complex, write out a real-valued general solution.
- Use either general solution to obtain the given initial data.

- (a) $\dot{x}_1 = 7x_1 - 4x_2, \dot{x}_2 = 20x_1 - 11x_2$ with $x_1(0) = 3, x_2(0) = 8$.
 (b) $\dot{x}_1 = -x_2, \dot{x}_2 = 6x_1 - 5x_2$ with $x_1(0) = 2, x_2(0) = 5$.
 (c) $\dot{x}_1 = x_1 - x_2, \dot{x}_2 = 5x_1 - 3x_2$ with $x_1(0) = 0, x_2(0) = 2$.
 (d) $\dot{x}_1 = -2x_1 - 3x_2, \dot{x}_2 = 3x_1 - 2x_2$ with $x_1(0) = 2, x_2(0) = -2$.
 (e) $\dot{x}_1 = 2x_1 - 1x_2 + 2x_3, \dot{x}_2 = 7x_1 + 6x_2 - 11x_3, \dot{x}_3 = 6x_1 + 6x_2 - 11x_3$ with $x_1(0) = 1, x_2(0) = -3, x_3(0) = -2$.
 (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 $x_1(0) = -2, x_2(0) = 2, x_3(0) = -4$.
 (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 $x_1(0) = 2, x_2(0) = 1, x_3(0) = 4, x_4(0) = 1$.

Exercise 6.2.2 The matrices below are defective. In each case

- Formulate the system as $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$, by explicitly writing out the matrix \mathbf{A} .
- Find the eigenvalues and eigenvectors of \mathbf{A} and use (6.33) 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 $x_1(0) = 1, x_2(0) = 3$.
 (b) $\dot{x}_1 = 7x_1 - 3x_2, \dot{x}_2 = 12x_1 - 5x_2$ with $x_1(0) = 1, x_2(0) = 1$.
 (c) $\dot{x}_1 = 5x_1 + 4x_2, \dot{x}_2 = -4x_1 - 3x_2$ with $x_1(0) = 1, x_2(0) = 0$.
 (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 $x_1(0) = 1, x_2(0) = 0, x_3(0) = -1$. (This is a 3×3 system, but the technique of Section 6.2.4 can easily be adapted.)

Exercise 6.2.3 Consider the system of linear ODE's $\dot{x}_1 = x_2, \dot{x}_2 = -x_2, \dot{x}_3 = x_1$, with initial data $x_1(0) = a_1, x_2(0) = a_2, x_3(0) = a_3$.

- (a) Formulate this as $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ with an appropriate initial condition.
- (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 6.2.4 Consider the double spring-mass-damper system of (6.10)-(6.13).

- (a) Write out the matrix \mathbf{A} that governs this linear system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$.
- (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} . What do they say about the motion of the system?
- (c) Repeat part (b) but use $c_1 = c_2 = 10$ (leave k_1, k_2, m_1, m_2 the same) 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 6.2.5 Consider a two compartment/salt tank problem in the arrangement of Figure 6.5. 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 $t = 0$.

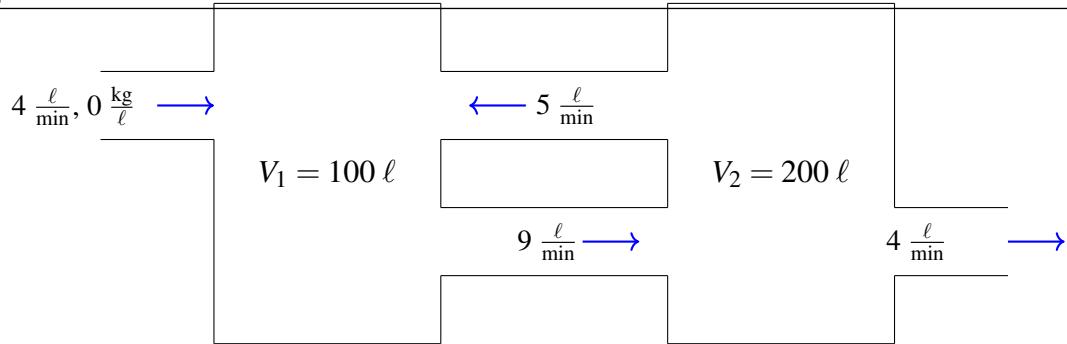


Figure 6.5: A salt tank problem with two tanks.

- (a) Formulate this as a homogeneous linear system of ODE's in the form $\dot{\mathbf{x}} = \mathbf{Ax}$. Explicitly write out \mathbf{A} . What are the initial conditions?
- (b) Solve the system using the method of Laplace transforms.
- (c) Solve the system using the method of undetermined coefficients.

6.3 Linear Constant Coefficient Nonhomogeneous Systems of Differential Equations

In this section we will use two different approaches to solve nonhomogeneous systems.

6.3.1 The Nonhomogeneous Equation $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f}$ via Laplace Transforms

A nonhomogeneous system of linear constant coefficient ODEs as in Definition 6.2.1 can be handled either by using the Laplace transform or by using the method of undetermined coefficients. Either approach is a very natural extension of the corresponding ideas for scalar equations. Let's first consider the method of Laplace transforms.

Laplace Transforming Systems of ODEs

If you did Exercise 6.1.6, you've already seen this technique in action. Consider a system of n ODEs of the form

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f} \quad (6.34)$$

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) (6.34) becomes

$$s\mathbf{X}(s) - \mathbf{x}_0 = \mathbf{AX}(s) + \mathbf{F}(s)$$

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

$$(s\mathbf{I} - \mathbf{A})\mathbf{X}(s) = \mathbf{F}(s) + \mathbf{x}_0 \quad (6.35)$$

where \mathbf{I} denotes the $n \times n$ identity matrix. The matrix equation (6.35) 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 the $X_k(s)$ and each inverse transformed to find $x_k(t)$. This also has the advantage of allowing us to deal with discontinuous and/or impulsive driving functions.

Here is an example that illustrates the technique.

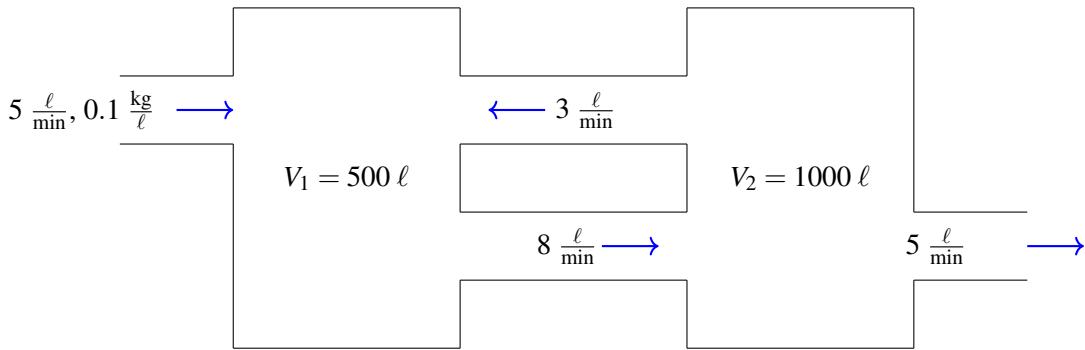


Figure 6.6: A salt tank problem with two tanks.

■ Example 6.12 Consider a salt tank problem with two tanks, volumes and flow rates as indicated in Figure 6.6. (See also Exercise 6.1.7). Let $x_1(t)$ denote the amount of salt in tank 1 at time t and $x_2(t)$ 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.

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 $(8 \frac{\ell}{\text{min}}) \times (\frac{x_1}{500} \frac{\text{kg}}{\ell}) = \frac{2x_1}{125} \frac{\text{kg}}{\text{min}}$. Since salt is conserved we conclude

$$\dot{x}_1 = -\frac{2}{125}x_1 + \frac{3}{1000}x_2 + \frac{1}{2}. \quad (6.36)$$

Similar conservation reasoning applied to tank 2 yields

$$\dot{x}_2 = \frac{2}{125}x_1 - \frac{1}{125}x_2. \quad (6.37)$$

Laplace transforming the system (6.36)-(6.37) and using $x_1(0) = x_2(0) = 0$ (the tanks start filled with pure water) yields an algebraic system of equations

$$\begin{aligned} sX_1(s) &= -\frac{2}{125}X_1(s) + \frac{3}{1000}X_2(s) + \frac{1}{2s} \\ sX_2(s) &= \frac{2}{125}X_1(s) + \frac{1}{125}X_2(s) \end{aligned} \quad (6.38)$$

where $X_1(s) = \mathcal{L}(x_1(t))$ and $X_2(s) = \mathcal{L}(x_2(t))$. The system (6.38) can be solved for $X_1(s)$ and $X_2(s)$ to yield

$$\begin{aligned} X_1(s) &= \frac{50(125s+1)}{s(12500s^2+300s+1)} \\ X_2(s) &= \frac{100}{s(12500s^2+300s+1)}. \end{aligned}$$

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 is 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. ■

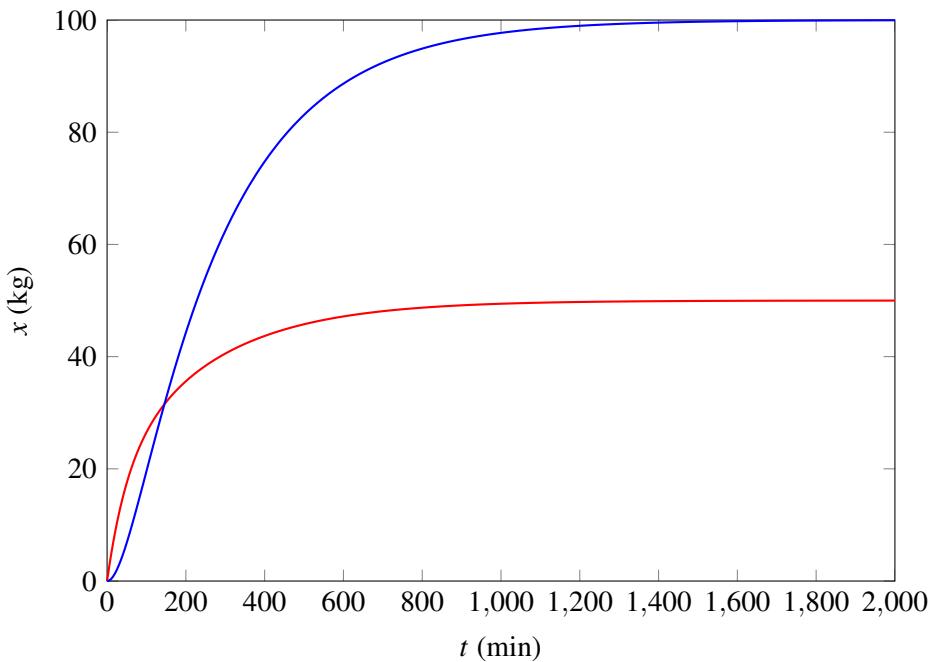


Figure 6.7: Solution to system (6.36)-(6.37), $x_1(t)$ in red, $x_2(t)$ in blue.

The Laplace transform offers a practical way to solve a nonhomogeneous system analytically, although a computer algebra system is helpful for the gory computations.

Remark 13 When using the Laplace transform to solve a system, (6.35) yields the transform of the solution as

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{F}(s) + (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}_0 \quad (6.39)$$

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

$$X(s) = F(s)/(s - a) + x_0/(s - a).$$

Note the parallel structure of this scalar equation with that of (6.39). Also observe that $1/(s - a)$ 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 Section 6.4.

6.3.2 Solution to $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f}$ via Undetermined Coefficients

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 (6.34). Let $\mathbf{x}_h(t)$ denote a general solution to the homogeneous system $\dot{\mathbf{x}} = \mathbf{Ax}$, formed according to (6.19) or other methods, as appropriate. This general solution comes with n arbitrary constants c_1, \dots, c_n that can be used to obtain any desired initial conditions. Let $\mathbf{x}_p(t)$ denote any particular solution to (6.34). We will use the method of undetermined coefficients to produce $\mathbf{x}_p(t)$ and then form the vector-valued function

$$\mathbf{x}(t) = \mathbf{x}_p(t) + \mathbf{x}_h(t). \quad (6.40)$$

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 $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f}$ since

$$\begin{aligned}\dot{\mathbf{x}} &= \dot{\mathbf{x}}_p + \dot{\mathbf{x}}_h \\ &= \mathbf{Ax}_p + \mathbf{f} + \mathbf{Ax}_h \\ &= \mathbf{Ax} + \mathbf{f}.\end{aligned}$$

Moreover, \mathbf{x} contains n arbitrary constants that can be used to obtain any desired initial conditions.

Finding the 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 . We next consider how such particular solutions can be found.

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 (6.34) 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 $x_1(0) = 1, x_2(0) = 2$.

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

$$\mathbf{x}_h(t) = c_1 e^{-t} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + c_2 e^{-3t} \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

Now we seek 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 seek a particular solution to $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{f}$ of the form

$$\mathbf{x}_p(t) = e^{-4t} \mathbf{v}$$

for some undetermined constant vector \mathbf{v} . To determine \mathbf{v} we substitute this 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

$$\mathbf{x}_p(t) = e^{-4t} \begin{bmatrix} -1 \\ -3 \end{bmatrix}.$$

A general solution to the nonhomogeneous system is, from (6.40), given by

$$\mathbf{x}(t) = e^{-4t} \begin{bmatrix} -1 \\ -3 \end{bmatrix} + c_1 e^{-t} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + c_2 e^{-3t} \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

Requiring $\mathbf{x}(0) = \langle 1, 2 \rangle$ leads to equations $-1 + c_1 + c_2 = 1$, $-3 + 2c_1 + 3c_2 = 2$ with solution $c_1 = 1, c_2 = 1$. The solution to the nonhomogeneous system is

$$\mathbf{x}(t) = e^{-4t} \begin{bmatrix} -1 \\ -3 \end{bmatrix} + e^{-t} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + e^{-3t} \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

■

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 (6.26). Suppose now that the second mass is acted on by an external force $f(t)$. In this case (6.13) 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

$$\dot{\mathbf{w}} = \mathbf{Aw} + \mathbf{f} \quad (6.41)$$

where \mathbf{f} is the vector

$$\mathbf{f} = \langle 0, 0, 0, f(t) \rangle.$$

A general solution $\mathbf{w}_h(t)$ to the homogeneous equation $\dot{\mathbf{w}} = \mathbf{Aw}$ was computed 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

$$\mathbf{w}_p(t) = \cos(\omega t) \mathbf{a} + \sin(\omega t) \mathbf{b} \quad (6.42)$$

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

$$-\omega \sin(\omega t) \mathbf{a} + \omega \cos(\omega t) \mathbf{b} = \cos(\omega t) \mathbf{Aa} + \sin(\omega t) \mathbf{Ab} + \sin(\omega t) \mathbf{e}_4. \quad (6.43)$$

As in the scalar case, match the $\sin(\omega t)$ and $\cos(\omega t)$ terms on the left and right in (6.43) to find that the vectors \mathbf{a} and \mathbf{b} must satisfy

$$\mathbf{Ab} + \omega \mathbf{a} = -\mathbf{e}_4 \quad (6.44)$$

$$\mathbf{Aa} - \omega \mathbf{b} = \mathbf{0}. \quad (6.45)$$

To find \mathbf{a} and \mathbf{b} , solve (6.45) for $\mathbf{b} = (\mathbf{Aa})/\omega$ and use this in (6.44) to obtain

$$\frac{1}{\omega} \mathbf{A}^2 \mathbf{a} + \omega \mathbf{a} = -\mathbf{e}_4$$

which can be written as

$$(\mathbf{A}^2 + \omega^2 \mathbf{I})\mathbf{a} = -\omega \mathbf{e}_4 \quad (6.46)$$

after multiplying through by ω . We can solve (6.46) to find the vector \mathbf{a} and then use (6.45) to compute $\mathbf{b} = (\mathbf{A}\mathbf{a})/\omega$. This assumes that the matrix $(\mathbf{A}^2 + \omega^2 \mathbf{I})$ on the left in (6.46) is invertible so that a solution for \mathbf{a} is guaranteed to exist.

As an illustration, suppose $\omega = 1$. In this case (6.46) 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 $\mathbf{b} = (\mathbf{A}\mathbf{a})/\omega$ we find $\mathbf{b} = \langle -1, 0, -1, 0 \rangle$. A particular solution to (6.41) is given by $\mathbf{w}_p(t)$ in (6.42) 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 $w_1(t) = -\sin(t), w_2(t) = -\cos(t), w_3(t) = -\sin(t), w_4(t) = -\cos(t)$ 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 $\dot{\mathbf{x}} = \mathbf{Ax}$ is dictated largely by the eigenvalues of \mathbf{A} . From the general solution (6.19) we see that if all of the eigenvalues have negative real part (this includes simply being 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 $c_k = 0$. This conclusion holds even for defective matrices. When the eigenvalues are all merely nonnegative, the situation is more delicate; see Exercise 6.2.3.

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, 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 may arise when we approximate nonlinear systems, something we'll consider in the next chapter.

6.3.4 Exercises

Exercise 6.3.1 For each system of ODE's below, Laplace transform the ODE's 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 $x_1(t), x_2(t), \dots$

- (a) $\dot{x}_1 = 7x_1 - 4x_2, \dot{x}_2 = 20x_1 - 11x_2$ with $x_1(0) = 3, x_2(0) = 8$. Also, compare to the result

of part (a) of Exercise 6.2.1.

- (b) $\dot{x}_1 = 7x_1 - 4x_2 + 3e^{-2t}$, $\dot{x}_2 = 20x_1 - 11x_2 + 7e^{-2t}$ with $x_1(0) = 2$, $x_2(0) = 3$.
- (c) $\dot{x}_1 = -6x_1 + 2x_2 - 3e^{-3t}$, $\dot{x}_2 = -15x_1 + 5x_2 - 9e^{-3t}$ with $x_1(0) = 1$, $x_2(0) = 2$.
- (d) $\dot{x}_1 = 3x_1 - 2x_2 + 1 - 5t$, $\dot{x}_2 = 10x_1 - 6x_2 - 1 - 16t$ with $x_1(0) = 1$, $x_2(0) = 2$.
- (e) $\dot{x}_1 = 3x_1 - 2x_2 - \sin(t) - \cos(t)$, $\dot{x}_2 = 10x_1 - 6x_2 - 3\sin(t)$ with $x_1(0) = 1$, $x_2(0) = 3$.
- (f) $\dot{x}_1 = 3x_1 - 2x_2 + 5\cos(t) - 3\sin(t)$, $\dot{x}_2 = 10x_1 - 6x_2 + 12\cos(t) - 12\sin(t)$ with $x_1(0) = 0$, $x_2(0) = 3$.
- (g) $\dot{x}_1 = 2x_1 - 4x_2 - 3x_3 - \sin(t) + \cos(t)$, $\dot{x}_2 = -1x_1 - 1x_2 + x_3$, $\dot{x}_3 = 6x_1 - 6x_2 - 7x_3 - \sin(t) + \cos(t)$ with $x_1(0) = 1$, $x_2(0) = 2$, $x_3(0) = 1$.
- (h) $\dot{x}_1 = 2x_1 - 4x_2 - 3x_3 - 5$, $\dot{x}_2 = -1x_1 - 1x_2 + x_3 + 2$, $\dot{x}_3 = 6x_1 - 6x_2 - 7x_3 - 13$ with $x_1(0) = 1$, $x_2(0) = 1$, $x_3(0) = -1$.

Exercise 6.3.2 For each system of ODE's below:

- 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 this to write out a general solution to the homogeneous system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ using (6.19).
- Form the general solution $\mathbf{x}(t)$ to the nonhomogeneous system and obtain the given initial data.

- (a) $\dot{x}_1 = 7x_1 - 4x_2$, $\dot{x}_2 = 20x_1 - 11x_2$ with $x_1(0) = 3$, $x_2(0) = 8$. Also, compare to the result of part (a) of Exercises 6.2.1 and 6.3.1.
- (b) $\dot{x}_1 = 7x_1 - 4x_2 + 3e^{-2t}$, $\dot{x}_2 = 20x_1 - 11x_2 + 7e^{-2t}$ with $x_1(0) = 2$, $x_2(0) = 3$.
- (c) $\dot{x}_1 = -6x_1 + 2x_2 - 3e^{-3t}$, $\dot{x}_2 = -15x_1 + 5x_2 - 9e^{-3t}$ with $x_1(0) = 1$, $x_2(0) = 2$.
- (d) $\dot{x}_1 = 3x_1 - 2x_2 + 1 - 5t$, $\dot{x}_2 = 10x_1 - 6x_2 - 1 - 16t$ with $x_1(0) = 1$, $x_2(0) = 2$.
- (e) $\dot{x}_1 = 3x_1 - 2x_2 - \sin(t) - \cos(t)$, $\dot{x}_2 = 10x_1 - 6x_2 - 3\sin(t)$ with $x_1(0) = 1$, $x_2(0) = 3$.
- (f) $\dot{x}_1 = 3x_1 - 2x_2 + 5\cos(t) - 3\sin(t)$, $\dot{x}_2 = 10x_1 - 6x_2 + 12\cos(t) - 12\sin(t)$ with $x_1(0) = 0$, $x_2(0) = 3$.
- (g) $\dot{x}_1 = 2x_1 - 4x_2 - 3x_3 - \sin(t) + \cos(t)$, $\dot{x}_2 = -1x_1 - 1x_2 + x_3$, $\dot{x}_3 = 6x_1 - 6x_2 - 7x_3 - \sin(t) + \cos(t)$ with $x_1(0) = 1$, $x_2(0) = 2$, $x_3(0) = 1$.
- (h) $\dot{x}_1 = 2x_1 - 4x_2 - 3x_3 - 5$, $\dot{x}_2 = -1x_1 - 1x_2 + x_3 + 2$, $\dot{x}_3 = 6x_1 - 6x_2 - 7x_3 - 13$ with $x_1(0) = 1$, $x_2(0) = 1$, $x_3(0) = -1$.

Exercise 6.3.3 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.

- (a) Formulate this as a nonhomogeneous linear system of ODE's in the form $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{f}$. Explicitly write out \mathbf{A} and \mathbf{f} . What are the initial conditions?
- (b) Solve the system using the method of Laplace transforms.
- (c) Solve the system using the method of undetermined coefficients.

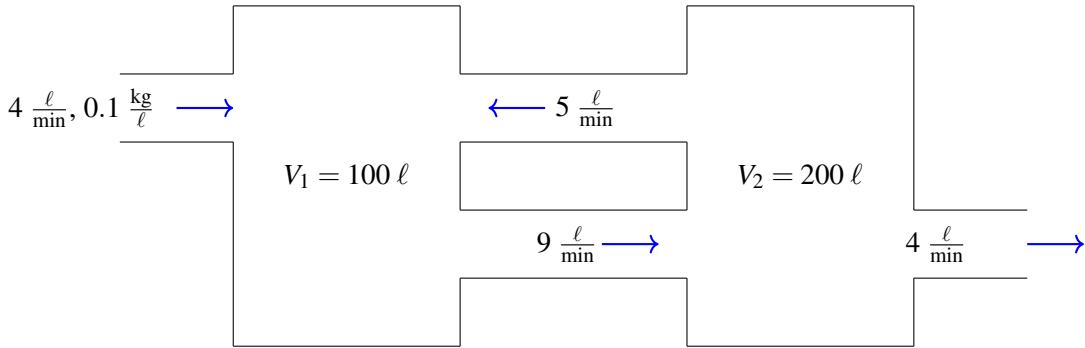


Figure 6.8: A salt tank problem with two tanks.

Exercise 6.3.4 Consider the system of Example 6.14.

- (a) Find a particular solution $\mathbf{w}_p(t)$ in the case that $\omega = 2$ (so the system is driven at $\omega = 2$ radians per second).
- (b) What difficulty arises when you seek a particular solution of the form (6.42) when $\omega = \sqrt{2}$?
- (c) There is one other (positive) value for ω in which a particular solution of the form (6.42) fails to exist. Find that value. Hint: look back at the homogeneous system solution in Example 6.9, and in particular 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 ODE's

$$\begin{aligned}\dot{q} &= \frac{V(t)}{R_1} - \frac{1}{R_1 C} q - I_2 \\ \dot{I}_2 &= \frac{1}{LC} q - \frac{R_2}{L} I_2.\end{aligned}\tag{6.47}$$

- (e) With $\mathbf{x} = \langle q, I_2 \rangle$, formulate (6.47) in the form $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{f}(t)$; show \mathbf{A} and $\mathbf{f}(t)$ explicitly.

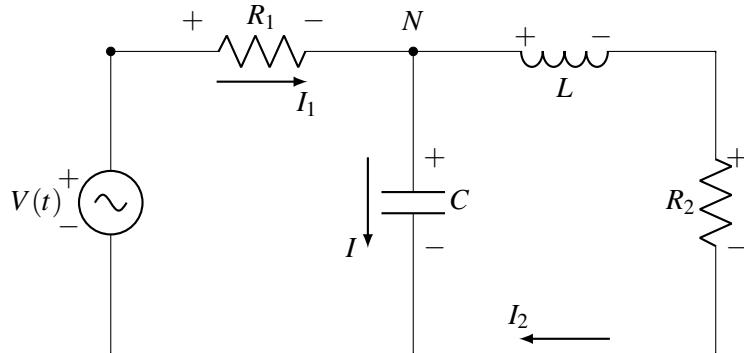


Figure 6.9: Double loop RC/RL circuit.

- (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. Plot $q(t)$ and $i_2(t)$ on the interval $0 \leq t \leq 0.05$ seconds.

6.4 The Matrix Exponential

The matrix exponential is a powerful computational and conceptual tool for analyzing systems of linear, constant coefficient, ODEs. This section offers a quick introduction to the technique, with examples and exercises. It also includes 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 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 13, in which it was shown that the Laplace transform of the vector $\mathbf{X}(s) = \mathcal{L}(\mathbf{x}(t))$ is given by $\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1} \mathbf{x}_0$, 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 t is a scalar and \mathbf{A} a matrix; the product $\mathbf{A}t$ is undefined. Then, in a nutshell

- $e^{t\mathbf{A}}$ 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 $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$. Note \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. Inspiration can be drawn from the Taylor series for e^t , which is

$$\begin{aligned} e^t &= 1 + t + \frac{t^2}{2} + \frac{t^3}{6} + \dots \\ &= \sum_{k=0}^{\infty} \frac{t^k}{k!}. \end{aligned} \quad (6.48)$$

The series (6.48) converges for any real number t . In fact, the series converges for any complex number as well.

We define $e^{\mathbf{B}}$ for an $n \times n$ matrix similarly as

$$\begin{aligned} e^{\mathbf{B}} &= \mathbf{I} + \mathbf{B} + \frac{\mathbf{B}^2}{2} + \frac{\mathbf{B}^3}{6} + \dots \\ &= \sum_{k=0}^{\infty} \frac{\mathbf{B}^k}{k!}. \end{aligned} \quad (6.49)$$

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 (6.49) is an $n \times n$ matrix, so the sum of any finite number of terms in (6.49)

$$\mathbf{S}_m = \sum_{k=0}^m \frac{\mathbf{B}^k}{k!} \quad (6.50)$$

makes sense as an $n \times n$ matrix.

3. It's not obvious that the sum on the right in (6.49) (the limit of (6.50) as $m \rightarrow \infty$) converges, but it does.

For any real number t the infinite sum on the right in the scalar Taylor series (6.48) converges to some real number, and this number is e^t . In fact the sum on the right in (6.48) can be taken as the very definition of e^t . Perhaps we can similarly show that the sum on the right in (6.49) 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 (6.50) 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 (6.50) 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 (6.50) 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

$$\mathbf{S}_{10} = \sum_{k=0}^{10} \frac{\mathbf{B}^k}{k!} \approx \begin{bmatrix} 5.069 & -2.350 \\ 4.700 & -1.983 \end{bmatrix}$$

to three significant figures. It seems that as m increases, the sum converges to something. As in the scalar case (6.48), it appears that this convergence is facilitated by the rapid growth of $k!$ in the denominator of each summand. ■

Reading Exercise 162 Let

$$\mathbf{B} = \begin{bmatrix} 5 & -2 \\ 6 & -2 \end{bmatrix}.$$

Compute the sum \mathbf{S}_m in (6.50) for $m = 1, 2, 5, 10$; a computer algebra system is helpful, and you should 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 (6.50) then it is a fact that these entries converge as $m \rightarrow \infty$. Specifically, for each j and k with $1 \leq j, k \leq n$,

$$\lim_{m \rightarrow \infty} (\mathbf{S}_m)_{jk} = E_{jk} \tag{6.51}$$

for some limit E_{jk} . For a proof of this see [30]. This phenomenon was clear in Example 6.15 and Reading Exercise 162, 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 equation (6.51).

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 (6.49). For a proof of Properties (2) and (4) see [30].

Reading Exercise 163 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 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 (6.49). When written out explicitly we find that

$$\begin{aligned} e^{t\mathbf{A}} &= \mathbf{I} + t\mathbf{A} + t^2 \frac{\mathbf{A}^2}{2} + t^3 \frac{\mathbf{A}^3}{6} + \dots \\ &= \sum_{k=0}^{\infty} t^k \frac{\mathbf{A}^k}{k!}. \end{aligned} \quad (6.52)$$

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 166) that $e^{t\mathbf{A}}$ is given by

$$e^{t\mathbf{A}} = \begin{bmatrix} -3 + 4e^{-t} & 2 - 2e^{-t} \\ 6e^{-t} - 6 & 4 - 3e^{-t} \end{bmatrix}.$$

Computing $e^{t\mathbf{A}}$ by using the Taylor series (6.52) 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 (6.52) and term-by-term differentiation (if permitted), and a bit of algebraic manipulation. We find

$$\begin{aligned} \frac{d(e^{t\mathbf{A}})}{dt} &= \sum_{k=0}^{\infty} \frac{d(t^k)}{dt} \frac{\mathbf{A}^k}{k!} \\ &= \sum_{k=1}^{\infty} k t^{k-1} \frac{\mathbf{A}^k}{k!} && \text{(the } k=0 \text{ term drops out)} \\ &= \sum_{k=1}^{\infty} t^{k-1} \frac{\mathbf{A}^k}{(k-1)!} && (k/k! = 1/(k-1)!) \\ &= \sum_{j=0}^{\infty} t^j \frac{\mathbf{A}^{j+1}}{j!} && \text{(let } j=k-1 \text{ above)} \\ &= \mathbf{A} \sum_{j=0}^{\infty} t^j \frac{\mathbf{A}^j}{j!} && \text{(factor an } \mathbf{A} \text{ out of the sum)} \\ &= \mathbf{A} e^{t\mathbf{A}}. \end{aligned} \quad (6.53)$$

We could also have factored the common \mathbf{A} term in the infinite sum above 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 above that lead to (6.53) (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 given by

$$\frac{d}{dt}(e^{t\mathbf{A}}) = \mathbf{A} e^{t\mathbf{A}} = e^{t\mathbf{A}}\mathbf{A}. \quad (6.54)$$

Both $\mathbf{A}e^{t\mathbf{A}}$ and $e^{t\mathbf{A}}\mathbf{A}$ in (6.54) 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. Note also the similarity of (6.54) to the scalar computation $d(e^{at})/dt = ae^{at} = e^{at}a$ that follows from the chain rule.

Reading Exercise 164

- 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 .
- For the matrix \mathbf{A} in Example 6.16, compute $\frac{d}{dt}(e^{t\mathbf{A}})$ by using (6.54), with each expression $\mathbf{A}e^{t\mathbf{A}}$ and $e^{t\mathbf{A}}\mathbf{A}$. Verify that you obtain the same answer as part (a).

6.4.5 Solving ODEs with the Matrix Exponential

The Fundamental Matrix Solution

The matrix exponential allows us to solve $\dot{\mathbf{x}} = \mathbf{Ax}$ 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 $\dot{\mathbf{x}} = \mathbf{Ax}$; according to (6.54), $\mathbf{X}(t)$ satisfies

$$\dot{\mathbf{X}} = \mathbf{AX}.$$

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

$$[\dot{\mathbf{x}}_1(t) \quad \dot{\mathbf{x}}_2(t) \quad \cdots \quad \dot{\mathbf{x}}_n(t)] = [\mathbf{Ax}_1(t) \quad \mathbf{Ax}_2(t) \quad \cdots \quad \mathbf{Ax}_n(t)].$$

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

$$\begin{aligned} \frac{d}{dt} \left(\sum_{k=1}^n c_k \mathbf{x}_k(t) \right) &= \sum_{k=1}^n c_k \dot{\mathbf{x}}_k \\ &= \sum_{k=1}^n c_k \mathbf{Ax}_k \\ &= \mathbf{A} \left(\sum_{k=1}^n c_k \mathbf{x}_k(t) \right). \end{aligned} \tag{6.55}$$

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, $\mathbf{x}_k(0) = \mathbf{e}_k$ 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

$$\mathbf{x}(t) = c_1 \mathbf{x}_1(t) + \cdots + c_n \mathbf{x}_n(t). \tag{6.56}$$

From (6.55) this function satisfies $\dot{\mathbf{x}} = \mathbf{Ax}$ and $\mathbf{x}(0) = c_1 \mathbf{e}_1 + \cdots + c_n \mathbf{e}_n = \mathbf{x}_0$, thus $\mathbf{x}(t)$ provides the solution we seek. From the definition of matrix-vector multiplication the quantity on the right in (6.56) can be written more succinctly as $\mathbf{X}\mathbf{x}_0$ or $e^{t\mathbf{A}}\mathbf{x}_0$.

We have shown

Theorem 6.4.1 The solution to $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ with initial data $\mathbf{x}_0 = \mathbf{x}_0$ is given by

$$\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}_0.$$

■ **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 given by

$$e^{t\mathbf{A}} = \begin{bmatrix} -3 + 4e^{-t} & 2 - 2e^{-t} \\ 6e^{-t} - 6 & 4 - 3e^{-t} \end{bmatrix}.$$

We can use this to solve the ODE system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ 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 $x_1 = u, x_2 = u'$ this is equivalent to the system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ with

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -4 & -4 \end{bmatrix}.$$

This matrix is defective, with double eigenvalue $\lambda = -2$ and eigenvector $\mathbf{v} = \langle -1, 2 \rangle$. As a result, special procedures were needed 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 this is computed when we consider Putzer's Algorithm. The solution to $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ with, for example, $\mathbf{x}(0) = \langle 1, -5 \rangle$ is

$$\mathbf{x}(t) = e^{t\mathbf{A}} \begin{bmatrix} 1 \\ -5 \end{bmatrix} = \begin{bmatrix} (2t+1)e^{2t} & te^{-2t} \\ -4te^{-2t} & (-2t+1)e^{-2t} \end{bmatrix} \begin{bmatrix} 1 \\ -5 \end{bmatrix} = \begin{bmatrix} (-3t+1)e^{-2t} \\ (6t-5)e^{-2t} \end{bmatrix}.$$

■

6.4.6 Computing The Matrix Exponential: The Diagonal Case

There is one case in which computing the matrix exponential using (6.49) 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

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & \lambda_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix}. \quad (6.57)$$

You can easily check that any power \mathbf{D}^k is given by

$$\mathbf{D}^k = \begin{bmatrix} \lambda_1^k & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2^k & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & d_{n-1}^k & 0 \\ 0 & 0 & \cdots & 0 & d_n^k \end{bmatrix}.$$

In this case we find that when we substitute \mathbf{D} into the Taylor series (6.49) 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} , $1 \leq m \leq n$. Thus

$$e^{\mathbf{D}} = \begin{bmatrix} e^{\lambda_1} & 0 & \cdots & 0 & 0 \\ 0 & e^{\lambda_2} & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & e^{\lambda_{n-1}} & 0 \\ 0 & 0 & \cdots & 0 & e^{\lambda_n} \end{bmatrix}. \quad (6.58)$$

Of course if we want to compute $e^{t\mathbf{D}}$ we simply note that $t\mathbf{D}$ is a diagonal matrix with m th diagonal entry $\lambda_m t$, so from (6.58) $e^{t\mathbf{D}}$ is the diagonal matrix with m th diagonal entry $e^{\lambda_m t}$.

Reading Exercise 165 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 (6.58) (with λ_m replaced by $\lambda_m t$) to solve the system with initial conditions $x_1(0) = 2, x_2(0) = 5$.

6.4.7 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 can then apply the method with $\mathbf{B} = t\mathbf{A}$ to solve systems of ODEs. We will assume that the matrix \mathbf{B} diagonalizable as detailed in Appendix B. Specifically, suppose \mathbf{B} has eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$, allowing repeated and/or complex eigenvalues. We suppose it is possible to choose a set $S = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ of eigenvectors (\mathbf{v}_k with 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 (6.57), ordered so λ_k is the eigenvalue for eigenvector \mathbf{v}_k . Under these assumptions the matrix \mathbf{B} can be written in the form

$$\mathbf{B} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1} \quad (6.59)$$

or *diagonalized*, as described in Appendix B. Moreover, as shown in Appendix B,

$$\mathbf{B}^k = \mathbf{P}\mathbf{D}^k\mathbf{P}^{-1}. \quad (6.60)$$

By using (6.60) 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 $m \rightarrow \infty$ the i th diagonal sum approaches e^{λ_i} . We conclude that

$$e^{\mathbf{B}} = \mathbf{P}e^{\mathbf{D}}\mathbf{P}^{-1} \quad (6.61)$$

where $e^{\mathbf{D}}$ is the diagonal matrix with i th diagonal entry e^{λ_i} . We should note that if \mathbf{v}_k is an eigenvector for \mathbf{B} with eigenvalue λ_k then any nonzero multiple of \mathbf{v}_k is also an eigenvector. This changes the matrix \mathbf{P} , of course, but \mathbf{P}^{-1} is also altered and (6.61) will remain valid.

In the case in which we want to compute $e^{t\mathbf{A}}$, simply substitute $\mathbf{B} = t\mathbf{A}$ in (6.61) and note that $t\mathbf{A}$ is just t times \mathbf{A} and so has the same eigenvectors as \mathbf{A} but with eigenvalues $\lambda_k t$. Then

$$e^{t\mathbf{A}} = \mathbf{P}e^{t\mathbf{D}}\mathbf{P}^{-1}. \quad (6.62)$$

■ **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 $x_1(0) = 1, x_2(0) = 2$. In matrix terms we have $\dot{\mathbf{x}} = \mathbf{Ax}$ 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 (6.59) 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 166 Verify that $e^{t\mathbf{A}}$ in Example 6.16 is correct, by diagonalizing \mathbf{A} and using (6.62).

Reading Exercise 167 What goes wrong with (6.62) when you try to compute $e^{t\mathbf{A}}$ for the matrix of Example 6.18?

6.4.8 Computing The Matrix Exponential: Putzer's Algorithm

Reading Exercise 167 illustrates what can go wrong with using diagonalization to compute the matrix exponential: not all matrices are diagonalizable. Nonetheless, the series (6.49) converges for any matrix, and so all matrices can be exponentiated. What do we do if 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 [85].

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

$$e^{t\mathbf{A}} = \sum_{j=0}^{n-1} r_{j+1}(t) \mathbf{P}_j \quad (6.63)$$

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

$$\mathbf{P}_j = \prod_{k=1}^j (\mathbf{A} - \lambda_k \mathbf{I})$$

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

$$\dot{\mathbf{r}}(t) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 & 0 \\ 1 & \lambda_2 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \lambda_{n-1} & 0 \\ 0 & 0 & \cdots & 1 & \lambda_n \end{bmatrix} \mathbf{r}(t) \quad (6.64)$$

with initial condition $\mathbf{r}(0) = \langle 1, 0, 0, \dots, 0 \rangle$.

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 $\lambda_1 = -1, \lambda_2 = -2$. We find that

$$\mathbf{P}_0 = \mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mathbf{P}_1 = \prod_{k=1}^1 (\mathbf{A} - \lambda_k \mathbf{I}) = (\mathbf{A} + \mathbf{I}) = \begin{bmatrix} 8 & -6 \\ 12 & -9 \end{bmatrix}.$$

Equation (6.64) 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 $r_1(0) = 1, r_2(0) = 0$. 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 $r_2(0) = 0$. This is a scalar constant coefficient linear ODE, easy to solve via the integrating factor approach. We find $r_2(t) = e^{-t} - e^{-2t}$. From (6.63) 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 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 was

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -4 & -4 \end{bmatrix}$$

with eigenvalues $\lambda_1 = \lambda_2 = -2$ (defective, since there is only a single eigenvector). Then

$$\begin{aligned} \mathbf{P}_0 &= \mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ \mathbf{P}_1 &= \prod_{k=1}^1 (\mathbf{A} - \lambda_k \mathbf{I}) = (\mathbf{A} - \mathbf{I}) = \begin{bmatrix} 2 & 1 \\ -4 & -2 \end{bmatrix}. \end{aligned}$$

Equation (6.64) 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 $r_1(0) = 1, r_2(0) = 0$. 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 $r_2(0) = 0$. This is a scalar constant coefficient linear ODE, easy to solve via the integrating factor approach. We find $r_2(t) = te^{-2t}$. From (6.63) 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 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, but Putzer's algorithm will. With $n = 3$ we find

$$\begin{aligned} \mathbf{P}_0 &= \mathbf{I} \\ \mathbf{P}_1 &= (\mathbf{A} + \mathbf{I}) = \begin{bmatrix} 3 & 0 & 1 \\ -3 & 0 & 2 \\ 0 & 0 & 3 \end{bmatrix} \\ \mathbf{P}_2 &= (\mathbf{A} + \mathbf{I})(\mathbf{A} - 2\mathbf{I}) = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & -3 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Equation (6.64) 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 $r_1(0) = 1$, $r_2(0) = 0$, $r_3(0) = 0$. 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, easy to solve via 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 via the integrating factor approach. We find $r_3(t) = (3t - 1)e^{2t}/9 + e^{-t}/9$. From (6.63) we have

$$\begin{aligned} e^{t\mathbf{A}} &= \sum_{j=0}^2 r_{j+1}(t)\mathbf{P}_j \\ &= r_1(t)\mathbf{P}_0 + r_2(t)\mathbf{P}_1 + r_3(t)\mathbf{P}_2 \\ &= e^{-t} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \left(\frac{e^{2t} - e^{-t}}{3} \right) \begin{bmatrix} 3 & 0 & 1 \\ -3 & 0 & 2 \\ 0 & 0 & 3 \end{bmatrix} + \left(\frac{(3t - 1)e^{2t} + e^{-t}}{9} \right) \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & -3 \\ 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} e^{2t} & 0 & te^{2t} \\ e^{-t} - e^{2t} & e^{-t} & e^{2t} - e^{-t} + te^{-2t} \\ 0 & 0 & e^{2t} \end{bmatrix}. \end{aligned}$$

In each example above notice how 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 $r_3(t)$, 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.

6.4.9 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 $a > 0$ then x_0 is multiplied

by a factor e^{at} that grows in time, while $a < 0$ means x_0 is diminished 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 fixed 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.

6.4.10 Exercises

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 this to solve $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ with $x_1(0) = 1, x_2(0) = 2$. ■

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 this to solve $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ with $x_1(0) = 4, x_2(0) = 2$. ■

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 this to solve $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ with $x_1(0) = 0, x_2(0) = -2$. ■

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 this to solve $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ with $x_1(0) = 1, x_2(0) = 2$. ■

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 this to solve $\dot{\mathbf{x}}(t) = \mathbf{Ax}(t)$ with $x_1(0) = 1, x_2(0) = 0, x_3(0) = -1$. ■

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 with $x_1(0) = 1, x_2(0) = 0, x_3(0) = -1$. Hint: the eigenvalues and eigenvectors here are simple, e.g., integer eigenvalues. ■

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 this to the case of n constant coefficient linear nonhomogeneous ODE's, of the form

$$\dot{\mathbf{x}}(t) = \mathbf{Ax}(t) + \mathbf{b}(t) \quad (6.65)$$

where $\mathbf{b}(t) = \langle b_1(t), \dots, b_n(t) \rangle$. In particular, show that

$$\mathbf{x}(t) = e^{t\mathbf{A}} \int_0^t e^{-s\mathbf{A}} \mathbf{b}(s) ds + e^{t\mathbf{A}} \mathbf{x}_0 \quad (6.66)$$

provides a solution to (6.65) with $\mathbf{x}(0) = \mathbf{x}_0$. The integral in (6.66) is to be done component-by-component.

Then use (6.66) 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 $x_1(0) = 1, x_2(0) = 2$. ■

6.5 Modeling Projects

6.5.1 Project: LSD Compartment Model

You should begin by re-reading the two-compartment model for LSD metabolism in Section 6.1. The data in Table 6.1 is provided on the book web site [6].

The model (6.1)-(6.2) we presented in Section 6.1 was based upon 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 [74, 93] formulate the 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

$$V_P \dot{c}_P(t) = -k_b V_P c_P(t) - k_e V_P c_P(t) + k_a V_T c_T(t) \quad (6.67)$$

$$V_T \dot{c}_T(t) = k_b V_P c_P(t) - k_a V_T c_T(t) \quad (6.68)$$

where we take $g(t) = 0$ (after the initial dose, no additional LSD is administered).

It is also estimated in [74, 93] 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 masses. Then (6.67)-(6.68) become

$$0.163M \dot{c}_P(t) = -0.163M k_b c_P(t) - 0.163M k_e c_P(t) + 0.115M k_a c_T(t) \quad (6.69)$$

$$0.115M \dot{c}_T(t) = 0.163M k_b c_P(t) - 0.115M k_a c_T(t). \quad (6.70)$$

Note that in (6.69)-(6.70) 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 1 Show that (6.69)-(6.70) can be reformulated as

$$\dot{c}_P(t) = -k_b c_P(t) - k_e c_P(t) + 0.706 k_a c_T(t) \quad (6.71)$$

$$\dot{c}_T(t) = 1.407 k_b c_P(t) - k_a c_T(t). \quad (6.72)$$

(note M drops out) with initial conditions $c_P(0) = 12.27$ mg per liter and $c_T(0) = 0$ mg per liter. Why is this initial condition for c_T appropriate?

Equations (6.71)-(6.72) have a unique solution for any given 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 those parameter values that give the best fit to the data, in a least-squares sense.

Modeling Exercise 2 Use a computer algebra system to solve (6.71)-(6.72) symbolically in terms of t , with parameters k_a, k_b, k_e (it will be quite messy).

Modeling Exercise 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 subject at time t_k . By the way, 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, e.g., Maple's `Minimize` command, or Mathematica's `FindMinimum` command. It may be helpful to specify that the parameters k_a, k_b, 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 performance? 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 [92], 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 [71]. The model yields a linear system of ODE's, so we can readily calculate solutions and determine long-term trends, even without using technology. In the next chapter 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 [80]. 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. in 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” [38]. 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 certain city has 118,000 non-homeowner households and 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

$$\begin{aligned} r(t) &= R(t)/N \text{ is the fraction of } \textit{renting} \text{ households at time } t, \\ e(t) &= E(t)/N \text{ is the fraction of } \textit{evicted} \text{ households at time } t. \end{aligned}$$

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.

Modeling Exercise 1 Find equations for $\frac{dr}{dt}$ and $\frac{de}{dt}$ that satisfy the above assumptions, under the assumption that t is measured in years. Explain the meanings of $\frac{dr}{dt}$ and $\frac{de}{dt}$ and what each component of your equations represents.

Modeling Exercise 2 Formula 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 3 Compute the eigenvalues and eigenvectors for \mathbf{A} , and use them to write out the general solution to $\mathbf{u} = \langle r(t), e(t) \rangle$.

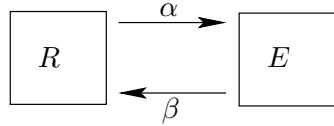


Figure 6.10: Flow diagram for eviction model.

Modeling Exercise 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 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, $r(0) = r_0, e(0) = 1 - r_0$) explicitly in terms of α, β , and r_0 .

What does the model predict as $t \rightarrow \infty$?

Modeling Exercise 5 Choose a value for α from the 2016 data for city evictions from the Eviction Lab [71], 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 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 [67, 68] and the article [29].

Tuned Mass Dampers

The Taipei 101 tower in Taiwan was completed in 2004, and was until 2010 the tallest skyscraper in the world. The tower is 509 meters tall, roughly 1670 feet. One of challenges in building such a tall and yet light-weight 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 simply 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, bridges, etc., was designed and built using rather conservative design processes that resulted in stiff, rigid 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 merely a small damped spring-mass system attached to a larger spring-mass system (e.g., the skyscraper). The tuned mass damper prevents the larger spring-mass system from vibrating excessively, especially near 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 first uses of TMD's in the United States for large structures were in the John Hancock Building in Boston in 1977 [47] and City Corp Center [77] in New York in 1978. Since that time many different styles, including active TMD's and pendulum TMD's, have been employed, while

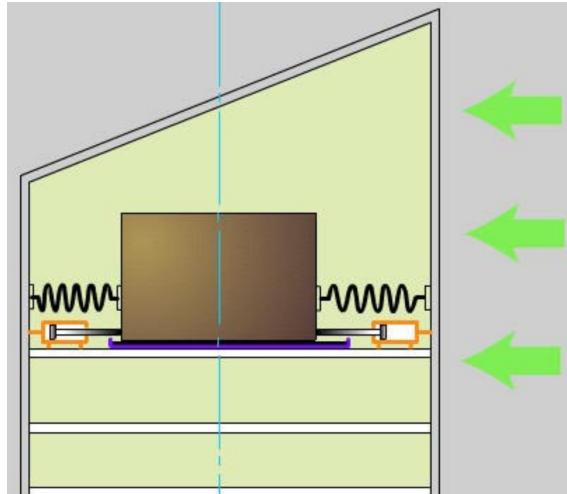


Figure 6.11: Schematic of the TMD atop the Citicorp building in New York City from [59].

diverse applications have been found through retro-fitting on 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 [13] 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 [14, 58]. As an example of the diversity of uses, TMDs are also used in the design of surgery tables to mitigate the vibrations of surroundings during surgery [76]—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 [79]. An excellent narrative, “What is a Tuned Mass Damper,” [42] with both technical, laboratory, and cultural elements is offered by the Practical Engineering Project at its YouTube site.

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 equation (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 fixed altitude, e.g., the top of the building. We thus have a model

$$m_1\ddot{x}_1 + c_1\dot{x}_1 + k_1x_1 = f(t)$$

where $f(t)$ is a driving force, perhaps the wind or a seismic disturbance, and m_1, c_1, 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 $t = t_0$. The variables are subscripted with a 1 because we are going to add a second mass.

When stimulated by a driving force, the building will sway. The goal is to control or damp out this swaying by including a second mass m_2 in the system, attached to the m_1 mass, but with $m_2 \ll m_1$. The situation in practice is illustrated in Figure 6.11, which depicts the tuned mass damper in the Citicorp building in New York City.

As suggested by Figure 6.11, we will model the TMD as a second mass m_2 , attached to the building mass m_1 , that can move horizontally. A simplified and abstracted version of the problem is illustrated in Figure 6.12.

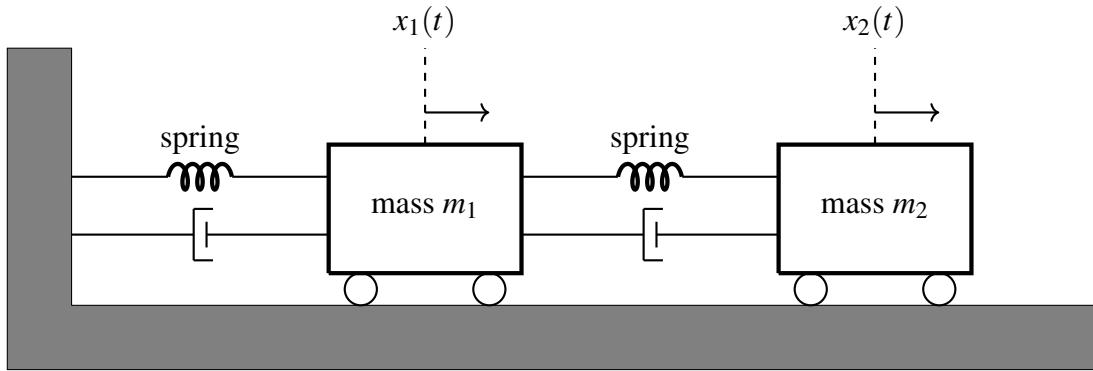


Figure 6.12: Tuned mass damper system.

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 the springs each obey 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 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

$$m_1\ddot{x}_1 = -k_1(x_1 - L_1) + k_2(x_2 - x_1 - L_2) - c_1\dot{x}_1 + c_2(\dot{x}_2 - \dot{x}_1) + f(t) \quad (6.73)$$

where \$f(t)\$ is the driving force on \$m_1\$. It may be helpful to review Reading Exercise 157 and the derivation of equations (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

$$m_2\ddot{x}_2 = -k_2(x_2 - x_1 - L_2) - c_2(\dot{x}_2 - \dot{x}_1). \quad (6.74)$$

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 (6.73)-(6.74) become

$$\begin{aligned} m_1\ddot{u}_1 &= -k_1u_1 + k_2(u_2 - u_1) - c_1\dot{u}_1 + c_2(\dot{u}_2 - \dot{u}_1) + f(t) \\ m_2\ddot{u}_2 &= -k_2(u_2 - u_1) - c_2(\dot{u}_2 - \dot{u}_1). \end{aligned} \quad (6.75)$$

Analysis of the Undamped Case

Modeling Exercise 2 Consider equations (6.75) 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. 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 (6.75) 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 (6.75) from part (a) with initial conditions $u_1(0) = 1, \dot{u}_1(0) = 0, u_2(0) = 0, \dot{u}_2(0) = 0$; you may wish to use a numerical solver. Plot $u_1(t)$ (the displacement of mass 1). Reconcile what you see with the answers to part (b).
- (d) Now take $f(t) = \cos(3t)$ in part (a) and solve the resulting forced system with zero initial data $u_1(0) = 0, \dot{u}_1(0) = 0, u_2(0) = 0, \dot{u}_2(0) = 0$. 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 3 We again consider equations (6.75) 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 fix driving force $f(t) = \cos(3t)$; perhaps this is excitation from the wind, or seismic in nature. 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 (6.75) from part (a) with these parameters and with initial conditions $u_1(0) = 0, \dot{u}_1(0) = 0, u_2(0) = 0, \dot{u}_2(0) = 0$, and plot $u_1(t)$ (the motion of mass 1). Compare what you see to the plot obtained in part (d) of Modeling Exercise 2.
- (b) Formulate the system in part (a) as a system of four first order ODE's $\dot{\mathbf{u}} = \mathbf{Au} + \mathbf{f}(t)$, 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 in part (d) of Modeling Exercise 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 plotting 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 4 Again consider equations (6.75) 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 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 Based on your analysis in Reading Exercises 2 to 4, offer a description of how to design a TMD to stop resonant phenomena in the case in which the system has no damping.

Some Analysis of the Damped Case

Modeling Exercise 6 Let us perform some analysis of the damped case. Take $m_1 = 10, k_1 = 90, c_1 = 3, m_2 = 0.1$ in what follows.

- (a) With $k_2 = 0, c_2 = 0$ and $f(t) = 0$ (so no TMD is present, and no forcing) solve (6.75) with zero initial data. 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 (6.75) 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 (6.75) 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 is the best choice 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 [68].

7. Nonlinear Systems of Differential Equations

This chapter is devoted to the analysis of autonomous systems of differential equations, 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 7.0.1 A system of 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

$$\mathbf{f}(\mathbf{x}) = \langle f_1(\mathbf{x}), \dots, f_n(\mathbf{x}) \rangle.\tag{7.2}$$

Autonomous systems of ODEs are important for at least two reasons: they provide good models for many real-world physical systems, and they are amenable to qualitative analysis using graphical and geometric techniques. We can use these methods to determine the nature of solutions, for example, long-term behavior. Do solutions approach equilibrium, grow without bound, cycle periodically? The ideas are a natural extension of those developed in Section 2.3 for autonomous scalar ODEs. We'll focus initially on applying these techniques to linear systems and the relation of these methods to the eigenvalues for the system, and then we'll proceed to the analysis of nonlinear systems. In each section we'll focus on systems of two ODEs for two unknown functions, but we will indicate how the ideas extend in some concrete and some conceptual ways to higher dimensions.

7.1 Autonomous Nonlinear Systems and Direction Fields

We've already encountered models involving linear systems of ODEs, for example, the LSD metabolism compartment model, double spring-mass models and tuned-mass-dampers, multiple-loop RLC circuits, even models for the homelessness problem. Let's now consider a few physical situations in which systems of nonlinear ODEs arise. These will provide motivation and illustration for the techniques we develop in this chapter.

7.1.1 Some Nonlinear ODE Models

The Struggle for Existence

This material is based on the SIMIODE project [105]. 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. The equation was

$$\dot{u}(t) = ru(t)(1 - u(t)/K), \quad (7.3)$$

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 equation (1.11). In Exercises 2.2.16 and/or 3.5.8 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, e.g., nutrients? How would each population grow? Would one dominate, perhaps drive the other to extinction? In the 1930s in the former Soviet Union the scientist G. F. Gause considered this question in the works [45] and [46], in service of improving vodka production.

The model adopted by Gause 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 and $u_2(t)$ the population of the second species, in this examples two different types of yeast. In this model we assume that each species would, in the absence of the other competing species, grow according to equation (7.3). 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 . However, when present in the same environment the species do interact and compete for resources, e.g., food or space. As a result, the presence of the second yeast species should have a negative impact on the population of the first species. One way to model this impact is by altering the logistic ODE (7.3) for $u(t) = u_1(t)$ as

$$\dot{u}_1(t) = r_1 u_1(t) \left(\frac{K_1 - u_1(t) - au_2(t)}{K_1} \right) \quad (7.4)$$

for some constant $a \geq 0$; 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

$$\dot{u}_2(t) = r_2 u_2(t) \left(\frac{K_2 - u_2(t) - bu_1(t)}{K_2} \right) \quad (7.5)$$

for some constant $b \geq 0$.

Reading Exercise 168 Justify the modifications of (7.3) that lead to (7.4) and (7.5). In particular, examine (7.4) when $u_2 = 0$, and the behavior of the right side of this ODE when u_2 increases. How does the parameter a affect things? Similar considerations apply to the ODE (7.5).

Equations (7.4)-(7.5) form an autonomous pair of coupled nonlinear ODEs for the functions $u_1(t), u_2(t)$, and this system typically has no elementary analytical solution. But with the techniques

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 .

of this chapter we will be able to make certain conclusions concerning the behavior of solutions. Most importantly, we can 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 7.4 we’ll consider some of the data collected by Gause and use this data to estimate the various parameters in (7.4)-(7.5).

Reading Exercise 169 Suppose that $u_1(t)$ and $u_2(t)$ in (7.4)-(7.5) are both constant functions, say $u_1(t) = p_1$ and $u_2(t) = p_2$. Show that each of the three pairs (p_1, p_2) given by

$$(0,0), \quad (K_1,0), \quad (0,K_2) \tag{7.6}$$

yield a constant solution $u_1(t) = p_1, u_2(t) = p_2$ to (7.4)-(7.5); note that in each case $\dot{u}_1 = \dot{u}_2 = 0$. What physical interpretation would you attach to each of these three solutions—what does it mean for the populations involved? As in the scalar case, constant solutions are called *fixed points* or *equilibrium solutions*.

Then verify that there is a fourth equilibrium solution

$$\left(\frac{K_1 - K_2 a}{1 - ab}, \frac{K_2 - K_1 b}{1 - ab} \right), \tag{7.7}$$

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.

Epidemic Models

This material is drawn from the SIMIODE project [75], which itself draws on material from [78].

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 were compiled from the Communicable Disease Surveillance Center [34, p. 587] and are given as an example by Murray in his book on mathematical biology [78]. The data values were extracted from the graph found in [87].

There were 763 boys at the English boarding school from which our data was obtained. We can see from Table 7.1 that the initial number of bedridden students is one. The data given is number of bedridden boys rather than 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 bedridden¹. 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

¹One 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.

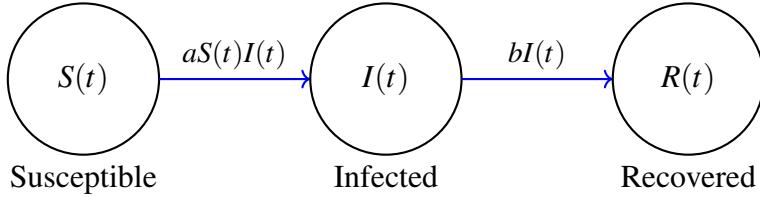


Figure 7.1: SIR compartmental model an epidemic.

various parameters that appear in the mode. Such a model can be used to examine various strategies for controlling the epidemic, e.g., isolating sick students for a period of time. One very common approach to modeling an epidemic is 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 $I(t)$ indicated the number of infected persons, so $S(t)$ and $R(t)$ will indicate the number of susceptible and recovered students, respectively.

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 S , if we double I 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, each such interaction carries a fixed risk of the susceptible person getting 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 to the I compartment in Figure 7.1, where the constant a depends on, for example, the rate at which people interact and the infectiousness of the disease. Movement from the I 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

$$\begin{aligned}\dot{S} &= -aSI \\ \dot{I} &= aSI - bI \\ \dot{R} &= bI.\end{aligned}\tag{7.8}$$

Reading Exercise 170 In light of the above assumptions, defend the model (7.8). What critiques of these assumptions can you make?

Reading Exercise 171 Compute $\dot{S} + \dot{I} + \dot{R}$ for equations (7.8), and explain why the result makes sense.

Equations (7.8) constitute a system of three coupled autonomous nonlinear ODEs for the functions $S(t)$, $I(t)$, and $R(t)$. We will analyze and improve upon 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 vertical as it swings obeys the nonlinear second order ODE

$$\ddot{\theta}(t) + c\dot{\theta}(t) + \frac{g}{L} \sin(\theta(t)) = 0 \quad (7.9)$$

where $c \geq 0$ is a frictional or damping constant. Equation (7.9) has no analytical solution. However, the behavior of solutions can be deduced using the techniques of this chapter. To proceed, we convert (7.9) 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

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{g}{L} \sin(x_1) - cx_2.\end{aligned}\quad (7.10)$$

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 172 If the angle $\theta(t)$ that the pendulum makes with respect to vertical remains close to zero then $x_1 \approx 0$ and the approximation $\sin(x_1) \approx x_1$ is reasonable. If we make this substitution in (7.10), show that the resulting linear system can be formulated as $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ with

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -g/L & -c \end{bmatrix}$$

and that \mathbf{A} has eigenvalues

$$-\frac{c}{2} \pm \sqrt{c^2/4 - g/L}.$$

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 notion of *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. Something similar can be done 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 that revolves around a spring-mass-damper system, since we have some intuition about the behavior of the 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

$$\begin{aligned}\dot{x}_1 &= \underbrace{x_2}_{f_1(x_1, x_2)} \\ \dot{x}_2 &= \underbrace{-x_1 - x_2}_{f_2(x_1, x_2)}\end{aligned}\quad (7.11)$$

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 the 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 (7.11) passes through a given point $(x_1, x_2) = (a_1, a_2)$ in the phase space at a time $t = t_0$ we can determine the past and future trajectory of the solution by solving (7.11) with initial condition $x_1(t_0) = a_1, x_2(t_0) = a_2$.

But even without solving, we can determine something about the solution to (7.11) that passes through the point $(x_1, x_2) = (a_1, a_2)$ at a time $t = t_0$. Let us write $\mathbf{x}(t) = \langle x_1(t), x_2(t) \rangle$ for a vector-valued description of a solution pair $x_1(t), x_2(t)$ to (7.11). Recall from multivariable calculus that the vector

$$\dot{\mathbf{x}}(t_0) = \langle \dot{x}_1(t_0), \dot{x}_2(t_0) \rangle$$

is tangent to the curve parameterized by $\mathbf{x}(t)$ at the point $\mathbf{x}(t_0)$, so $\dot{\mathbf{x}}(t_0)$ tells us what direction the solution is moving. We can compute this vector for a solution curve passing through $x_1 = a_1, x_2 = a_2$ at some time t_0 by using (7.11) as

$$\begin{aligned} \dot{\mathbf{x}}(t_0) &= \langle \dot{x}_1(t_0), \dot{x}_2(t_0) \rangle \\ &= \langle f_1(x_1(t_0), x_2(t_0)), f_2(x_1(t_0), x_2(t_0)) \rangle \\ &= \langle f_1(a_1, a_2), f_2(a_1, a_2) \rangle. \end{aligned} \tag{7.12}$$

Note that because the system is autonomous the functions f_1 and f_2 do not depend on t and so the vector $\dot{\mathbf{x}}(t_0)$ defined by (7.12) doesn't depend on t_0 , but only on the point (a_1, a_2) itself.

■ **Example 7.1** Suppose that a solution for the system (7.11) passes through the point $(1, 2)$ in the phase plane; this corresponds physically to a mass position $u = 1$, velocity $\dot{u} = 2$. From (7.12) 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 $\ddot{x}_2 = \ddot{u}$, the mass has acceleration -3 . ■

Reading Exercise 173 Use (7.12) to compute the vector $\dot{\mathbf{x}}$ for a solution that passes through the point $(-1, 1)$ in the phase plane. Use this to express the direction in which the solution is moving, as a unit vector. The interpret the situation physically; if $(x_1, x_2) = (-1, 1)$, what is the mass position and velocity? 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 $(1, 2)$, or the vector found in Reading Exercise 173 with its tail at $(-1, 1)$, to indicate the direction the solution to (7.11) is moving as it passes through the corresponding point. 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 shows these vectors scaled to $1/5$ 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, since $\dot{\mathbf{x}}(a_1, a_2)$ varies widely in magnitude. Since our primary interest is the actual 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 length 0.2. We will generally adopt the convention of scaling all vectors to the same length, determined by whatever gives the most insightful depiction of how solutions behave.

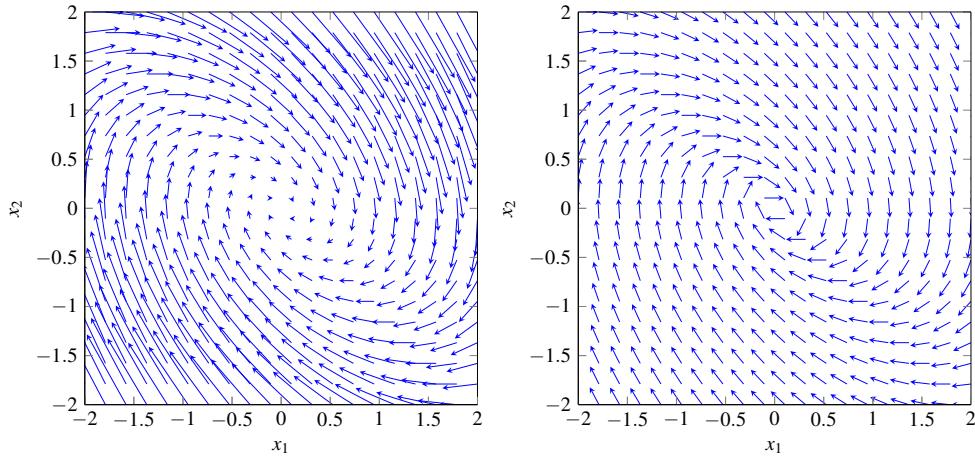


Figure 7.2: Left panel: Direction field for (7.11) with vectors given by (7.12), scaled to $1/5$ actual length. Right panel: Same, but with all vectors scaled to fixed length 0.2.

Recall Remark 11 in Section 6.1; a solution to the ODE system (7.11) that passes through a given 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 given point (a_1, a_2) , by starting at this point and following the arrows of the direction field. An Example is shown in Figure 7.3, 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 (7.11) that passes through the point $(x_1, x_2) = (2, 0)$.

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 given time, the curve does indicate the long-term behavior of the solution. Specifically, it is clear that the solution curve parameterized by $\mathbf{x}(t)$ starting at $(2, 0)$ spirals into the origin. It is also essential to note that this gives us 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 $x_2(t) = \dot{u}(t)$ is the velocity of the mass. 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 . At some point in time we pass though a point where $x_1 = u \approx -0.6 < 0$ and $x_2 = \dot{u} = 0$; the spring is in compression and the mass is momentarily stopped. After this time $x_1 = u$ begins to increase and $x_2 = \dot{u} > 0$, as the mass moves in the direction of increasing u , until coming to rest momentarily, 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 (7.11) or equivalently, $\ddot{u} + \dot{u} + u = 0$.

Reading Exercise 174 Solve the system $\dot{x}_1 = x_2$, $\dot{x}_2 = -x_1 - x_2$ with $x_1(0) = 2, x_2(0) = 0$ (or equivalently, $\ddot{u} + \dot{u} + u = 0$ with $u(0) = 2, \dot{u}(0) = 0$). 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.

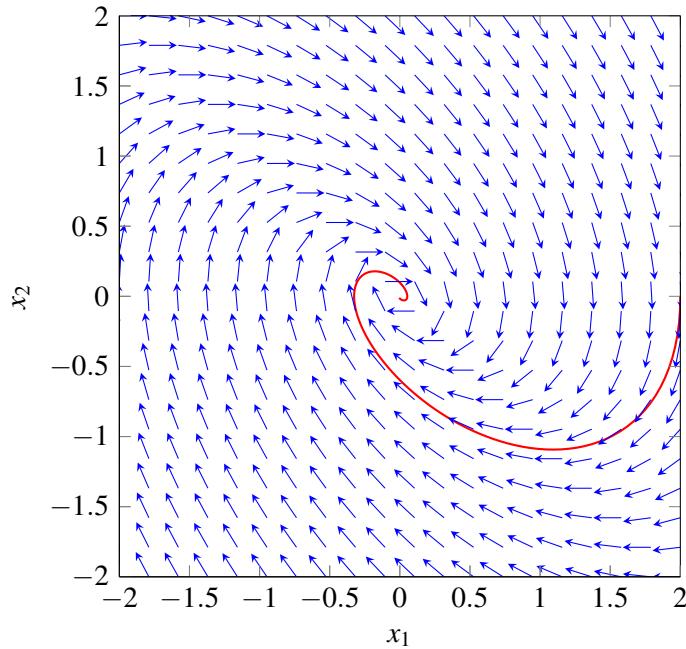


Figure 7.3: Direction field and a solution curve to (7.11); the solution passes through $(x_1, x_2) = (2, 0)$.

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 (7.4)-(7.5) in the specific case that $r_1 = 1, K_1 = 2, r_2 = 2, K_2 = 3, a = 0.2$, and $b = 0.45$. With these parameters the equations are

$$\begin{aligned}\dot{u}_1 &= \underbrace{u_1(1 - u_1/2) - 0.1u_1u_2}_{f_1(u_1, u_2)} \\ \dot{u}_2 &= \underbrace{2u_2(1 - u_2/3) - 0.3u_1u_2}_{f_2(u_1, u_2)}\end{aligned}\tag{7.13}$$

for the species populations $u_1(t), u_2(t)$, with f_1 and f_2 as indicated. Figure 7.4 shows the direction field for the system (7.13), in the range $0 \leq u_1, u_2 \leq 5$; 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 range, computing the vector $\langle \dot{u}_1, \dot{u}_2 \rangle = \langle f_1(u_1, u_2), f_2(u_1, u_2) \rangle$, and then plotting this vector with its tail at (u_1, u_2) . Note that this computation does not require us to solve the ODEs (7.13), which we can't do anyway; it merely requires arithmetic.

Based on the direction field we can sketch solutions with any given initial condition by following the arrows. Several solution curves are shown in red 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 $(u_1, u_2) \approx (1.6, 2.3)$, and based on the appearance of the direction field it seems this will happen for any starting populations. 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 175 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 $(u_1, u_2) \rightarrow (p_1, p_2)$ then $\dot{u}_1 \rightarrow 0$ and $\dot{u}_2 \rightarrow 0$, so that from (7.13) $p_1(1 - p_1/2) - 0.1p_1p_2 = 0$ and $2p_2(1 - p_2/3) - 0.3p_1p_2 = 0$. Find all points

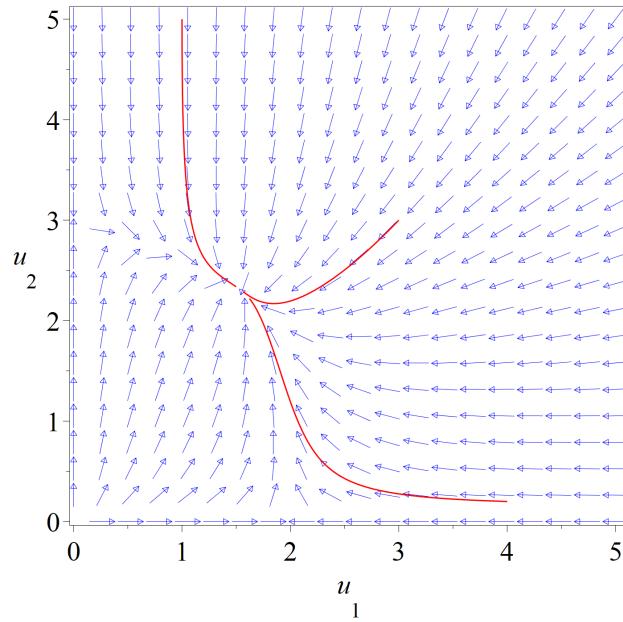


Figure 7.4: Direction field for the system (7.13), with several solution trajectories.

(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 169.

If we increase the competition parameters a and b from $a = 0.2, b = 0.45$ to $a = 2, b = 1.5$ 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 $(u_1, u_2) = (0, 3)$ in which 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, 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 various parameters in the ODE. In the competing species model we will be able determine what choices for the parameters allow 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 176 Repeat Reading Exercise 175 but with the parameters $a = 2, b = 1.5$ in (7.13).

7.1.4 Direction Fields in Higher Dimensions

For an autonomous system of n ODEs in n functions $x_1(t), \dots, x_n(t)$ of the form (7.1), the vector-valued function

$$\mathbf{f}(\mathbf{x}) = \langle f_1(\mathbf{x}), \dots, f_n(\mathbf{x}) \rangle$$

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 (7.12) does in \mathbb{R}^2 . The difficulty, of course, is that if $n > 3$ we can’t sketch or easily visualize this direction field. Indeed, even though many

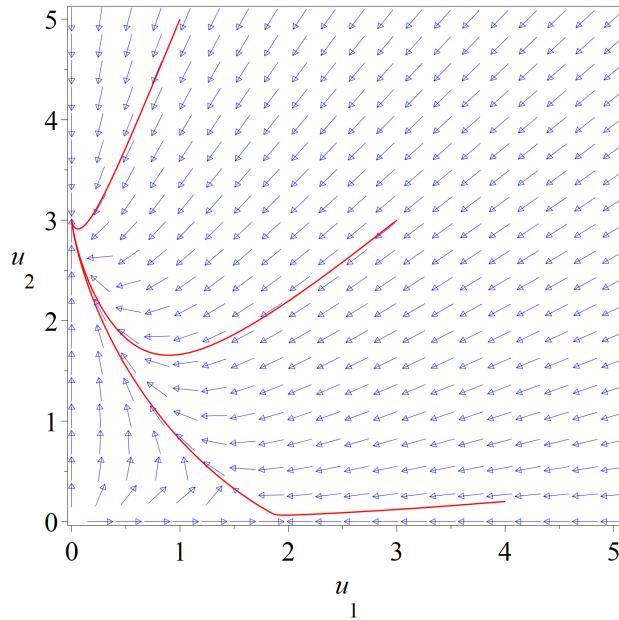


Figure 7.5: Direction field for the system (7.13) but with $a = 2, b = 1.5$, with several solutions.

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 of the ODE system dictate the direction 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 ODE's $\dot{x}_1 = f_1(x_1, x_2), \dot{x}_2 = f_2(x_1, x_2)$ 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 (7.12)).
 - 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.
- $f_1(x_1, x_2) = (x_1 + x_2)/2, f_2(x_1, x_2) = x_2/2$.
 - $f_1(x_1, x_2) = x_1^2/2 - x_2, f_2(x_1, x_2) = x_1 + 1$.
 - $f_1(x_1, x_2) = x_2 - 3/2, f_2(x_1, x_2) = -x_1 + 3/2$.
 - $f_1(x_1, x_2) = x_2, f_2(x_1, x_2) = x_1$.

Exercise 7.1.2 Use whatever technology you have to sketch a direction field for equations (7.4)-(7.5) with parameters values $r_1 = 1, r_2 = 1, K_1 = 3, K_2 = 3, a = 2$, and $b = 2$, on the range $0 \leq u_1, u_2 \leq 5$. Sketch a few representative solutions. What do all solutions converge to in this case? Does 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 equations (7.4)-(7.5) with parameters 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 (7.10) with $L = 1, c = 1, g = 9.8$ on the range $-2 \leq x_1, x_2 \leq 2$. Sketch a few solution trajectories. What do solutions do as $t \rightarrow \infty$? 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 (7.10) with $L = 1, c = 0, g = 9.8$ on the range $-2 \leq x_1, x_2 \leq 2$. Sketch a few solution trajectories. What do solutions do as $t \rightarrow \infty$? What does this say about the motion of the pendulum, and why does this make sense? ■

Exercise 7.1.6 The SIR epidemic model (7.8) might be considered a system of two ODE's $\dot{S} = -aSI$ 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 $t \rightarrow \infty$? What does this say about the epidemic, in particular, the number of susceptible and infected 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 $\dot{\mathbf{x}} = \mathbf{Ax}$, as well as $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{b}$ where \mathbf{b} is a (necessarily) constant vector. The focus is on how the eigenvalues of \mathbf{A} affect the direction field and what this says about the behavior of solutions. Our initial and primary interest is the case in which the system is two-dimensional and \mathbf{A} is invertible, but we will indicate how the conclusions extend to higher dimensions. The case in which \mathbf{A} is singular will then be summarized. We'll also 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 in the next section for analyzing nonlinear systems of ODEs.

7.2.1 Direction Fields for Homogeneous Linear Systems

Consider a linear constant coefficient system of n ODEs of the form $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{b}(t)$. If the system is autonomous, which we assume, then \mathbf{b} must a constant vector, and so the system can be expressed as

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{b}. \quad (7.14)$$

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, this means that all of the eigenvalues of \mathbf{A} are nonzero. Finally, 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 that 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 , or equivalently, 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 the general solution $\mathbf{x}_h(t)$ to the homogeneous system $\dot{\mathbf{x}} = \mathbf{Ax}$ is

$$\mathbf{x}_h(t) = c_1 e^{\lambda_1 t} \mathbf{v}_1 + c_2 e^{\lambda_2 t} \mathbf{v}_2. \quad (7.15)$$

Our primary interest is how these eigenvalues affect the direction field for the system.

There are three main cases to consider:

- If λ_1 and λ_2 are both negative then from (7.15) we see that all nontrivial solutions satisfy $\mathbf{x}_h(t) \rightarrow 0$ as $t \rightarrow \infty$ and $|\mathbf{x}_h(t)| \rightarrow \infty$ as $t \rightarrow -\infty$. This 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 in red; all solutions approach the origin. 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 (7.15) we see that all nontrivial solutions satisfy $|\mathbf{x}_h(t)| \rightarrow \infty$ as $t \rightarrow \infty$ and $\mathbf{x}_h(t) \rightarrow 0$ as $t \rightarrow -\infty$. This 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 in red; all solutions here radiate away from the equilibrium point $\langle 0, 0 \rangle$. In this case the origin $\langle 0, 0 \rangle$ is 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 $\mathbf{x}_h(t) \rightarrow c_2 e^{\lambda_2 t} \mathbf{v}_2$ as $t \rightarrow \infty$; that is, $\mathbf{x}_h(t)$ approaches a larger and larger multiple of the eigenvector \mathbf{v}_2 , as long as the initial conditions dictate $c_2 \neq 0$ (generically the case). But as $t \rightarrow -\infty$ 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*. The 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 $\mathbf{v}_1 = \langle 1, 1 \rangle$ (eigenvalues -1) and $\mathbf{v}_2 = \langle 1, -2 \rangle$ (eigenvalues 1); you can see them in the direction field behavior.

■ **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

$$\lambda_1 = \frac{-c + \sqrt{c^2 - 4mk}}{2m}, \quad \lambda_2 = \frac{-c - \sqrt{c^2 - 4mk}}{2m}. \quad (7.16)$$

It's not hard to see that $\lambda_2 < 0$ always, since $\sqrt{c^2 - 4mk}$ is positive. The eigenvalue λ_1 is also negative, because $c^2 - 4mk < c^2$ implies $\sqrt{c^2 - 4mk} < c$ and then $-c + \sqrt{c^2 - 4mk} < 0$. The numerator of λ_1 is thus negative and so is λ_1 since $2m > 0$.

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 $u(t) \rightarrow 0$ and $\dot{u}(t) \rightarrow 0$ as $t \rightarrow \infty$, for any initial data. 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 177 Suppose the matrix \mathbf{A} that governs $\dot{\mathbf{x}} = \mathbf{Ax}$ has two equal eigenvalues $\lambda_1 = \lambda_2 = \lambda$ and only one eigenvector \mathbf{v} , so the matrix \mathbf{A} is defective.

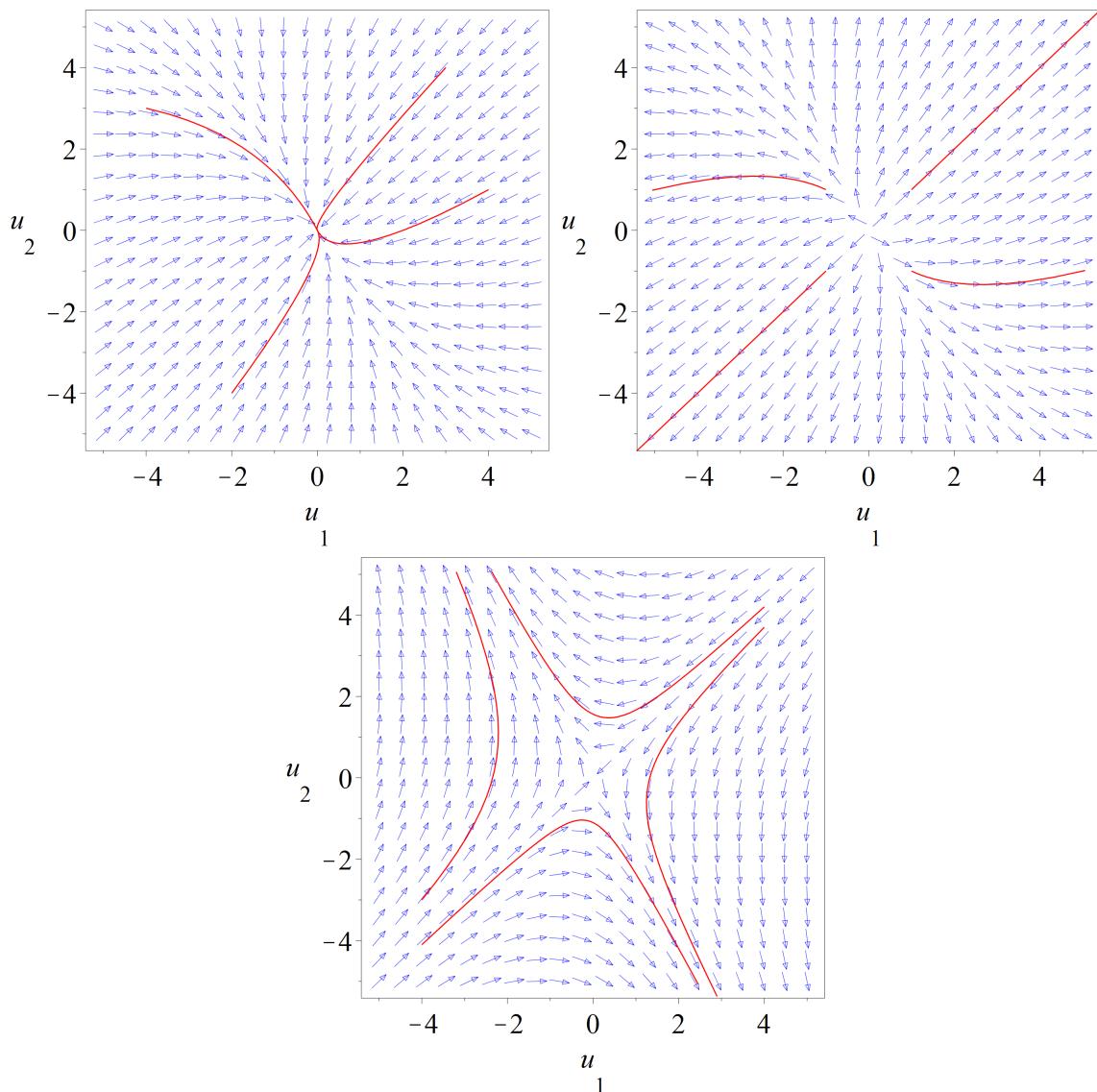


Figure 7.6: Direction field for a linear system with real nonzero eigenvalues. An asymptotically stable sink (top left), an unstable source (top right) and a saddle point (bottom) panel.

- (a) Use (6.33) to argue that all nontrivial solutions decay to the origin as $t \rightarrow \infty$ if $\lambda < 0$. In this case the origin is called an *asymptotically stable improper node*.
- (b) Use (6.33) to argue that all nontrivial solutions limit to infinity as $t \rightarrow \infty$ if $\lambda > 0$. In this case the origin is called an *unstable improper node*.

The Case in Which \mathbf{A} Is Invertible With Complex Eigenvalues

HERE

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 (7.15).

There are three cases to consider, depending on whether the real part of the eigenvalues for \mathbf{A} have negative, positive, or zero real parts.

- If the eigenvalues of the form $\lambda = -\alpha \pm \omega i$ with $\alpha > 0$ then the general solution (7.15) 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 (7.15) 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 an *asymptotically stable spiral point* or *spiral sink*.

- If the eigenvalues of the form $\lambda = -\alpha \pm \omega i$ with $\alpha > 0$ then the general solution (7.15) 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 (7.15) 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 an *unstable spiral point* or *spiral source*.

- If the eigenvalues are purely imaginary, of the form $\lambda = \pm \omega i$, then the general solution (7.15) can be written as

$$x_h(t) = c_1 \cos(\omega t) + c_2 \sin(\omega t).$$

In this case all solutions (7.15) 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 $c^2 - 4mk < 0$. We formulate this as a system $\dot{\mathbf{x}} = \mathbf{Ax}$ as in Example 7.2; the matrix \mathbf{A} and eigenvalues are unchanged, but now the eigenvalues in (7.16) are complex and can be written as

$$\lambda_1 = \frac{-c}{2m} + i \frac{d}{2m}, \quad \lambda_2 = \frac{-c}{2m} - i \frac{d}{2m}$$

where $d = \sqrt{4mk - c^2}$ is real, since $4mk - c^2 > 0$. If $c > 0$ 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. ■

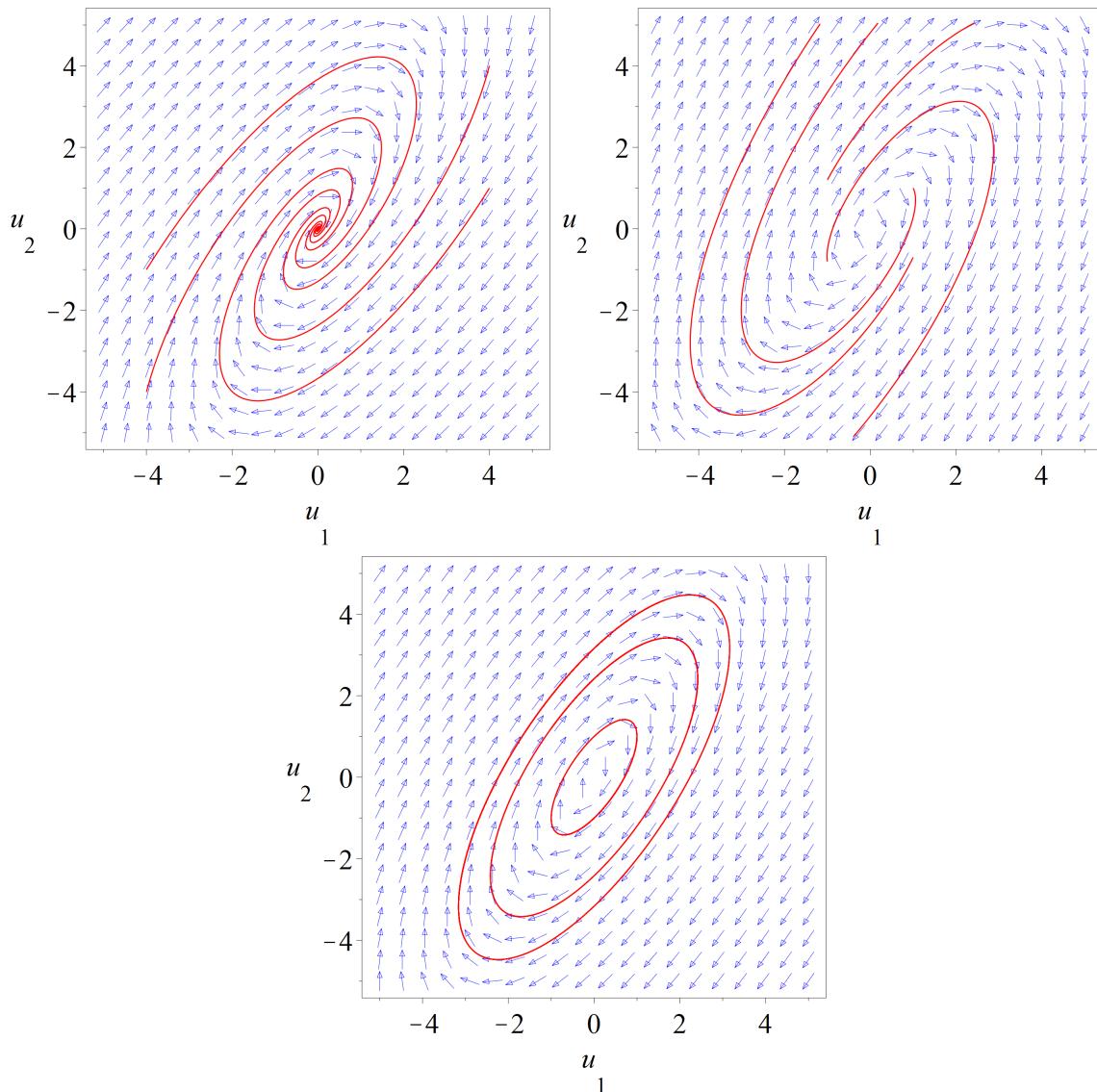


Figure 7.7: Direction field for a linear system with complex-conjugate eigenvalues with real parts negative (left panel), positive (right panel), and zero (bottom panel).

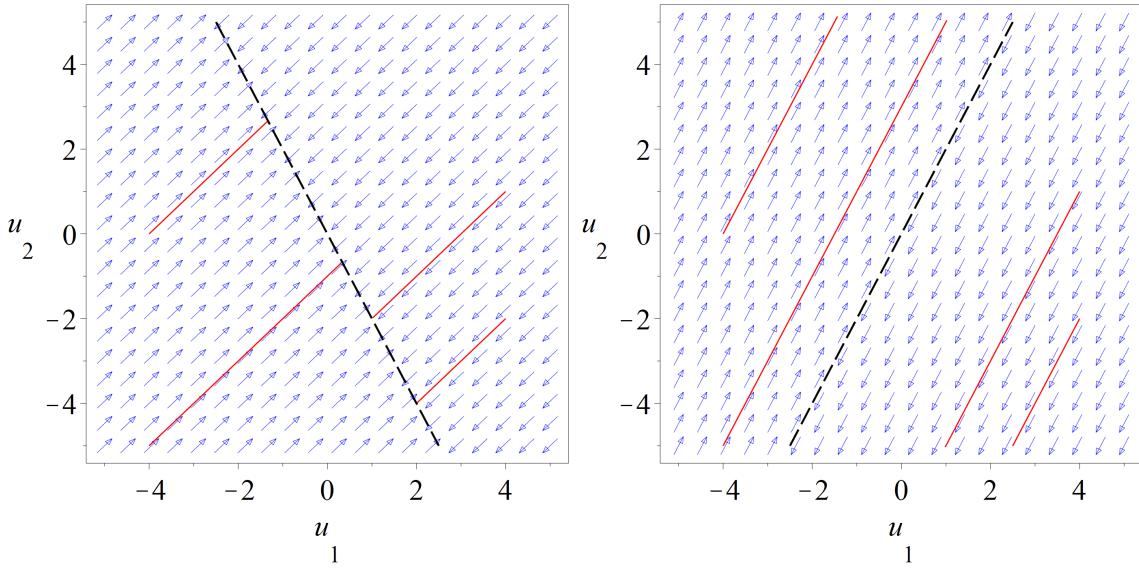


Figure 7.8: Direction field for a system with eigenvalues $0, \lambda$ with $\lambda < 0$ (left panel) and defective double eigenvalue 0 (right panel).

The Case in Which \mathbf{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 \mathbf{A} has at least one eigenvalue that equals zero. Suppose \mathbf{A} has exactly one zero eigenvalue, say $\lambda_1 = 0$ and $\lambda_2 = \lambda \neq 0$ (λ is necessarily real) with eigenvectors \mathbf{v}_1 and \mathbf{v}_2 , the general solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ is

$$\mathbf{x}_h(t) = c_1 \mathbf{v}_1 + c_2 e^{\lambda t} \mathbf{v}_2.$$

Every point on the line spanned by $c_1 \mathbf{v}_1$ is an equilibrium solution. If $\lambda < 0$ the all solutions decay to this line, while if $\lambda > 0$ all solutions radiate away from this line. The case $\lambda < 0$ is illustrated in Figure 7.8, in which the line spanned by the eigenvector \mathbf{v}_2 is shown as a dashed black line. The case $\lambda > 0$ is similar, just reverse the direction arrows.

If \mathbf{A} has only 0 as an (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 simply 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 (6.33), 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.

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 dictated assumptions that $m, k > 0$ and $c \geq 0$. This is true for many types of ODEs; we can make conclusions about the qualitative and long-term behavior of solutions without choosing specific values for the physical constants. 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 such a system without specifying these parameters. These techniques will be extended to nonlinear systems in Section 7.3.

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

$$\dot{x}_1 = -(k_b + k_e)x_1 + k_a x_2 \quad (7.17)$$

$$\dot{x}_2 = k_b x_1 - k_a x_2 \quad (7.18)$$

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 (7.17) and (7.18), $x_1(t)$ 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 . Note that given the physical nature of the problem we can confine our attention to the first quadrant $x_1, x_2 \geq 0$ in the phase plane.

Matrix Formulation and Eigenvalue Analysis

The system (6.1)-(6.2) can be formulated as $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ where $\mathbf{x} = \langle x_1, x_2 \rangle$ and

$$\mathbf{A} = \begin{bmatrix} -(k_b + k_e) & k_a \\ k_b & -k_a \end{bmatrix}. \quad (7.19)$$

It's easy to check that the matrix \mathbf{A} is invertible (for example, it has determinant $k_a k_b > 0$), so that the origin $\mathbf{x} = \mathbf{0}$ is the only equilibrium solution for this system.

The stability of the origin is dictated by the eigenvalues for \mathbf{A} . Despite the simplicity of \mathbf{A} , these eigenvalues are rather complicated, and determining their nature requires a bit of messy algebra for which a computer algebra system isn't terribly helpful. In Exercise 7.2.5 we provide guidance in showing that these eigenvalues are always real, negative, and distinct, for any positive choices of k_a, k_b , and k_e . This kind of task is fairly common in analyzing eigenvalues. There is no magic recipe; it requires perseverance, experimentation, and a bit of inspiration.

However, based on the fact that the eigenvalues are distinct negative real numbers 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 with the easy linear case, by illustrating these graphical methods on (7.17)-(7.18).

The Nullclines

First, consider equation (7.17) and suppose we are at a point in the $x_1 x_2$ 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 vertically-oriented. 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 on a vertical trajectory (unless $\dot{x}_2 = 0$ too, in which case we are at an equilibrium point). Physically, this corresponds to a point in the $x_1 x_2$ phase plane where the amount of LSD in the subject's plasma is (momentarily) not changing. From (7.17) the set of (x_1, x_2) points in the phase plane where $\dot{x}_1 = 0$ is described by the equation

$$-(k_b + k_e)x_1 + k_a x_2 = 0 \quad (7.20)$$

or equivalently, $x_2 = \frac{k_b + k_e}{k_a}x_1$. This curve is called the \dot{x}_1 -nullcline for this ODE system.

In this example the \dot{x}_1 -nullcline is a line through $(0,0)$ with positive slope $(k_b + k_e)/k_a$, illustrated in red in the left panel of Figure 7.9. The short red vertical tick marks 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 178 Show that if $x_2 > \frac{k_b+k_e}{k_a}x_1$ (we're above the \dot{x}_1 -nullcline in red 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 (7.17) 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 \dot{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.

Note that we can determine the \dot{x}_1 -nullcline and how solutions behavior on either side of it without specific values for the constants k_a, k_b , and k_e ; all we need to know is that these constants are positive.

The same analysis can be performed for equation (7.18). Specifically, $\dot{x}_2 = 0$ at those points in the phase plane where

$$k_b x_1 - k_a x_2 = 0 \quad (7.21)$$

or equivalently, $x_2 = \frac{k_b}{k_a}x_1$. Equation (7.21) describes the \dot{x}_2 -nullcline for this system, and this nullcline is a line through the origin with slope k_b/k_a , illustrated as the blue line in the left panel of Figure 7.9. The short blue horizontal tick marks indicate that the direction field vectors with tail on this nullcline are horizontal, and solution curves crossing the \dot{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 \dot{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 \dot{x}_2 -nullcline has a slope less than that of the \dot{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 left panel of Figure 7.9.

Sketching the Phase Portrait

Based on the above, we can make a few conclusions about the direction field for (7.17)-(7.18) 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 \dot{x}_1 -nullcline (given by (7.20)) satisfy $\dot{x}_1 = 0$ at that point, and so are vertically oriented. Solution curves that cross the \dot{x}_1 -nullcline do so vertically.
2. Direction field arrows based at points not on the \dot{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 say nothing about their motion in the x_2 direction, though.
3. The direction field arrows with tail on the \dot{x}_2 -nullcline (given by (7.21)) satisfy $\dot{x}_2 = 0$ at that point, and so are horizontally oriented. Solution curves that cross the \dot{x}_2 -nullcline do so horizontally.
4. Direction field arrows based at points not on the \dot{x}_2 -nullcline satisfy $\dot{x}_2 > 0$ or $\dot{x}_2 < 0$. As a result solutions passing through such points are moving generally to the up ($\dot{x}_2 > 0$) or down ($\dot{x}_2 < 0$). This say 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, the \dot{x}_1 -nullcline is shown in red and the \dot{x}_2 -nullcline in blue, with vertical or horizontal tick marks to indicate the

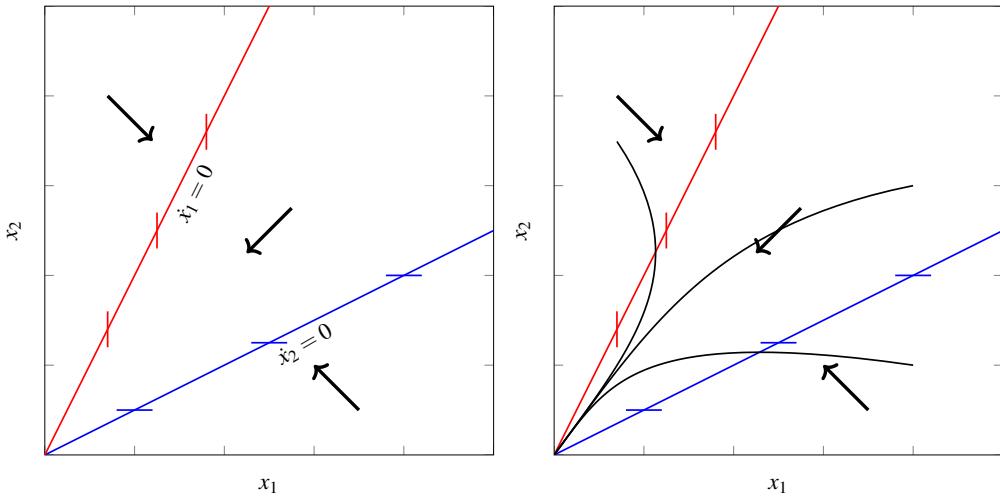


Figure 7.9: Left panel: Qualitative sketch of direction field for system (7.20)-(7.21). Right panel: Some representative solution curves superimposed (black), all converging to the origin.

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 black arrows in each region, to provide a rough indication of whether solutions are moving generally up/right, up/left, down/right, or down/left. The left panel of Figure 7.9 is a supremely economical direction field, that shows which direction solution curves move in the phase plane, in this case using only three arrows.

Sketching Solution Curves

Based on the left panel of Figure 7.9 we can sketch plausible solution trajectories, using 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, it is the inescapable conclusion is that all solutions converge to $(x_1, x_2) = (0, 0)$, or $(u_P, u_T) = (0, 0)$ in the original dependent variables. 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, in which we indicate the behavior of solutions to a system of ODEs by drawing direction arrows (whether manually or with a computer) and show typical solution trajectories, is called a *phase portrait* for the ODE system. 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 this analysis 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. In this case 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

$$\mathbf{x}_p = -\mathbf{A}^{-1}\mathbf{b}. \quad (7.22)$$

The general solution to $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}$ is then as given in (6.40) and is $\mathbf{x}(t) = \mathbf{x}_p + \mathbf{x}_h(t)$ or

$$\mathbf{x}(t) = -\mathbf{A}^{-1}\mathbf{b} + \mathbf{x}_h(t) \quad (7.23)$$

where $\mathbf{x}_h(t)$ is the general solution to the homogeneous system $\dot{\mathbf{x}} = \mathbf{Ax}$. The analysis for the homogeneous case based on (7.15), in conjunction with (7.23), shows that the direction field for the nonhomogeneous system $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{b}$ 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 (7.22). 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, the vector

$$\mathbf{f}(x_1, \dots, x_n) = \begin{bmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_n(x_1, \dots, x_n) \end{bmatrix}$$

dictates the direction a solution to (7.1) moves at any given point (x_1, \dots, x_n) in n -dimensional phase space, since $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$. In the linear homogeneous case this 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 \mathbf{A} has a full set of eigenvectors that span \mathbb{R}^n (that is, \mathbf{A} is diagonalizable) then a general solution $\mathbf{x}_h(t)$ to the homogeneous system $\dot{\mathbf{x}} = \mathbf{Ax}$ was derived in (6.19) and is given by

$$\mathbf{x}_h(t) = \sum_{k=1}^n c_k e^{\lambda_k t} \mathbf{v}_k \quad (7.24)$$

where the \mathbf{v}_k are the eigenvectors for \mathbf{A} and λ_k 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 $t \rightarrow \infty$.
- If any λ_k has positive real part then typical solutions grow without bound as $t \rightarrow \infty$, unless the corresponding c_k just happens to be zero.

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 [48] for more information.

7.2.5 Exercises

Exercise 7.2.1 Each pair of ODE's below is of the form

$$\dot{x}_1 = A_{1,1}x_1 + A_{1,2}x_2$$

$$\dot{x}_2 = A_{2,1}x_2 + A_{2,2}x_2$$

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 by computing the eigenvalues.

(a) $\mathbf{A} = \begin{bmatrix} 1 & -1 \\ 6 & -4 \end{bmatrix}$

(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) $\mathbf{A} = \begin{bmatrix} 3 & -2 \\ 4 & -1 \end{bmatrix}$

(g) $\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 6 & -3 \end{bmatrix}$

(h) $\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 2 & -1 \end{bmatrix}$

(i) $\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix}$

Exercise 7.2.2 Apply the technique used on (7.17)-(7.18) to sketch solution curves 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

- 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/left, up/right, down/left, or down/right and then sketch an appropriate arrow.
- Based on your picture, sketch solution curves with the given initial conditions; make sure to follow the arrows, and cross the \dot{x} -nullcline vertically, the \dot{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 of solution behavior to the eigenvalues of the matrix \mathbf{A} .

(a) $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 12 & -1 \end{bmatrix}$, initial conditions $x(0) = 2, y(0) = -2$ and $x(0) = 1, y(0) = -4$.

(b) $\mathbf{A} = \begin{bmatrix} 1 & -1 \\ 6 & -4 \end{bmatrix}$, initial conditions $x(0) = 2, y(0) = -2$ and $x(0) = -2, y(0) = 3$.

(c) $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$, initial conditions $x(0) = 1, y(0) = -1$ and $x(0) = 1, y(0) = -2$.

(d) $\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 4 & -3 \end{bmatrix}$, initial conditions $x(0) = 1, y(0) = -1$ and $x(0) = 1, y(0) = -2$.

(e) $\mathbf{A} = \begin{bmatrix} 3 & -2 \\ 4 & -1 \end{bmatrix}$, 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 $\dot{x} = y, \dot{y} = -kx/m - cy/m$ by letting $y = \dot{x}$.
- Show that the \dot{x} -nullcline is the x -axis in the phase plane, and the \dot{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 $k/c = 1$ here. The nullclines should divide the plane into 4 regions; sketch direction arrows in each (up/left, up/right, down/left, down/right).

- (d) Use this to sketch some solution trajectories. Does your sketch conform to how a damped spring-mass system should behave?
- (e) Repeat parts (c)-(d) assuming c is close to zero, so k/c is large. Does your sketch make sense?
- (f) Repeat parts (c)-(d) assuming c is very large, so k/c is close to zero. Does your sketch make sense?
- (g) Repeat parts (c)-(d) assuming $c = 0$ (so the ODE is $m\ddot{x} + kx = 0$). What does the \dot{y} -nullcline become? Does your sketch and solution curves reflect how an undamped system should behave?

Exercise 7.2.4 In Exercise 6.1.7 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.

- (a) Show that the \dot{x}_1 -nullcline here is the line $x_2 = -50 + 5x_1/6$ in the x_1x_2 phase plane. Also show that the \dot{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 $0 \leq x_1, x_2 \leq 200$. (Note that only $x_1, x_2 \geq 0$ is physically relevant here.)
- (b) Find the point where the nullclines intersect, and show that this is an equilibrium solution for the system.
- (c) The nullclines should divide the first quadrant into four regions. Sketch direction arrows in each (up/left, up/right, down/left, down/right), then sketch some plausible solution trajectories. What do solutions seem to do as $t \rightarrow \infty$?
- (d) 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 (7.17)-(7.18)) has distinct negative eigenvalues for any choice of $k_a, k_b, k_e > 0$, as strongly suggested by Figure 7.9, by following these steps:

- (a) Show that the eigenvalues of \mathbf{A} can be expressed as

$$\lambda_1 = \frac{-D - \sqrt{D^2 - 4k_a k_e}}{2} \quad \text{and} \quad \lambda_2 = \frac{-D + \sqrt{D^2 - 4k_a k_e}}{2}$$

where $D = k_a + k_b + k_e$. Note this implies $D > 0$.

- (b) Show that $D^2 - 4k_a k_e = (k_a + k_b - k_e)^2 + 4k_b k_e$. Why does this imply that $D^2 - 4k_a k_e > 0$?

- (c) Use parts (a) and (b) to argue that both λ_1 and λ_2 are real, distinct, and negative. It's easy to see that $\lambda_1 < 0$. Why is $\lambda_2 < 0$? (You may find it helpful to look back at Example 7.2.) ■

7.3 Autonomous Nonlinear Systems and Phase Portraits

The method of finding nullclines and deciding on solution directions, as we did to sketch solutions to (7.17)-(7.18), can also be used to sketch phase portraits for other systems of ODEs, even nonlinear systems. In this section we will refine and extend these techniques, and use them to determine the behavior of solutions to the competing species model (7.4)-(7.5), to illustrate the power of this approach for analyzing system of ODEs using little more than high-school algebra. Although the system (7.4)-(7.5) is two-dimensional, some concepts extend to higher dimensions and will be discussed.

7.3.1 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 (7.4)-7.5) of ODEs. Let us revisit the specific case in which the parameters were given as $r_1 = 1, K_1 = 2, r_2 = 2, K_2 = 3, a = 0.2$, 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

$$f_1(u_1, u_2) = u_1(1 - u_1/2) - 0.1u_1u_2 \quad (7.25)$$

$$f_2(u_1, u_2) = 2u_2(1 - u_2/3) - 0.3u_1u_2 \quad (7.26)$$

(these were given in (7.13).) A computer-drawn direction field for this system 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 the species' populations change over time, but we'll do it without the aid of a computer. This will also be a stepping-stone to determining how solutions to (7.4)-7.5) behave for any choice of the parameters r_1, K_1, r_2, K_2, a and b .

Our ultimate goal is a sketch that illustrates how solutions behave in the u_1u_2 phase plane, which we can interpret 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 finding the nullclines for the system.

The \dot{u}_1 -Nullcline

Our first step is to determine the \dot{u}_1 -nullcline, that is, the set of points in the u_1u_2 -plane at which $\dot{u}_1 = 0$ for solutions to (7.25)-(7.26). Given that $\dot{u}_1 = f_1(u_1, u_2)$ we find this nullcline by setting $f_1(u_1, u_2) = 0$. This yields equation

$$u_1(1 - u_1/2) - 0.1u_1u_2 = 0. \quad (7.27)$$

A little algebra (factor out u_1) shows that (7.27) 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 \dot{u}_1 -nullcline thus consists of two pieces, the vertical line $u_1 = 0$ (the u_2 axis) and the line $1 - u_1/2 - 0.1u_2 = 0$. This nullcline is shown in red in the left panel of Figure 7.10, in which we are now plotting on the range $0 \leq u_1, u_2 \leq 10$ to show the entire relevant portion of the nullcline. Any solution to (7.25)-(7.26) that crosses or touches this nullcline must move with $\dot{u}_1 = 0$, that is, vertically in the u_1u_2 phase plane. We indicate this with a few vertical tick marks on the \dot{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 \dot{u}_1 -nullcline, the u_1 population is not changing, at least momentarily.

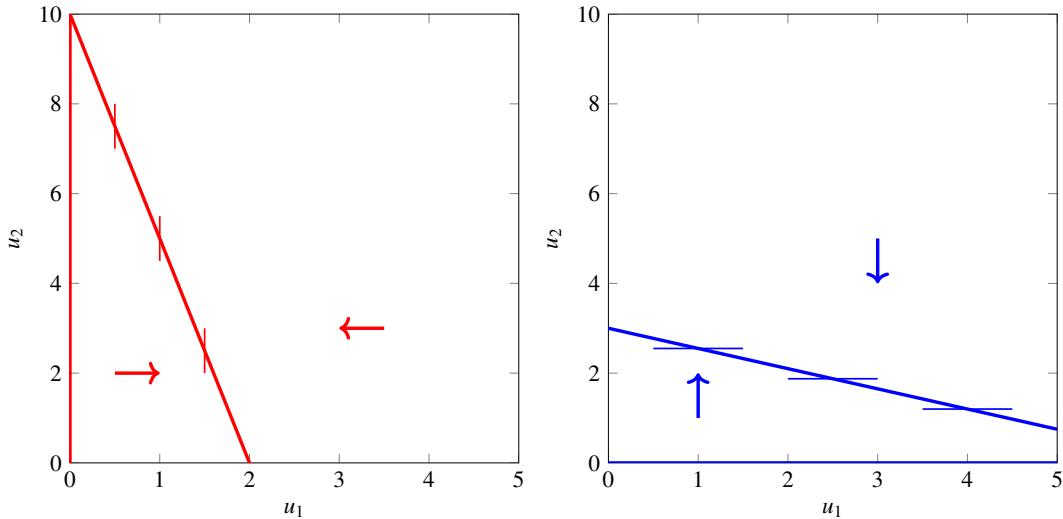


Figure 7.10: Left panel: Sketch of \dot{u}_1 -nullcline for (7.25)-(7.26). Right panel: Sketch of \dot{u}_2 -nullcline for (7.25)-(7.26).

Reading Exercise 179 In the left panel of Figure 7.10 it appears that if $u_1(t_0) = 0$ and $u_2(t_0) > 0$ (we are on the positive u_2 -axis at time $t = t_0$) then the solution to (7.25)-(7.26) 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 (7.25)-(7.26) with data $u_1(t_0) = 0, u_2(t_0) = u_0$.

What is the physical interpretation of this situation?

The \dot{u}_1 -nullcline divides the first quadrant into two distinct pieces, one of which lies below the line $1 - u_1/2 - 0.1u_2 = 0$, 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. We indicate this 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 $f_1(1, 1) = 0.3 > 0$, 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 vertical motion of the solution (this is determined using the $\dot{u}_2 = 0$ nullcline, which we discuss shortly).

Reading Exercise 180 Consider the left panel in Figure 7.10. If a solution $u_1 = u_1(t), u_2 = u_2(t)$ passes through the point $u_1 = 3, u_2 = 2$, is the u_1 population increasing or decreasing?

The \dot{u}_2 -Nullcline

We can perform precisely the same analysis for the u_2 nullcline. Setting $f_2(u_1, u_2) = 0$ yields $2u_2(1 - u_2/3) - 0.3u_1u_2 = 0$ or

$$u_2(2 - 2u_2/3 - 0.3u_1) = 0.$$

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 blue in the right panel of Figure 7.10. Any solution curve on that touches this nullcline does so with $\dot{u}_2 = 0$, and so is moving horizontally. In terms of the populations, the second species numbers are momentarily

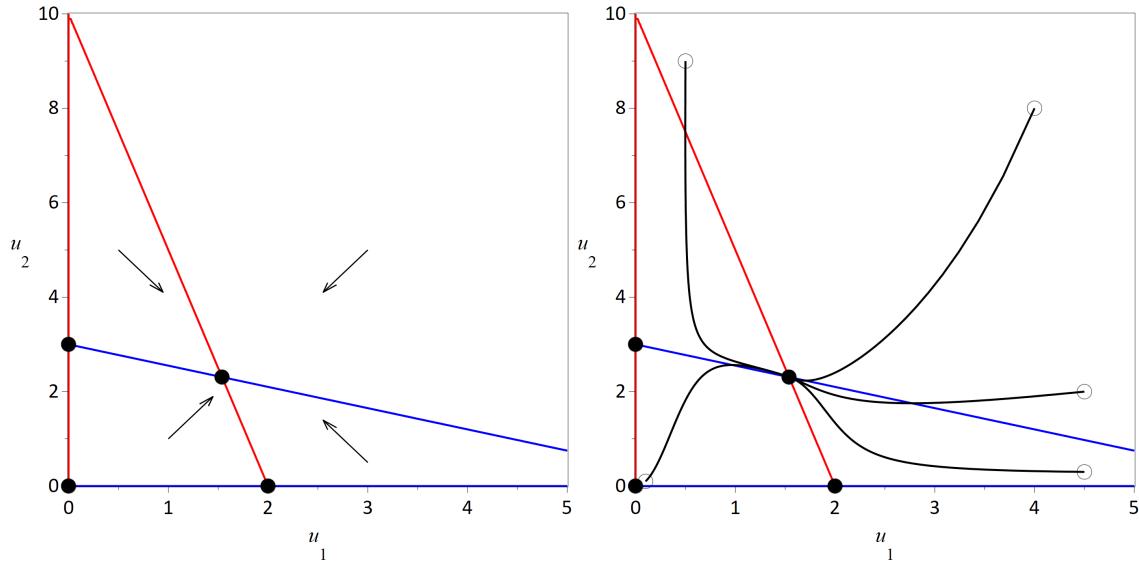


Figure 7.11: Left panel: Nullclines and direction arrows for (7.25)-(7.26), with equilibrium solutions shown as black dots. Right panel: Typical solution trajectories.

not changing at such a point. The nullcline divides the region of interest into two pieces, and in each we have either $\dot{u}_2 = f_2(u_1, u_2) < 0$ or $\dot{u}_2 = f_2(u_1, u_2) > 0$, that is, the u_2 population is either decreasing or increasing. To figure out which, we can 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 we draw an arrow with tail at $(3, 5)$ pointing downward.

Reading Exercise 181 Suppose that at some instant in time $u_1 = 2$ and $u_2 = 4$. Is \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 182 In the right panel of Figure 7.10 it appears that if $u_2(t_0) = 0$ and $u_1(t_0) > 0$ (we are on the positive u_1 -axis at time $t = t_0$) then the solution to (7.25)-(7.26) 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 (7.25)-(7.26) with data $u_1(t_0) = u_0, u_2(t_0) = 0$.

What is the physical interpretation of this situation?

Reading Exercise 183 Use Reading Exercises 179 and 182 in conjunction with the Existence-Uniqueness Theorem 6.1.1 to argue that solutions to (7.25)-(7.26) that start with initial data $u_1(t_0) > 0, u_2(t_0) > 0$ can never leave the first quadrant $u_1, u_2 > 0$, or even touch the axes. In this model if we start with positive populations for each species we can never find ourselves with a negative population.

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 up/left, up/right, down/left, or down/right arrows, to indicate the general motion of solutions.

Observe in the left panel of Figure 7.11 that there are four fixed points or equilibrium solu-

tions 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 about $(1.54, 2.31)$. These points are highlighted with black dots and can be found by solving the equations $f_1(u_1, u_2) = 0$, $f_2(u_1, u_2) = 0$ simultaneously, where f_1 and f_2 are given by (7.25) and (7.26).

Reading Exercise 184 What is the physical interpretation, in terms of the two species' populations, of each fixed point?

Based on the arrows in the left panel and using the nullclines as a guide, we can sketch solutions curves to (7.25)-(7.26) starting at any initial point. Although the curves in the text were drawn using a computer, this can easily be done 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. Note that in each case the trajectory follows the general left/up, left/down, right/up, right/down 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 \dot{u}_1 nullcline (in red) vertically. The solution that starts near $(u_1, u_2) = (4.5, 2)$ crosses the \dot{u}_2 nullcline (in blue) horizontally. The sketch in the right panel of Figure 7.11 is known as a *phase portrait*. 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 simply drawing a lot of arrows.

Based on our sketch it is clear that all solutions approach the fixed point at $u_1 \approx 1.54, u_2 \approx 2.31$ as $t \rightarrow \infty$. 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 the fixed point; for this ODE system it seems that all solutions approach this fixed point. . The fixed point at $(0, 0)$ appears to be unstable, for solutions that start near (but not at) near $(0, 0)$ move away. The fixed points at $(2, 0)$ and $(0, 3)$ are slightly more subtle, but not much. A little experimentation with drawing trajectories should convince you that although solutions may initially approach these fixed points, most solutions eventually turn away toward the stable fixed point at $(1.54, 2.31)$.

Based on this analysis we can make an extremely important conclusion regarding the physical behavior of this system (with the particular parameters r_1, K_1, r_2, K_2, a , and b):

For any starting populations the species will approach stable coexistence.

The Behavior of Solutions and Physical Interpretation

It's important to note that the phase portrait is not an end unto itself, but a tool that lets us understand the behavior of solutions to the system, and these solutions are functions $u_1(t)$ and $u_2(t)$ of time t that indicate the population of each species. 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 (moves to the left) and converges to $u_1 \approx 1.54$. The function u_2 starts at $u_2(0) = 2$, initially decreases, then increases to $u_2 \approx 2.31$. Based on this we can sketch u_1 and u_2 individually as functions of t , as shown in Figure 7.12. Although these graphs were drawn with a computer, we could easily sketch qualitatively correct versions using only the phase portrait on the right in Figure 7.11. 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 $u_2 \approx 2.31$.

Reading Exercise 185 Based on the 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

- $u_1(0) = 4, u_2(0) = 8$.
- $u_1(0) = 1, u_2(0) = 8$.
- $u_1(0) = 0.5, u_2(0) = 0.5$.

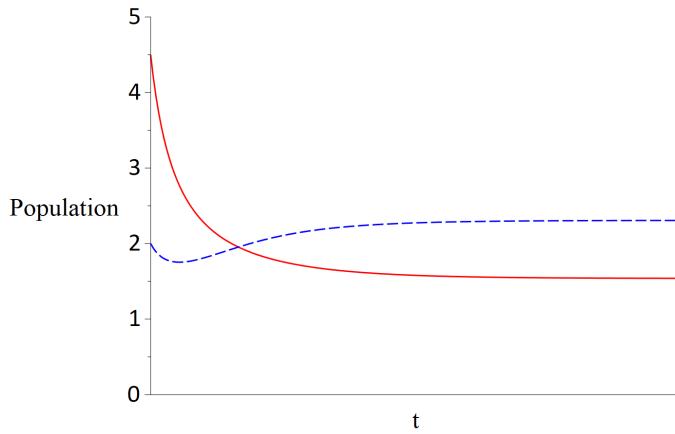


Figure 7.12: Species 1 population $u_1(t)$ (red, solid) and species 2 population $u_2(t)$ (blue, dashed) versus t for (7.25)-(7.26) with initial populations $u_1(0) = 4.5, u_2(0) = 2$.

7.3.2 Changing the Parameters

For the system (7.25)-(7.26) we could simply have had the computer sketch the direction field, then used that to visualize the solutions. Why trouble ourselves with the nullclines, manually computed direction arrows, and hand-sketched solution curves?

To answer this, consider the phase portrait in Figure 7.13. This is the same system (7.4)-7.5) of ODEs that led to (7.25)-(7.26), 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 direction, 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. But these conclusions are now a bit less obvious. Why did the stability of the fixed points, especially the fixed point with $u_1, u_2 > 0$ that represents stable coexistence, change? Is it because the competition is fiercer? If so, what combinations of a and/or b doom one species or the other to extinction? What if a and b are both large? How do the other parameters r_1, r_2, K_1 , and K_2 factor into this?

To answer this we need to perform the same analysis that led to Figure 7.11, but without assuming specific values for the parameters (aside from the physical necessity of requiring all to be positive in this model.) Happily, this can be done using exactly the approach we've already developed.

Rescaling the ODEs

Before we proceed it will be helpful to make a change in the dependent variables by replacing u_1 and u_2 in (7.4)-7.5) with rescaled functions

$$v_1 = u_1/K_1 \quad \text{and} \quad v_2 = u_2/K_2, \quad (7.28)$$

in spirit of Section 4.5. We are, in effect, changing the units we use to measured population, so that v_1 quantifies the first species' population 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 186 Verify that if we use $\dot{u}_1 = K_1 \dot{v}_1$ and $\dot{u}_2 = K_2 \dot{v}_2$ along with (7.28) in (7.4)-

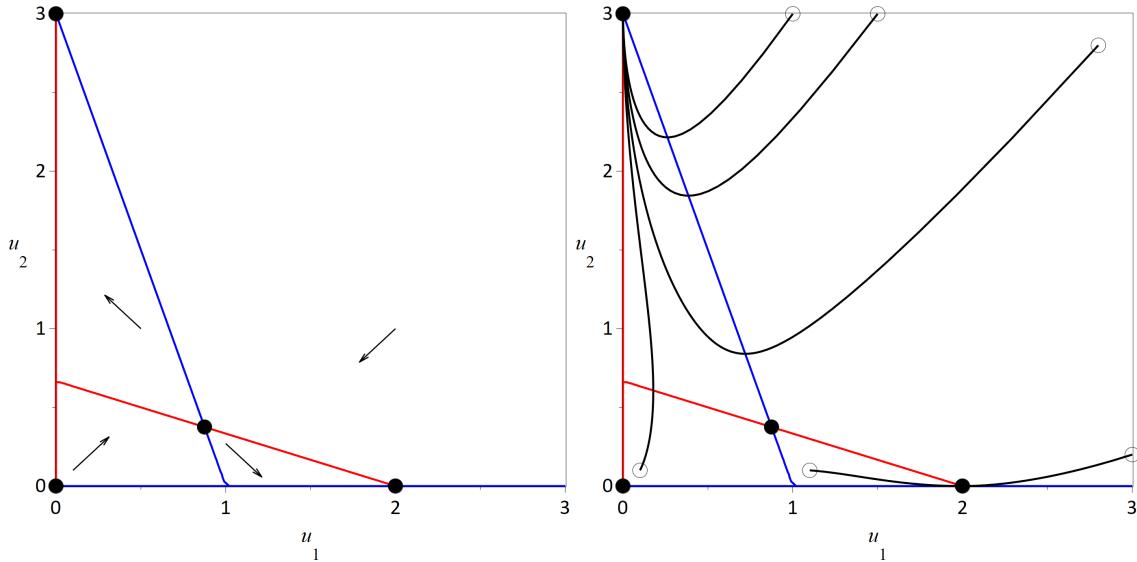


Figure 7.13: Left panel: Nullclines and direction arrows for (7.4)-(7.5) with $r_1 = 1, K_1 = 2, r_2 = 2, K_2 = 3, a = 3$, and $b = 3$; equilibrium solutions shown as black dots. Right panel: Typical solution trajectories.

7.5) we obtain the system

$$\dot{v}_1 = r_1 v_1 (1 - v_1 - \bar{a} v_2) \quad (7.29)$$

$$\dot{v}_2 = r_2 v_2 (1 - v_2 - \bar{b} v_1) \quad (7.30)$$

where

$$\bar{a} = K_2 a / K_1 \quad \text{and} \quad \bar{b} = K_1 b / K_2. \quad (7.31)$$

We will analyze the ODE system (7.29)-(7.30), from which we can use (7.28) to make conclusions about the original system (7.4)-(7.5). The advantage of the (7.29)-(7.30) is that we have only 4 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.

7.3.3 Sketching Phase Portraits with Unspecified Parameters

Figure 7.11 was constructed from the ODEs (7.25)-(7.26) using the following steps:

1. Compute the \dot{u}_1 -nullcline by graphing those points in the $u_1 u_2$ phase plane that satisfy $f_1(u_1, u_2) = 0$. This nullcline divides the plane into various regions in which $f_1(u_1, u_2) > 0$ (so $\dot{u}_1 > 0$) or $f_1(u_1, u_2) < 0$ (so $\dot{u}_1 < 0$). Put a right or left arrow in each such region, to indicate the horizontal motion of solutions.
2. Compute the \dot{u}_2 -nullcline by graphing those points in the $u_1 u_2$ phase plane that satisfy $f_2(u_1, u_2) = 0$. This nullcline divides the plane into various regions in which $f_2(u_1, u_2) > 0$ (so $\dot{u}_2 > 0$) or $f_2(u_1, u_2) < 0$ (so $\dot{u}_2 < 0$). Put an up or down arrow in each such region, to indicate the vertical motion of solutions.
3. Synthesize the plots from steps (1) and (2) by overlaying the nullcline graphs, which divide the phase plane into a number of regions. In each region use the nullcline plots to decide whether solutions move up/left, up/right, down/left, or down/right and sketch an appropriate arrow. Equilibrium solutions occur at the intersection of the nullclines.

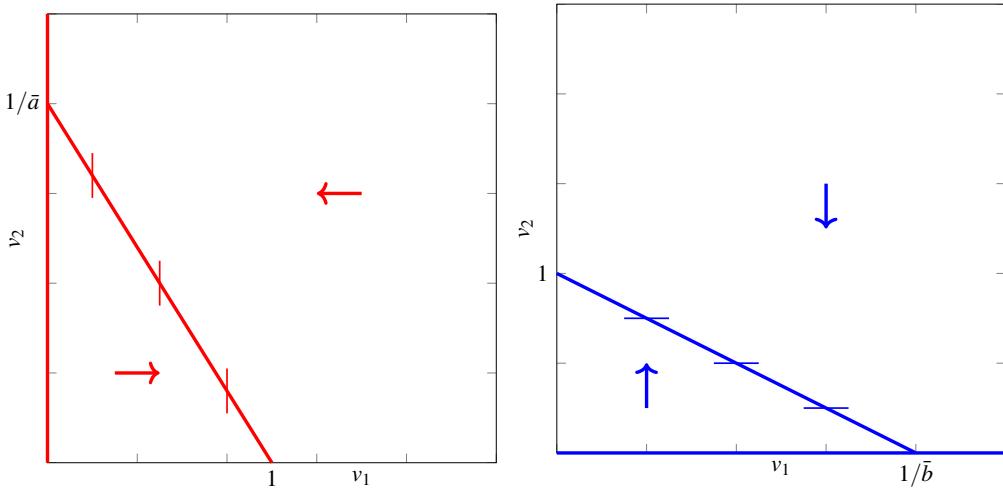


Figure 7.14: Left panel: Sketch of \dot{v}_1 -nullcline for (7.29)-(7.30). Right panel: Sketch of \dot{v}_2 -nullcline for (7.29)-(7.30).

4. Choose a representative sample of initial points and sketch solution trajectories consistent with the arrows in the plot from step (3), as well as with the nullclines.

But the punchline for this whole procedure is this final step:

Interpret what the solutions you sketched say about the physical behavior of the system.

A Phase Portrait for (7.29)-(7.30)

Let's carry out these steps on (7.29)-(7.30) 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.

For step 1, we set $f_1(v_1, v_2) = 0$ which leads to

$$r_1 v_1 (1 - v_1 - \bar{a} v_2) = 0.$$

Since $r_1 > 0$ we find that the 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 $-1/\bar{a}$. This nullcline is shown in red in the left panel of Figure 7.14, with a few vertical tick marks to indicate solution directions on the nullcline. We can sketch this nullcline without choosing specific values for the system parameters, as long as we label intercepts 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 $\dot{v}_1 < 0$; 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 the v_1 population is increasing. The situation is indicated by the appropriate left and right pointing vectors in the left panel of Figure 7.14.

For step 2 we sketch the \dot{v}_2 nullcline and proceed similarly. If we set $f_2(v_1, v_2) = 0$ then from (7.30) we find that 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 7.14. 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/down arrows in the right panel.

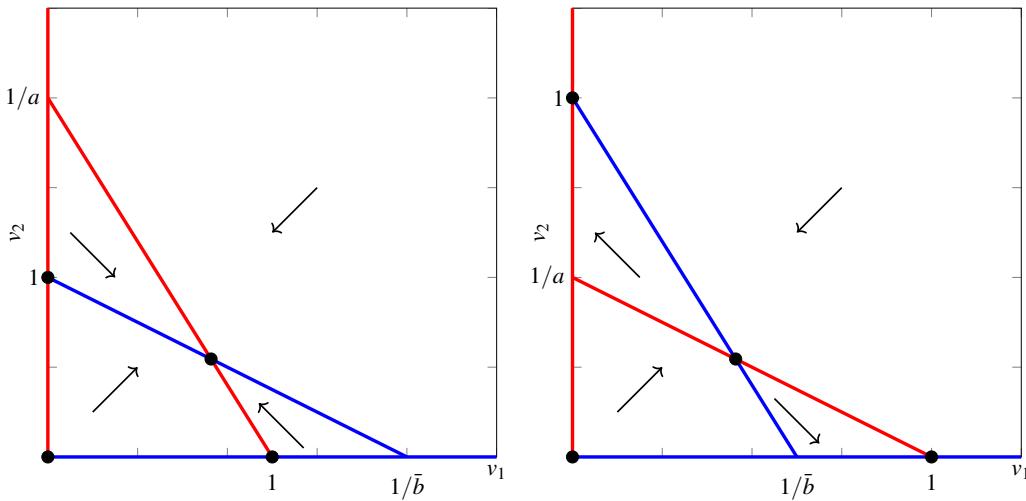


Figure 7.15: Left panel: Nullclines and direction arrows for (7.29)-(7.30) when $1 < 1/\bar{a}$ and $1 < 1/\bar{b}$. Right panel: Nullclines and direction arrows for (7.29)-(7.30) when $1 > 1/\bar{a}$ and $1 > 1/\bar{b}$. Equilibrium solutions in each case are shown as black dots.

Reading Exercise 187 Show that the equilibrium solutions to (7.29)-(7.30) are given by

$$(0,0), (1,0), (0,1), ((\bar{a}-1)/(\bar{a}\bar{b}-1), (\bar{b}-1)/(\bar{a}\bar{b}-1)).$$

The situation is now analogous to that after we constructed Figure 7.10. In that example we superimposed the nullcline graphs to produce the left panel in Figure 7.11 from which we could sketch representative solutions. However, there is now one important difference: we don't know the precise relation between the diagonal lines in the nullclines of Figure 7.14. As drawn in that figure they represent the case in which we have $1 < 1/\bar{a}$ 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 we can amalgamate the nullclines to produce the left panel in Figure 7.15. 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 188 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 (7.25)-(7.26) that led to Figures 7.10 and 7.11 yield $\bar{a} = 0.3$ and $\bar{b} = 0.3$ in (7.29)-(7.30) and satisfy $1 < 1/\bar{a}$ and $1 < 1/\bar{b}$.

However, consider the possibility that $1 > 1/\bar{a}$ and $1 > 1/\bar{b}$. In this case the situation is similar to that that led to Figure 7.13 with $r_1 = 1, r_2 = 2, K_1 = 2, K_2 = 3, a = 3$ and $b = 2$, corresponding to $\bar{a} = 4.5, \bar{b} = 2$. The resulting overlayed nullclines and direction arrows appear as shown in the right panel of Figure 7.15.

Reading Exercise 189 Sketch the nullclines for (7.29)-(7.30) under the assumption that $1 > 1/\bar{a}$ and $1 > 1/\bar{b}$, and verify that the picture in the right panel of Figure 7.15 is correct.

The two cases illustrated in Figure 7.15 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 190 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 7.15).
2. $1 > 1/\bar{a}$ and $1 > 1/\bar{b}$ (examined above in the right panel of Figure 7.15).

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 possibilities like $\bar{a} = 1$ and/or $\bar{b} = 1$.

Reading Exercise 191 Based on the analysis above, and in particular, Figure 7.15, determine the behavior of the system if $1 < 1/\bar{a}$ and $1 < 1/\bar{b}$ (case 1 in Reading Exercise 190.) In particular, sketch some representative solution curves; how do solutions behave as $t \rightarrow \infty$? Translate your conclusions back to the original population variables u_1 and u_2 . What is the fate of each species? Repeat your analysis for the case in which $1 > 1/\bar{a}$ and $1 > 1/\bar{b}$ (case 2 in Reading Exercise 190.)

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, depending on the initial conditions.

7.3.4 Linearizing Multivariable Functions

The conclusions above aren't fully ironclad, though, as they are based on an examination of the graphs in Figure 7.15; how solutions behave, especially near fixed points, is not entirely clear. For example, the behavior of solutions near the fixed point at $v_1 = 1, v_2 = 0$ in the right panel of Figure 7.15 seem to indicate that this fixed point is stable (solutions that start close to that fixed point approach it asymptotically), but it's not as painfully clear as the situation in, say, the left panel of Figure 7.15 with the equilibrium solution in which $v_1, v_2 > 0$.

We can gain further insight into the behavior of solutions near fixed points by using *linearization*. Recall from multivariable calculus that a function $f(x,y)$ of two variables can often be well-approximated near a point $x = c, y = d$, as $f(x,y) \approx L(x,y)$ where

$$L(x,y) = f(c,d) + \frac{\partial f}{\partial x}(c,d)(x-c) + \frac{\partial f}{\partial y}(c,d)(y-d). \quad (7.32)$$

The function L is called the *linearization of f at the point (c,d)* . This of course assumes that both partial derivatives exist, and technically, we would like them to be continuous near $x = c, y = d$ as well. 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 act as a stand in for f in certain computations. Because L is linear, the computation may be greatly simplified.

■ **Example 7.4** Let $f(x,y) = x(1-x/2) - 0.1xy$. To linearize f at the point $x = 2, y = 0$ we compute $\frac{\partial f}{\partial x} = 1 - x - 0.1y$, $\frac{\partial f}{\partial y} = -0.1x$. From (7.32) with $c = 2, d = 0$ we find that $\frac{\partial f}{\partial x}(2,0) = -1$ and $\frac{\partial f}{\partial y}(2,0) = -0.2$, so that

$$L(x,y) = 0 - (x-2) - 0.2(y-0) = -x - 0.2y + 2.$$

A graph of the functions $f(x,y)$ and $L(x,y)$ near the point $x = 2, y = 0$ is shown in Figure 7.16. From a geometric perspective, the graph $z = L(x,y)$ is the tangent plane to the surface $z = f(x,y)$ at the point $x = c, y = d, z = f(c,d)$. ■

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 [37].

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

$$L(\mathbf{x}) = f(\mathbf{c}) + \sum_{k=1}^n \frac{\partial f}{\partial x_k}(\mathbf{c})(x_k - c_k). \quad (7.33)$$

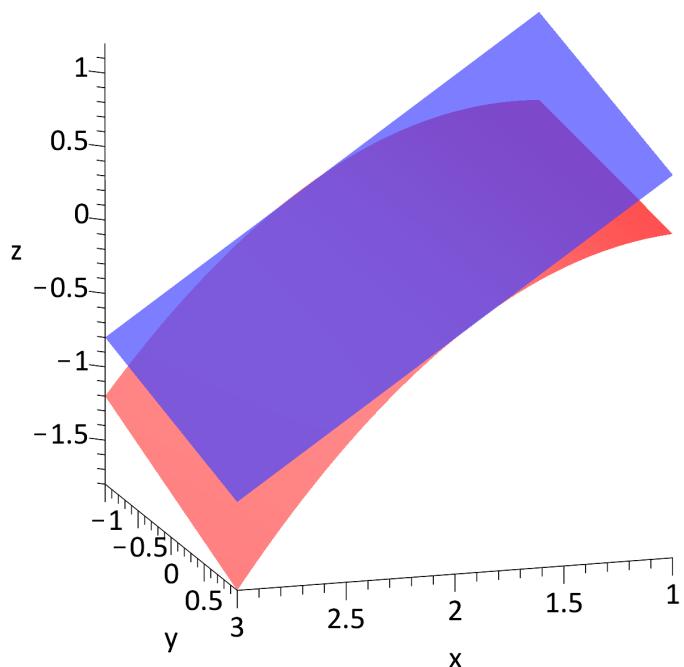


Figure 7.16: Nonlinear function $f(x,y) = x(1-x/2) - 0.1xy$ graphed in red and linearization $L(x,y) = -x - 0.2y + 2$ at $x = 2, y = 0$ graphed in blue.

7.3.5 Linearizing ODEs at Equilibrium Points

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 (7.25)-(7.26), 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 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 then use this to infer the stability of the original nonlinear system near the relevant equilibrium solution. We carry out this computation in Example 7.5 below. Under the right circumstances the nonlinear and linearized system share the same stability properties at the fixed point, as detailed in the Hartman-Grobman Theorem, stated below.

■ Example 7.5 For the system (7.25)-(7.26) 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 $L_1(u_1, u_2) = -u_1 - 0.2u_2 + 2$. By using $f_2(u_1, u_2) = 2u_2(1 - u_2/3) - 0.3u_1u_2$ we find linearization $L_2(u_1, u_2) = 1.4u_2$. The resulting linearized system is $\dot{u}_1 = -u_1 - 0.2u_2 + 2, \dot{u}_2 = 1.4u_2$. In matrix form the linearized system is $\dot{\mathbf{u}} = \mathbf{A}\mathbf{u} + \mathbf{b}$ where $\mathbf{u}(t) = \langle u_1(t), u_2(t) \rangle$ and

$$\mathbf{A} = \begin{bmatrix} -1 & -0.2 \\ 0 & 1.4 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = -\begin{bmatrix} 2 \\ 0 \end{bmatrix}. \quad (7.34)$$

Based on the discussion in Section 7.2 we can determine the stability of the fixed point $c = 2, d = 0$ for the linearized system by computing the eigenvalues for \mathbf{A} , which turn out to be $\lambda_1 = -1, \lambda_2 = 1.4$. Since these are real and of mixed sign, the point $u_1 = 2, u_2 = 0$ is a saddle point. ■

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 (7.25)-(7.26) as well. Of course this computation can be carried out for the other fixed points for (7.25)-(7.26), 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 $f_1(c, d) = f_2(c, d) = 0$, so 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

$$\mathbf{A} = \begin{bmatrix} \frac{\partial f_1}{\partial u_1}(c, d) & \frac{\partial f_1}{\partial u_2}(c, d) \\ \frac{\partial f_2}{\partial u_1}(c, d) & \frac{\partial f_2}{\partial u_2}(c, d) \end{bmatrix} \quad \text{and} \quad \mathbf{b} = -\mathbf{A} \begin{bmatrix} c \\ d \end{bmatrix}. \quad (7.35)$$

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 192 Linearize (7.25)-(7.26) at the fixed point $u_1 \approx 1.54, u_2 \approx 2.31$ and write the matrix \mathbf{A} that embodies the linearized system. Compute the eigenvalues for \mathbf{A} . What is the stability 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 mutual coexistence of the species?

The Jacobian Matrix

The process of linearizing a nonlinear system of two ODEs that gave rise to the matrix \mathbf{A} in Example 7.5 works for larger systems. When we linearize each function f_j in a larger system $\dot{u}_j = f_j(u_1, \dots, u_n)$, $1 \leq j \leq n$, using (7.33) we are led to a linearized system $\dot{\mathbf{u}} = \mathbf{A}\mathbf{u} + \mathbf{b}$ where \mathbf{A} is computed using the so-called *Jacobian matrix*:

Definition 7.3.1 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

$$\mathbf{J} = \begin{bmatrix} \frac{\partial f_1}{\partial u_1} & \cdots & \frac{\partial f_1}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial u_1} & \cdots & \frac{\partial f_n}{\partial u_n} \end{bmatrix} \quad (7.36)$$

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 (7.36) are functions of u_1, \dots, u_n , and hence so is \mathbf{J} .

■ **Example 7.6** For the system (7.25)-(7.26) we have

$$\frac{\partial f_1}{\partial u_1} = -u_1 - 0.1u_2, \quad \frac{\partial f_1}{\partial u_2} = -0.1u_1, \quad \frac{\partial f_2}{\partial u_1} = -0.3u_2, \quad \frac{\partial f_2}{\partial u_2} = -4u_2/3 - 0.3u_1.$$

The Jacobian matrix is then

$$\mathbf{J}(u_1, u_2) = \begin{bmatrix} -u_1 - 0.1u_2 & -0.1u_1 \\ -0.3u_2 & -4u_2/3 - 0.3u_1 \end{bmatrix}.$$

Note that 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 equation 7.34 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 $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}$ 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 this lets us determine the stability of the nonlinear system.

The Hartman-Grobman Theorem

A condition under which the linearized system and the nonlinear system share the same stability properties at a given fixed point is that the fixed point be *hyperbolic*:

Definition 7.3.2 Suppose \mathbf{p} is an equilibrium 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 has a nonzero real part then \mathbf{p} is called a *hyperbolic equilibrium point*.

If an eigenvalue λ has a nonzero real part this means either that λ is real and nonzero or 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. That is, 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. We'll see some examples below.

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 are given in

the exercises. Our statement of the Hartman-Grobman Theorem isn't as precise as it could be, but will suffice for our purposes. For a more careful statement see [49].

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 $\dot{u}_j = f_j(u_1, \dots, u_n)$ as follows:

1. Compute the Jacobian matrix $\mathbf{J}(\mathbf{u})$ defined by (7.36) 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 7.3.2) then the nonlinear and linearized system have the same qualitative stability properties at \mathbf{p} .

■ **Example 7.7** Let's finish our analysis of the system (7.25)-(7.26). 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 $\lambda_2 = 1.4$. Neither is zero, so this is an equilibrium point, a saddle for the linearized system. Based on the Hartman-Grobman Theorem we can conclude this is a saddle point for the nonlinear system. This point is unstable, and solutions will generally not approach it. Physically, the situation in which species one is near its carrying capacity and relatively few of species two are present is an unstable situation here; solutions will move away from this fixed point. What the solution trajectory does and where it goes in the long run is 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, 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, and unstable. Solutions that start nearby will not generally tend toward $u_1 = 0, u_2 = 3$ (the extinction of species one.)

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, a stable sink. Solutions to the linearized system approach this point, and so do solutions to the nonlinear system that start sufficiently close. ■

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 (7.25)-(7.26) all initial conditions results in the long-term stable coexistence of the species. Neither will drive the other to extinction.

Reading Exercise 193 Compute the Jacobian matrix for the system (7.4)-7.5) with parameters choices $r_1 = 1, K_1 = 2, r_2 = 2, K_2 = 3$ but with $a = b = 3$. 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 stable sinks, while the coexistence fixed point $(7/8, 3/8)$ is an unstable saddle.

Convince yourself this is consistent with Figure 7.13. What can you conclude about the long-term behavior of the populations here?

7.3.6 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 (7.29)-7.30) with unspecified (positive) parameters r_1, r_2, \bar{a} , and \bar{b} (as obtained from (7.4)-7.5) with positive parameters r_1, r_2, K_1, K_2, a , and b). We've already sketched the nullclines and a simple phase portrait in Figure 7.15; we will now perform linearization on the equilibrium solutions and determine precisely how the various parameters affect the fate of each species.

Equations (7.29)-7.30) are reproduced here for convenience:

$$\dot{v}_1 = r_1 v_1 (1 - v_1 - \bar{a} v_2) \quad (7.37)$$

$$\dot{v}_2 = r_2 v_2 (1 - v_2 - \bar{b} v_1) \quad (7.38)$$

Recall that $\bar{a} = K_2 a / K_1$ and $\bar{b} = K_1 b / K_2$, and $\bar{a}, \bar{b} > 0$.

The equilibrium solutions for this system were given in Reading Exercise 187 and are

$$(0, 0), (1, 0), (0, 1), ((\bar{a} - 1)/(\bar{a}\bar{b} - 1), (\bar{b} - 1)/(\bar{a}\bar{b} - 1)). \quad (7.39)$$

This first three points above are always physically relevant; the fourth fixed point is relevant only when both components are nonnegative. Some typical possibilities for the nullclines were shown in Figure 7.15 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 7.15 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 ($(0, 0)$ is unstable in both cases.) Linearization can provide additional insight into this phenomena.

The Jacobian matrix for (7.37)-(7.38) is given by

$$\mathbf{J}(v_1, v_2) = \begin{bmatrix} r_1(1 - 2v_1 - \bar{a}v_2) & -r_1\bar{a}v_1 \\ -r_2\bar{b}v_2 & r_2(1 - 2v_2 - \bar{b}v_1) \end{bmatrix}. \quad (7.40)$$

We evaluate $\mathbf{J}(v_1, v_2)$ at each equilibrium solution, compute the relevant eigenvalues, and use this to determine the nature of the equilibrium solution.

Stability of the Origin

At $v_1 = v_2 = 0$ we find

$$\mathbf{J}(0, 0) = \begin{bmatrix} r_1 & 0 \\ 0 & r_2 \end{bmatrix}.$$

The eigenvalues are r_1 and r_2 , both 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 mutual extinction point.

Stability When One Species Is Extinct

At $v_1 = 1, v_2 = 0$ we find

$$\mathbf{J}(1, 0) = \begin{bmatrix} -r_1 & -r_1\bar{a} \\ 0 & r_2(1 - \bar{b}) \end{bmatrix}.$$

This matrix is upper triangular, so the eigenvalues are exactly the diagonal elements $-r_1$ and $r_2(1 - \bar{b})$. Of course $-r_1 < 0$ always, 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 a stable sink. 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

$$\mathbf{J}(1, 0) = \begin{bmatrix} r_1(1 - \bar{a}) & 0 \\ -r_2\bar{b} & -r_2 \end{bmatrix}.$$

This matrix is lower triangular, so the eigenvalues are exactly the diagonal elements $-r_2$ and $r_1(1 - \bar{a})$. This is similar to the previous fixed point. If $\bar{a} < 1$ this is a saddle, and if $\bar{a} > 1$ this is a stable sink, and if $\bar{a} = 1$ this equilibrium solution is not hyperbolic and this analysis will not reveal its nature.

Reading Exercise 194 Summarize the stability analysis above for the fixed point $(1, 0)$ and $(0, 1)$ in terms of \bar{a} and \bar{b} . What does each case imply for the two species' populations? Given the interpretation of \bar{a} and \bar{b} as competition parameters, why does this make sense?

Stability for Mutual Coexistence

To examine the linearized system at the last equilibrium solution in the list (7.39), 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 $v_1^*, v_2^* \geq 0$. We compute

$$\mathbf{J}(v_1^*, v_2^*) = \frac{1}{\bar{a}\bar{b} - 1} \begin{bmatrix} -r_1(\bar{a} - 1) & -r_1\bar{a}(\bar{a} - 1) \\ -r_2\bar{b}(\bar{b} - 1) & -r_2(\bar{b} - 1) \end{bmatrix}. \quad (7.41)$$

This matrix is not upper or lower triangular, and the eigenvalues are a bit messy. Even with the aid of a computer algebra system, the result is undecipherable, unless we are a bit clever.

Define variables $p = -r_1 v_1^*$ and $q = -r_2 v_2^*$ and take note of the fact that $v_1^*, v_2^* \geq 0$ corresponds to $p, q \leq 0$. From (7.41) we obtain

$$\mathbf{J}(v_1^*, v_2^*) = \begin{bmatrix} p & \bar{a}p \\ \bar{b}q & q \end{bmatrix}. \quad (7.42)$$

The eigenvalues of this matrix are given by

$$\lambda_1 = \frac{(p + q) - \sqrt{(p + q)^2 + 4(\bar{a}\bar{b} - 1)pq}}{2}, \quad \lambda_2 = \frac{(p + q) + \sqrt{(p + q)^2 + 4(\bar{a}\bar{b} - 1)pq}}{2}. \quad (7.43)$$

It's easy to see that if either $p = 0$ or $q = 0$ then at least one of these eigenvalues is zero and our analysis will not be sufficient to deduce stability, so let's assume $p, q < 0$; this corresponds to $v_1^*, v_2^* > 0$, so that both populations are actually positive. Our goal is to determine the nature of these eigenvalues and how they depend on p, q, \bar{a} , and \bar{b} ; specifically, are they real, and if so, what is the sign of the eigenvalues?

As it turns out, under the present circumstances both eigenvalues in (7.43) are real. To see why, note that if $\bar{a}\bar{b} > 0$ then $4(\bar{a}\bar{b} - 1) > -4$ and since $pq > 0$ we find that $4(\bar{a}\bar{b} - 1)pq > -4pq$. Thus

$$(p + q)^2 + 4(\bar{a}\bar{b} - 1)pq > (p + q)^2 - 4pq = (p - q)^2 \geq 0.$$

This means the discriminant under the square root in the eigenvalues in (7.43) is always nonnegative; both eigenvalues are thus real numbers.

It is then easy to see that $\lambda_1 < 0$ always (since $p + q < 0$ and we subtract a nonnegative square root), and the stability of this equilibrium point comes down to the sign of λ_2 . If $\bar{a}\bar{b} - 1 > 0$ then $4(\bar{a}\bar{b} - 1)pq > 0$ and so $(p + q)^2 + 4(\bar{a}\bar{b} - 1)pq > (p + q)^2$. Then

$$\sqrt{(p + q)^2 + 4(\bar{a}\bar{b} - 1)pq} > \sqrt{(p + q)^2} = -(p + q) \quad (7.44)$$

and so λ_2 in (7.43) satisfies $\lambda_2 > 0$; in this case (v_1^*, v_2^*) is a saddle point. But if $\bar{a}\bar{b} - 1 < 0$ then $4(\bar{a}\bar{b} - 1)pq < 0$ and so $(p + q)^2 + 4(\bar{a}\bar{b} - 1)pq < (p + q)^2$. Then

$$\sqrt{(p + q)^2 + 4(\bar{a}\bar{b} - 1)pq} < \sqrt{(p + q)^2} = -(p + q) \quad (7.45)$$

and so $\lambda_2 < 0$ and this point is a stable sink.

7.3.7 Conclusions for Competing Species

We can now make some firm conclusions concerning the fate of each species in (7.4)-7.5), 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 $\bar{a}, \bar{b} > 0$. Note that since $\bar{a} = K_2 a / K_1$ and $\bar{b} = K_1 b / K_2$ we have

$$\bar{a}\bar{b} = ab \quad (7.46)$$

and so conclusions below 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 (7.31) into, for example, $K_2 a / K_1 < 1$ or $a < K_1 / K_2$.

For any parameter choices in the system (7.29)-(7.30) the origin $(v_1, v_2) = (0, 0)$ is an unstable source. From (7.28) 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 (7.4)-7.5). 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

$$(v_1^*, v_2^*) = \left(\frac{\bar{a} - 1}{\bar{a}\bar{b} - 1}, \frac{\bar{b} - 1}{\bar{a}\bar{b} - 1} \right) \quad (7.47)$$

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 (7.29)-(7.30) are saddle points, and hence so are the points $(K_1, 0)$ and $(0, K_2)$ for (7.4)-7.5). But the discussion leading to (7.45) then shows that since $\bar{a}\bar{b} - 1 < 0$, the point (v_1^*, v_2^*) is a stable sink. In the original system (7.4)-7.5) the coexistence equilibrium point defined by (7.7) is also a stable sink.

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 (7.47) 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 (7.29)-(7.30) are stable sinks, and hence so are the points $(K_1, 0)$ and $(0, K_2)$ for (7.4)-7.5). But the discussion leading to (7.44) then shows that since $\bar{a}\bar{b} - 1 > 0$, the point (v_1^*, v_2^*) is a saddle. The same conclusion applies to the coexistence equilibrium (7.7) for (7.4)-7.5).
3. Suppose $\bar{a} > 1$ while $\bar{b} < 1$. In this case exactly one of v_1^* or v_2^* in (7.47) is negative and the equilibrium solution (v_1^*, v_2^*) is not physically relevant; the same holds for the solution (7.7) for (7.4)-7.5). In this case since $\bar{a} > 1$ the fixed point $(0, 1)$ is a stable sink and since $\bar{b} < 1$ the point $(1, 0)$ is an unstable source. This means that $(0, K_2)$ is a stable sink and $(K_1, 0)$ is an unstable source 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 (7.4)-7.5)) then the species populations will approach mutual coexistence as defined by (7.7). 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 conditions/populations. If the competition is lop-sided ($a < K_1/K_2$ 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.

7.3.8 Higher Dimensional Systems

7.3.9 Exercises

Exercise 7.3.1 For each system of ODE's $\dot{x}_1 = f_1(x_1, x_2), \dot{x}_2 = f_2(x_1, x_2)$ below.

- Find and sketch the \dot{x}_1 nullcline on the indicated range for x_1, x_2 . The nullcline will divide the plane into a some 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 \dot{x}_2 nullcline on the indicated range for x_1, x_2 . The nullcline will divide the plane into a some 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) .
- 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 function of t (as was done in Figure 7.12 for the system (7.25)-(7.26).)

- a. $\dot{x}_1 = 2 - x_1^2 - x_2, \dot{x}_2 = x_1 - x_2$ on the range $-5 \leq x_1, x_2 \leq 5$. Sketch a solution trajectory with $x_1(0) = 0, x_2(0) = 4$, as well as $x_1(t)$ versus t and $x_2(t)$ versus t . Repeat for the solution with $x_1(0) = -1, x_2(0) = 3$.
- b. $\dot{x}_1 = -2x_1 - x_2 - 2, \dot{x}_2 = -x_1 x_2$ on the range $-5 \leq x_1, x_2 \leq 5$. Sketch a solution trajectory with $x_1(0) = 2, x_2(0) = -2$, as well as $x_1(t)$ versus t and $x_2(t)$ versus t . Repeat for the solution with $x_1(0) = -1, x_2(0) = 3$.
- c. $\dot{x}_1 = x_1 x_2 + x_2^2, \dot{x}_2 = x_1 - 2x_2 + 3$ on the range $-5 \leq x_1, x_2 \leq 5$. Sketch a solution trajectory with $x_1(0) = -3, x_2(0) = 1$, as well as $x_1(t)$ versus t and $x_2(t)$ versus t . Repeat for the solution with $x_1(0) = -1, x_2(0) = 3$.
- d. $\dot{x}_1 = x_1^3 - 3x_1 - x_2, \dot{x}_2 = x_1 - x_2$ on the range $-5 \leq x_1, x_2 \leq 5$. Sketch a solution trajectory with $x_1(0) = -2, x_2(0) = -1$, as well as $x_1(t)$ versus t and $x_2(t)$ versus t . Repeat for the solution with $x_1(0) = -1, x_2(0) = 3$.

Exercise 7.3.2 The epidemic model (7.8) is a system of three ODE's in three functions $S(t)$, $I(t)$, and $R(t)$, but can effectively be considered a system of two differential equations

$$\begin{aligned}\dot{S} &= -aSI \\ \dot{I} &= aSI - bI\end{aligned}\tag{7.48}$$

where $a, b > 0$, since the first two equation don't involve R . We can analyze (7.48) to determine the behavior of $S(t)$ and $I(t)$, then use this to determine the behavior of $R(t)$ using $\dot{R} = bI$. We can concern ourselves only with the first quadrant $S, I \geq 0$.

- a. Show that the \dot{S} nullcline is given by the coordinate axes $S = 0$ and $I = 0$. Show that the \dot{I} nullcline is given by the horizontal coordinate axis $I = 0$ and vertical line $S = b/a$.
- b. Sketch the \dot{S} nullcline and appropriate arrow(s) to indicate the horizontal motion of any solution.
- c. Sketch the \dot{I} nullcline and appropriate arrow(s) 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 the 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 $\lambda = 0$ 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 7.3.3 Consider the damped nonlinear pendulum equation (7.9). 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 (7.10). Here $g > 0$ and $L > 0$ denote gravitational acceleration and the length of the pendulum, respectively, $c \geq 0$ is a damping constant, and $x_1(t), x_2(t)$ are the angular position and angular velocity of the pendulum with respect to vertical.

- a. Show that the \dot{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 \dot{x}_2 nullcline is the sine curve with graph $x_2 = -\frac{cg}{L} \sin(x_1)$. 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 $x_1 = k\pi, x_2 = 0$ where k 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 $(x_1, x_2) = (2\pi, 0)$? ■

Exercise 7.3.4 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 stability. Of course, you can't really draw nullclines, but can you guess how solutions typically behave? Try solving the system numerically and plotting the results to confirm your analysis. ■

7.4 Modeling Projects

7.4.1 Project: Homelessness Revisited

Recap of the Linear Model

In the Section 6.5 modeling project “Homelessness” we examine a linear system of ODE’s to model the fraction of homeless people in a city. We defined

$R(t)$ 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 considered the fraction of households in each category: if N is the total number of non-homeowner households 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 .

Of course $r(t) + e(t) = 1$ for all t .

A Nonlinear Model

In our first model we assumed that a constant fraction of the renting group, αr , transitioned to the evicted

that the city has exactly M rental units (apartments and houses).

In this section 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 decrease (i.e. when R is close to M).
- The flow rate from the evicted group to the renting group decreases as the number of vacancies decrease.

Modeling Exercises

Modeling Exercise 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 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 3 Suppose that $M = 118,500$. Using the same values of α and β from Part I, use technology to sketch the phase plane and the solution curves that satisfy the initial conditions $R(0) = 112,100$ and $E(0) = 5900$ (so initially, 95% of the non-homeowners are renting). 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?

Modeling Exercise 4 Something with general parameters.

7.4.2 Project: Predator-Prey Model

In this project we analyze one model of a “predator-prey” system in which two species exist. Unlike the competing yeast species of Section 7.1, the setting is quite asymmetric: one species is the “prey,” the other the “predator.”

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

$$\dot{x}_1 = r_1 x_1 (1 - x_1/K) \quad (7.49)$$

where $r_1 > 0$ is a 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 impact the prey species growth. We modify (7.49) in the same spirit as was done in (7.4) or (7.5). Specifically, we now take

$$\dot{x}_1 = \underbrace{r_1 x_1 \left(\frac{K - x_1(t) - ax_2(t)}{K} \right)}_{f_1(x_1, x_2)} \quad (7.50)$$

for some constant $a > 0$.

Modeling Exercise 1 Provide some justification for equation (7.50). 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 constant $r_2 > 0$, 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

$$\dot{x}_2 = \underbrace{(-r_2 + bx_1)x_2}_{f_2(x_1, x_2)} \quad (7.51)$$

for some constant $b > 0$.

Modeling Exercise 2 Provide some justification for equation (7.51). 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 3 Show that the equilibrium solutions for the system (7.50)-(7.51) are given by

$$(0,0), (K,0), \left(\frac{r_2}{b}, \frac{Kb - r_2}{ab} \right). \quad (7.52)$$

What is the physical interpretation of each fixed point in (7.52)? Show that the last fixed point is relevant only when $Kb - r_2 > 0$. Hint: If $Kb - r_2 \leq 0$, what does this say about the predator population?

How do each of K, b , and r_2 influence the truth of $Kb - r_2 > 0$, 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 4 Show that the \dot{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 . The draw appropriate arrows to indicate the horizontal motion of solutions in each region into which the nullcline divides the plane.

Modeling Exercise 5 Show that the \dot{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 . The draw appropriate arrows to indicate the vertical motion of solutions in each region into which the nullcline divides the plane.

Modeling Exercise 6 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 $r_1, r_2, a, b, K > 0$.

Modeling Exercise 7 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 ($Kb - r_2 > 0$). Hint: Show the eigenvalues of the Jacobian are given by

$$\lambda_1 = \frac{r_1 r_2}{2Kb} (-1 - \sqrt{1 - 4Kb(Kb - r_2)/(r_1 r_2)}), \lambda_2 = \frac{r_1 r_2}{2Kb} (-1 + \sqrt{1 - 4Kb(Kb - r_2)/(r_1 r_2)}).$$

Since we are assuming $Kb - r_2 > 0$ the discriminant $1 - 4Kb(Kb - r_2)/(r_1 r_2)$ under the square root is always less than 1.

Modeling Exercise 8 Based on your analysis above, sketch a typical phase portrait for this system under the assumption $Kb - r_2 > 0$. What is the long-term fate of each species?

The Case $Kb - r_2 < 0$

Modeling Exercise 9 Analyze the case in which $Kb - r_2 < 0$. What happens to each species' population in the long run?

7.4.3 Project: Parameter Estimation for Competing 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 (7.4)-(7.5) from data collected for two competing species of yeast.

The data in Table 7.2 comes from [44]. The data comes from several different experiments performed by Gause concerning the populations of two species of yeast, *Saccharomyces cerevisiae* and *Schizosaccharomyces kefir*. In each experiment a nutrient-filled vessel was inoculated with a fixed amount of either the *Saccharomyces* species, the *Schizosaccharomyces* species, or both. The population of each species was monitored at the listed times in column 1 of Table 7.2 (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 [44] for precise experimental procedures.

The data in column 2 of Table 7.2 are for the *Saccharomyces* species alone in the vessel, while column 4 is for the *Schizosaccharomyces* species alone in the vessel. Column 3 tabulates the *Saccharomyces* population when both yeast species are present, and Column 5 tabulates the *Schizosaccharomyces* 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 *Schizosaccharomyces* alone, or both, were grown and populations measured at the times 6, 16, ..., 141 hours listed in the table; these are the first 11 rows in Table 7.2. Then another set of three experiments was performed under identical conditions, with the results tabulated in the last 7 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

	<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 7.2: The growth of the yeast volume and the number of cells in pure cultures of *Saccharomyces cerevisiae* (column 1), *Schizosaccharomyces kefir* (column 3) and in the mixed population of these species (column 2 and 4 respectively). [44, p. 395]

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 7.2 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 *Schizosaccharomyces* species when alone, and finally the competition parameters a and b in (7.4)-(7.5). Gause had no computer to aid his analysis, but rather 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

$$\dot{u}_1 = r_1 u_1 (1 - u_1/K_1) \quad (7.53)$$

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 (7.53) with $u(0) = u_0$ is given by $u_1(t) = K_1 u_0 / (u_0 + e^{-r_1 t} (K_1 - u_0))$. However, we don't initial data at time $t = 0$ in Table 7.2, rather initial data $u(t_0) = u_0$. In this case the solution to (7.53) is

$$u_1(t) = \frac{K_1 u_0}{u_0 + e^{-r_1(t-t_0)} (K_1 - u_0)}. \quad (7.54)$$

Modeling Exercise 1 Use the data in column 2 of Table 7.2, along with (7.54), 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 3.5. Gause obtained estimates $r_1 = 0.21827$ and $K_1 = 13.0$, but these may not be the “best.”

Modeling Exercise 2 Repeat Modeling Exercise 1 for the *Schizosaccharomyces* species, using the data in column 4 of Table 7.2, to estimate the growth rate r_2 and carrying capacity K_2 for this species. Gause obtained estimates $r_2 = 0.06069$ and $K_1 = 5.8$, but your estimate may well be different.

Estimating the Competition Parameters a and b

Modeling Exercise 3 Fix the values for r_1, K_1, r_2 , and K_2 as you obtain them in Modeling Exercises 1 and 2. Use the data in columns 3 and 5 of Table 7.2 to estimate a and b in the equations (7.4)-(7.5). However, unlike the previous two Modeling Exercises, we do not have the luxury of a closed-form solution. A numerical solution to (7.4)-(7.5) may be required, in conjunction with plotting. Gauss estimated $a = 3.15$ and $b = 0.439$, but these estimates depend very much on his approach.

Modeling Exercise 4 Based on your estimates for r_1, K_1, r_2, K_2, a and b and the analysis summarized at the end of Section 7.3, 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. A Brief Introduction to Partial Differential Equations

Appendices

A. Complex Arithmetic

i.

Maybe remark or exercise on when roots of $ax^2 + bx + c = 0$ with $a, b, c > 0$ are both negative.

Make sure to include Fundamental Theorem of Algebra, roots conjugate.

Also, multiplicity of roots, poles.

Properties of conjugation, including exponentials.

B. Matrix Algebra Review

A matrix.

Invertible means nonzero eigenvalues.

Eigenvalues/vectors in complex conjugate pairs.

Linear independence of eigenvectors same as P invertible.

Diagonalization.

C. Circuits

Current, Voltage, and Resistance

Our 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 ODE's, by using the notion of *impedance*.

A First Example

Let's start with a very simple circuit containing a voltage source (also called an *electromotive force*) and resistor, as illustrated in Figure C.1. The main physical quantities of interest are the *current* through the circuit and the *electric potential* or *voltage* differences between any two points in the circuit. The current through a wire at a given position is simply the net rate at which electric charge is flowing past that position in some reference direction. The voltage between two points in a circuit is a measure of 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 the water flow rate, say in "mass per time," past a given 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's 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 that current consists of positive charges that flow from higher potential to lower potential, just

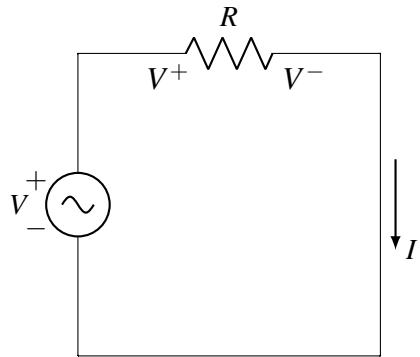


Figure C.1: Voltage source-resistor series circuit.

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 *amperes* or just “amps” and voltage/potential differences in *volts*.

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 ingredients to analyze the circuit of Figure C.1:

- Kirchhoff's Voltage Law:** The sum of all the voltage differences/changes around a closed loop in a circuit must be zero.
- 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.
- 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 *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 given 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 41.

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 resistance. A consequence of this assumption is that the voltage is the same at all points in a perfect conductor.

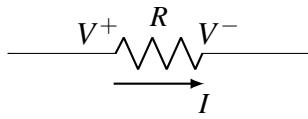


Figure C.2: Resistor schematic and Ohm's Law.

Circuit Analysis

To analyze the circuit of Figure C.1, let us designate one side of the voltage source (arbitrarily) as the “plus” side, the other is the “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 at 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 $V^- = 0$. 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.

It should be noted that if the voltage source is time dependent, so $V = V(t)$, the above analysis still holds. In this case the current through the circuit is also time-dependent and (C.2) becomes $V(t) = I(t)R$.

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 given by

$$q = CV \quad (\text{C.3})$$

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 41. The current $I(t)$ going into the left side of the capacitor equals the current

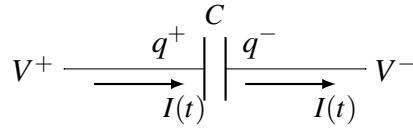
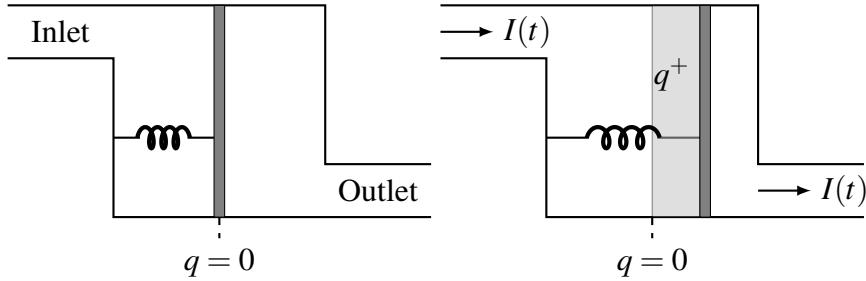


Figure C.3: Capacitor schematic.

Figure C.4: Mechanical/fluid analogue to a capacitor, uncharged (left) and charged to $q = q^+$ (right).

$I(t)$ leaving the right side; thus the net charge on the capacitor is always zero.

A mechanical/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 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 across the inlet/outlet.

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 has been applied across the inlet/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. Note that 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 across the inlet/outlet 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, note that in our mechanical/fluid analogy for the capacitor, the inlet and outlet are effectively reversible. This is true for many types of electrical capacitors, but not all!

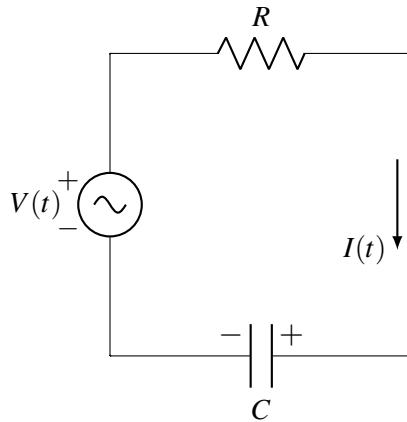


Figure C.5: Single loop RC series circuit.

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.

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 ($I > 0$ 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 equation (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" side of the source, designated as 0 V, and move clockwise around the circuit, back to our starting point. The net change in voltages over each component is $V(t) - RI(t) - q(t)/C$ (the voltage source, then the resistor, then the capacitor) which must be zero, so $V(t) - RI(t) - q(t)/C = 0$ or

$$RI(t) + \frac{q(t)}{C} = V(t). \quad (\text{C.5})$$

If we now make use of (C.4) we arrive at

$$Rq'(t) + \frac{q(t)}{C} = V(t) \quad (\text{C.6})$$

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 $q(t) = \frac{1}{200}(1 - e^{-125t})$. From equation (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.6) 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



Figure C.6: Inductor schematic.

$t \approx 5RC$ (note $1 - e^{-5} \approx 0.993$). The quantity RC has the dimension of time and is referred to as the “RC time constant” for the circuit. Note that the argument $-t/RC$ to 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 DE (C.6) becomes

$$Rq'(t) + \frac{q(t)}{C} = \cos(\omega t).$$

For an initial condition $q(0) = q_0$ the solution is

$$q(t) = \underbrace{De^{-t/(RC)}}_{\text{transient}} + \underbrace{A \cos(\omega t) + B \sin(\omega t)}_{\text{periodic}} \quad (\text{C.7})$$

where

$$D = q_0 - \frac{C}{1+C^2R^2\omega^2}, \quad A = \frac{C}{1+C^2R^2\omega^2}, \quad B = \frac{C^2R\omega}{1+C^2R^2\omega^2}.$$

The first term in (C.7) 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 given by

$$I_{per}(t) = -A\omega \sin(\omega t) + B\omega \cos(\omega t) = -\frac{C\omega}{1+C^2R^2\omega^2} \sin(\omega t) + \frac{C^2R\omega^2}{1+C^2R^2\omega^2} \cos(\omega t).$$

■

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 above is about $1/R$. That is, the current is given by $I_{per}(t) \approx V(t)/R$, 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.

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 $V = V^+ - V^-$ and current is given by

$$V = L \frac{dI}{dt} \quad (\text{C.8})$$

where L is the *inductance* of the inductor. Inductors are, in their simplest form, just coiled wire. The inductance depends on the size and geometry of the inductor, among other things. Equation (C.8) 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.8) to be dimensionally consistent,

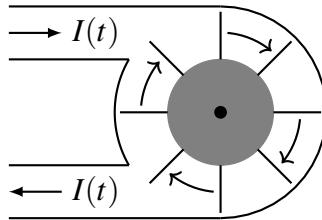


Figure C.7: Mechanical/fluid analogue to an inductor.

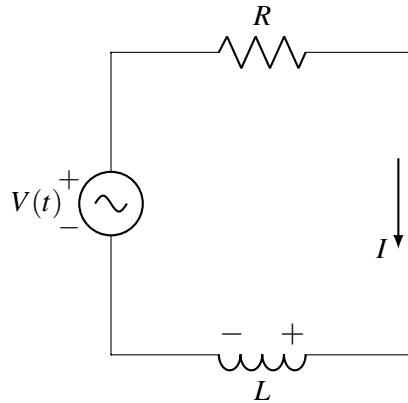


Figure C.8: Single loop RL series circuit.

inductance must have the dimension ML^2Q^{-2} . (Note we are using L for both inductance and the dimension of length, so 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 enclosed in a housing with an inlet and outlet as illustrated (the inlet/outlet are interchangeable). The wheel spins without friction. If the incoming fluid flow $I(t)$ 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 show that to decrease $I(t)$ we need to slow down the wheel, and apply a negative pressure differential (higher at the outlet, lower at the inlet) in order to accomplish this.

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.

We start of the “minus” side of the source, 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.8)) which must be zero, so $V(t) - RI(t) - LI'(t) = 0$ or

$$LI'(t) + RI(t) = V(t). \quad (\text{C.9})$$

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

$$0.001I'(t) + 10I(t) = 5$$

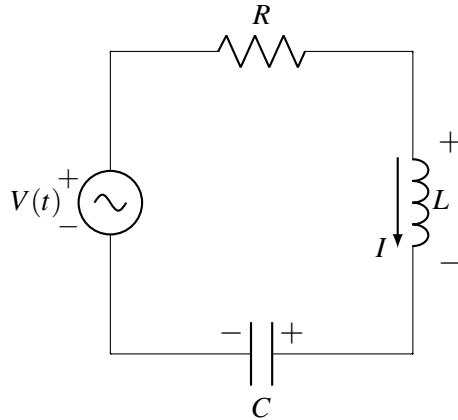


Figure C.9: Single loop RLC series circuit.

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 DE (C.9) becomes

$$LI'(t) + RI(t) = \cos(\omega t).$$

For an initial condition $I(0) = I_0$ the solution is

$$I(t) = \underbrace{De^{-Rt/L}}_{\text{transient}} + \underbrace{A \cos(\omega t) + B \sin(\omega t)}_{\text{periodic}} \quad (\text{C.10})$$

where

$$D = I_0 - \frac{R}{L^2\omega^2 + R^2}, \quad A = \frac{R}{L^2\omega^2 + R^2}, \quad B = \frac{L}{L^2\omega^2 + R^2}.$$

The first term in (C.10) 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 simply $I(t) = 1/R$, 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, 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.

RLC Circuits

Consider the circuit shown in Figure C.9, a single loop RLC circuit with the components in series.

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 we find, using Kirchhoff's Voltage Law and the voltage-current relations for each component, that $V(t) - RI(t) - LI'(t) - q(t)/C = 0$ or

$$LI'(t) + RI(t) + q(t)/C = V(t). \quad (\text{C.11})$$

Now use $I(t) = q'(t)$ (and so $I'(t) = q''(t)$) in (C.11) to obtain

$$Lq''(t) + Rq'(t) + q(t)/C = V(t), \quad (\text{C.12})$$

a second order, linear, constant coefficient DE very similar to the mass-spring equations of Chapter 4. Equation (C.12) needs two initial conditions, say of the form $q(0) = q_0$ and $I(0) = I_0$ where note that $I(0) = q'(0)$. We can solve (C.12) to find $q(t)$, from which we can compute the current $I(t) = q'(t)$ and then use (C.2), (C.3), and (C.8) to find the voltage across any of the components in the circuit at any given 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 DE (C.12) is

$$(5 \times 10^{-5})q''(t) + 2q'(t) + 10^6 q(t) = \cos(\omega t).$$

As with a damped spring-mass system, the solution will consist of a transient portion that satisfies the homogeneous DE, plus a periodic portion. The characteristic equation for the homogeneous DE 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 naturally oscillation frequency is determined by the imaginary parts 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.

Let's 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. Note that 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. ■

Practice Problems

Reading Exercise 195 A series RLC circuit has $L = 0.001$ henry, $R = 1$ ohms, $C = 0.00005$ farad). At time $t = 0$ the capacitor is uncharged and no current flows in the circuit. At time $t = 1$ a switch is closed that puts a 5 volt source in series with the other components. Set up and solve the relevant DE, then compute the current $I(t)$ through the circuit, and plot $I(t)$ from $t = 0.99$ to $t = 1.01$. Is this system over or underdamped?

Reading Exercise 196 Repeat Reading Exercise 195 if $V(t) = \cos(1000\pi t)$ (starting at time $t = 0$), and plot from $t = 0$ to $t = 0.02$. Identify the transient and periodic portions of $I(t)$, both visually and in the formula for the solution.

Reading Exercise 197 Repeat Reading Exercise 195, but suppose now that the voltage source is

$$V(t) = \begin{cases} 0, & t < 2 \\ 5, & 2 \leq t < 2.1 \\ 0, & t \geq 2.1 \end{cases}$$

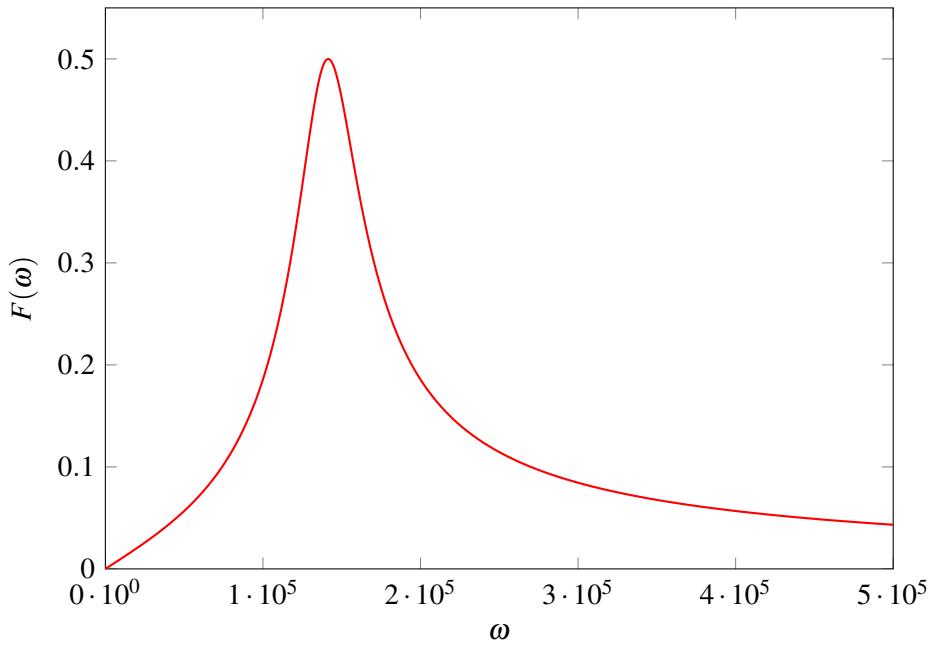


Figure C.10: Magnitude of periodic current through RLC circuit as a function of driving frequency.

Write down a DE for $q(t)$, the charge on the capacitor; it should involve a Heaviside function(s). Solve the DE and plot from $t = 1.95$ to $t = 2.15$.

Complex-Valued Solutions and Periodic Forcing

Up this point in the course when we've talked about 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

$$ax''(t) + bx'(t) + cx(t) = f(t) \quad (\text{C.13})$$

for some real numbers a, b, c , although what we're about to do works for linear constant coefficient ODE's 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.13) when $f(t)$ is complex-valued?

This means that the solution $x(t)$ will itself have a real and 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 . And it's easy to see 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 DE 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.13) but applied to the real and imaginary parts of $x(t)$ separately.

■ **Example C.6** Consider the DE

$$x'(t) + 2x(t) = f(t)$$

with $f(t) = t + ie^{-t}$ and initial condition $x(t) = 2 + 3i$. 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

$$x_r(t) = \frac{t}{2} - \frac{1}{4} + \frac{9}{4}e^{-2t} \quad \text{and} \quad x_i(t) = e^{-t} + 2e^{-2t}.$$

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.12), 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 as the forcing function f . Use $q'(t) = i\omega Ae^{i\omega t}$ and $q''(t) = -\omega^2Ae^{i\omega t}$ and substitute this information into the ODE, divide by $e^{i\omega t}$, and find

$$(-\omega^2L + i\omega R + 1/C)A = V_0$$

so that

$$A = \frac{V_0}{-\omega^2L + i\omega R + 1/C} \tag{C.14}$$

for the periodic response. This really is just 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.14) 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

$$q_r(t) \approx 1.0 \times 10^{-6} \cos(10000t) + 2.02 \times 10^{-8} \sin(10000t).$$

If we want the response with forcing $\sin(\omega t)$ we could take the imaginary part of $Ae^{i\omega t}$. ■

Impedance in Electrical Circuits

The notion of *impedance* generalizes that of resistance to circuits that contain capacitors, inductors, possibly other circuit elements. It quantifies more precisely 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.12), with voltage source $V(t) = V_0 e^{i\omega t}$. We have

$$Lq''(t) + Rq'(t) + q(t)/C = V_0 e^{i\omega t}. \quad (\text{C.15})$$

Differentiate both sides of (C.15) 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

$$LI''(t) + RI'(t) + I(t)/C = i\omega V_0 e^{i\omega t}. \quad (\text{C.16})$$

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 DE (C.16) (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

$$(-\omega^2 L + i\omega R + 1/C)I_0 = i\omega V_0. \quad (\text{C.17})$$

Finally, divide both sides above by $i\omega$ to obtain

$$\left(i\omega L + R + \frac{1}{i\omega C}\right)I_0 = V_0. \quad (\text{C.18})$$

Equation (C.18) 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.18) are not present), with $\omega = 0$ and V_0 and I_0 as real numbers. Then (C.18) becomes the familiar Ohm's law $V = IR$ —no ODE's required to find the current in the circuit. But even if the capacitor and inductor are present in the circuit, (C.18) 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 DE's.

The quantity R in (C.18) is, of course, the resistance 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 . Note 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.18) 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 $V = IZ$. When two components with impedances Z_1 and Z_2 are in parallel the resulting impedance Z obeys

$$\frac{1}{Z} = \frac{1}{Z_1} + \frac{1}{Z_2}$$

just like for resistors.

■ **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.18), $((20000)(5 \times 10^{-5})i + 2 + 1/((20000)(10^{-6})/i))I_0 = 5$ 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 ODE’s. 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 ODE’s. This is not unlike the machinery of the Laplace Transform, that allows us to reduce linear, constant coefficient ODE’s to simple algebra problems. If you want to know more about this stuff, take an AC circuits course!

Bibliography

- [1] Project MKULTRA. 2012 (accessed 2 April 2012). 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 8 December 2015). <https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2014/09/risky-decisions>.
- [3] https://www.youtube.com/watch?v=tzm_yyl13yo., (accessed 03 April 2020).
- [4] <https://www.youtube.com/watch?v=0ubvTOHWtms>, (accessed 03 April 2020).
- [5] Parke systems, active vibration isolation table. <https://parksystems.com/park-afm-options/active-vibration-isolation-table>, (accessed 03 May 2020).
- [6] SIMIODE Book Website, (accessed 04 October 2020). simiode.org.
- [7] Systemic Initiative for Modeling Investigations and Opportunities With Differential Equations (SIMIODE), (accessed 04 October 2020). simiode.org.
- [8] Wikipedia: Hydrogen Peroxide, (accessed 12 March 2010). http://en.wikipedia.org/wiki/Hydrogen_peroxide.
- [9] <https://www.youtube.com/watch?v=o-urnlaJp0A>, (accessed 13 May 2019).
- [10] <https://www.youtube.com/watch?v=V8W4Djz6jnY>, (accessed 15 July 2020).
- [11] <https://www.youtube.com/watch?v=kzVvd4Dk6sw>, (accessed 15 June 2020).
- [12] Mile high hearing. <https://milehighhearing.com/blog/types-of-hearing-loss>, (accessed 20 August 2018).

- [13] Taipei 101. <https://www.youtube.com/watch?v=B031ebdzRvc>, (accessed 26 August 2016).
- [14] Grand canyon skywalk. <https://www.youtube.com/watch?v=J0uoEXfS6ZA>, (accessed 27 August 2016).
- [15] 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 28 August 2019).
- [16] Podcast: Determining Time of Death, Air date 15 February 2015 (accessed 23 February 2020). <https://coronertalk.com/28>.
- [17] 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, 1964.
- [18] H. Akaike. A new look at the statistical model identification.
- [19] Christopher Arnold. *Chapter 4 in Designing For Earthquakes: A Manual for Architects, FEMA 454*. 2006. https://www.fema.gov/media-library-data/20130726-1556-20490-5679/fema454_complete.pdf.
- [20] Stealth Fighter Association. About the f117. <https://www.f117sfa.org/about-the-f117>, (accessed 22 August 2020).
- [21] Kendall Atkinson. *An Introduction to Numerical Analysis*. John Wiley and Sons, New York, second edition, 1989.
- [22] S. Axler. *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Springer, Princeton NJ, third edition, 2015.
- [23] G.I. Barenblatt. *Scaling*, volume 34. Cambridge Texts in Applied Mathematics, New York, second edition, 2003.
- [24] Athanassios Bissas, Josh Walker, Catherine Tucker, and Gorgios Paradisis. Biomechanical Report For The 100 m Women's IAAF World Championships, London 2017. <http://centrostudilombardia.com/wp-content/uploads/2018/10/1-100-donne.pdf>, (accessed 13 May 2019).
- [25] A. Björck. *Numerical Methods for Least Squares Problems*. SIAM, 1996.
- [26] Karen Bliss. 1-011A-T-Kinetics, 2015. <https://www.simiode.org/resources/850>.
- [27] M. Braun. *Differential Equations and Their Applications: An Introduction to Applied Mathematics*. Springer-Verlag, New York, fourth edition, 1993.
- [28] Kurt Bryan. Elementary inversion of the laplace transform. 1998 (accessed 29 July 2020). <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://www.simiode.org/resources/6424>.
- [31] Kurt Bryan. 3-095-S-ShotInWater, 2018. <https://www.simiode.org/resources/4498>.

- [32] Kurt Bryan. 1-092-S-DashItAll, 2019. <https://www.simiode.org/resources/6391>.
- [33] T. Carlson. Über geschwindigkeit und grösse der hefevermebrung in würze. *Biochem. Z.*, 57:313–334, 1913.
- [34] Communicable Disease Surveillance Center. News and notes: Influenza in a boarding school. *British Medical Journal*, 1(6112):586–590, 1978 (accessed 25 October 2014). <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1603269/pdf;brmedj00115-0064.pdf>.
- [35] C. W. Clark. *Mathematical bioeconomics: The optimal management of renewable resources*. John Wiley and Sons, New York, second edition, 1990.
- [36] J. Cloud. Was Timothy Leary Right? *Time Magazine*, 19 April 2007 (accessed 2 April 2012). <http://www.time.com/timemagazine/article/0,,9171,1612717,00.html>.
- [37] L. Corwin and R. Szczarba. *Calculus in Vector Spaces*. Marcel Dekker, New York, second edition, 1995.
- [38] M. Desmond. *Evicted: Poverty and Profit in the American City*. Broadway Books, New York, 2017.
- [39] W. Ding and S. Lenhart. Optimal harvesting of a spatially explicit fishery model. *Natural Resource Modeling*, 22(2):173–211, 2009.
- [40] Wandi Ding. 1-070-S-FisheryHarvest, 2015. <https://www.simiode.org/resources/1318>.
- [41] P. Dyke. *An Introduction to Laplace Transforms and Fourier Series*. Springer London, London, second edition, 2014.
- [42] Practical Engineering. What is a tuned mass damper? <https://www.youtube.com/watch?v=f1U4SAgy60c>, (accessed 27 August 2016).
- [43] R.L. Finney and D.E. Ostberg. *Elementary Differential Equations With Linear Algebra*. Addison-Wesley, Reading MA, 1968.
- [44] G.F. Gause. Experimental studies on the struggle for existence. *Journal of Experimental Biology*, 9(4):389–402, 1932.
- [45] G.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>.
- [46] G.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.
- [47] P. Goldberger. Architecture View; A Novel Design And Its Rescue From Near Disaster. *New York Times*, pages 45–53, 24 April 1988. <https://www.nytimes.com/1988/04/24/arts/architecture-view-a-novel-design-and-its-rescue-from-near-disaster.html>.
- [48] J. Hale. *Ordinary Differential Equations*. Dover, Mineola NY, 2009.

- [49] P. Hartman. *Ordinary Differential Equations*. SIAM Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, second edition, 2002.
- [50] Rogovchenko S.P. Hasanbulli, M. and Y.V. Rogovchenko. Dynamics of a single species in a fluctuating environment under periodic yield harvesting. *Journal of Applied Mathematics*, 2013. <https://doi.org/10.1155/2013/167671>.
- [51] K.L. Henold and F. Walmsley. *Chemicals: Principles, Properties, and Reactions*. Addison-Wesley, Reading MA, 1984.
- [52] Gudmund Hernes. The process of entry into first marriage. *American Sociological Review*, pages 173–182, 1972.
- [53] J. Hespanha. *Linear Systems Theory*. Princeton University Press, Princeton NJ, second edition, 2009.
- [54] R. Hilborn. *Overfishing: What Everyone Needs to Know*. Oxford University Press, Oxford, 2012.
- [55] R. Hilborn and C.J. Walters. *Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty*. Chapman and Hall, Inc., London, 1992.
- [56] A.V. Hill. The physiological basis of athletic records. *The Scientific Monthly*, 21(4):409–428, 1925.
- [57] P. Horowitz and W. Hill. *The Art of Electronics*. Cambridge University Press, Cambridge, England, third edition, 2015.
- [58] Motioneering Inc. Grand Canyon Skywalk and Custom TMDs. *At the Moment: Motion Control News and Views from Motioneering*, 5:1–2, 2006.
- [59] Tom Irvine. The Citicorp Building Tuned Mass Damper. *Vibration.com Newsletter*, pages 2–6, 2002. http://www.vibrationdata.com/Newsletters/January2002_NL.pdf.
- [60] H.R. Joshi, G.E. Herrera, S. Lenhart, and M.G. Neubert. Optimal dynamic harvest of a mobile renewable resource. *Natural Resource Modeling*, 22(2):322–343, 2009.
- [61] Erdi Kara. 1-126-S-MarriageMath-StudentVersion, 2020. <https://www.simiode.org/resources/7384>.
- [62] J.B. Keller. A theory of competitive running. *Physics Today*, 26(9):43, 1973.
- [63] S. Kemper. *Code Name Ginger: The Story Behind Segway and Dean Kamen's Quest to Invent a New World*. Harvard Business School Press.
- [64] Rose M. Kreider and Renee Ellis. Number, timing, and duration of marriages and divorces: 2009. *Current Population Reports, U.S. Census Bureau*, 2009. <https://www.census.gov/prod/2011pubs/p70-125.pdf>.
- [65] M. Kurlansky. *Cod: A biography of the fish that changed the world*. Penguin Books, London, 1998.
- [66] J.D. Lambert. *Numerical Methods for Ordinary Differential Systems: The Initial Value Problem*. Wiley, 1991.

- [67] Keith Alan Landry and Brian Winkel. 5-040-T-TunedMassDampers-Part I, 2016. <https://simiode.org/resources/2805>.
- [68] Keith Alan Landry and Brian Winkel. 5-040-T-TunedMassDampers-Part II, 2016. <https://simiode.org/resources/2807>.
- [69] O.A. Linares and A.L. Linares. Computational opioid prescribing: A novel application of clinical pharmacokinetics. *Journal of Pain and Palliative Care Pharmacotherapy*, 25:125–135, 2011.
- [70] 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. <https://doi.org/10.2514/6.2003-5762>.
- [71] Desmond M., A. Gromis, L. Edmonds, J. Hendrickson, K. Krywokulski, L. Leung, and A. Porton. Eviction Lab National Database: Version 1.0, (accessed on 12 August 2018). <https://www.evictionlab.org>.
- [72] 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.
- [73] M. McFarland. The Segway is Officially Over. *CNN Business*, 25 June 2020 (accessed 5 March 2021). <https://www.cnn.com/2020/06/23/tech/segway-pt-shut-down/index.html>.
- [74] C.M. Metzler. A mathematical model for the pharmacokinetics of LSD effect. *Clin Pharmacol Ther.*, 10(5):737–740, 1969.
- [75] Sheila Miller. 6-001-S-Epidemic, 2015. <https://simiode.org/resources/572>.
- [76] Lai Ming-Lai. Tuned mass damper. www.google.com/patents/US5558191, 1996.
- [77] Joe Morgenstern. The Fifty-Nine Story Crisis. *The New Yorker*, pages 45–53, 29 May 1995.
- [78] J. Murray. *Mathematical Biology*. Springer-Verlag, New York, second, corrected edition, 1993.
- [79] National Information Service for Earthquake Engineering (NISEE). The Earthquake Engineering Online Archive. <https://nisee.berkeley.edu>, (accessed 18 August 2016).
- [80] National Law Center on Homelessness and Poverty. Protect Tenants, Prevent Homelessness, (accessed on 12 August 2018). <https://www.nlchp.org>.
- [81] J. Norman. One-compartment kinetics. *British Journal of Anaesthesia*, 69:387–396, 1992.
- [82] R.Lynch Peastrel, M. and Jr. A. Armenti. Terminal velocity of a shuttlecock in vertical fall. *American Journal of Physics*, 48(7):511–513, 1980.
- [83] M. Peppi, A. Marie, C. Belline, and J.T. Borenstein. Intracochlear drug delivery systems: a novel approach whose time has come. *Expert Opin. Drug Deliv*, 15:319–324, 2018.
- [84] W.G. Pritchard. Mathematical models of running. *SIAM Review*, 35(3):359–379, 1993.
- [85] 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.

- [86] G. Simmons. *Differential Equations with Applications and Historical Notes*. Chapman and Hall/CRC, New York, third edition, 2017.
- [87] J. Switkes. A modified discrete sir model. *The College Mathematics Journal*, 34(5):399–402, 2003.
- [88] V. Tandon, W.S. Kang, T.A. Robbins, A.J. Spencer, E.S. Kim, M.J. McKenna, S.G. Kujawa, 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/>.
- [89] 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. <https://pubs.usgs.gov/of/2011/1053/>.
- [90] 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):123–144, Jul 2019.
- [91] TMC Vibration Control. Optical table advances quiet vibrations in highly sensitive applications. <https://www.techmfg.com/learning/whitepapers/optical-table-advances>, (accessed 24 August 2020).
- [92] Mary Vanderschoot. 5-026-S-Evictions, 2018. <https://simiode.org/resources/4872>.
- [93] 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.
- [94] Jue Wang. 1-138-S-InnerEarDrugDelivery, 2018. <https://www.simioide.org/resources/5068>.
- [95] Tracy Weyand. 1-136-S-MarriageAge, 2020. <https://simiode.org/resources/7693>.
- [96] Brian Winkel. Ants, tunnels, and calculus: An exercise in mathematical modeling. *Mathematics Teacher*, 87(4):284–287, 1994.
- [97] Brian Winkel. 1-006-S-FinancingSavingsAndLoans, 2015. <https://www.simioide.org/resources/278>.
- [98] Brian Winkel. 1-007-S-AntTunnelBuilding, 2015. <https://www.simioide.org/resources/289>.
- [99] Brian Winkel. 1-012-T-SublimationCarbonDioxide, 2015. <https://www.simioide.org/resources/435>.
- [100] Brian Winkel. 3-008-S-HangTime, 2015. <https://www.simioide.org/resources/304>.
- [101] Brian Winkel. 3-019-T-ShuttlecockFall, 2015. <https://simiode.org/resources/99>.
- [102] Brian Winkel. 5-001-T-LSDAndProblemSolving, 2015. <https://simiode.org/resources/395>.
- [103] Brian Winkel. 7-008-T-MachineReplacement, 2015. <https://simiode.org/resources/617>.

- [104] Brian Winkel. 1-061-T-PotatoCooling, 2016. <https://simiode.org/resources/2953>.
- [105] Brian Winkel. 6-040-StruggleForExistence, 2016. <https://simiode.org/resources/3115>.
- [106] Brian Winkel. 3-002-T-ModelsMotivatingSecondOrder, 2017. <https://www.simiode.org/resources/3428>.
- [107] World Health Organization. Deafness and hearing loss. <http://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>, 2018 (accessed 20 August 2018).
- [108] 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.
- [109] A.H. Zemanian. *Distribution Theory and Transform Analysis: An Introduction to Generalized Functions, with Applications*. Dover Publications, New York, 2010.
- [110] S.S. Zumdahl. *Chemical Principles*. D. C. Heath and Company, Lexington MA, 1992.

Index

- s -domain, 244
 - t -domain, 243
 - sgn function, 148
 - adaptive step sizing, 104, 105
 - Akaike Information Criterion, 118, 138
 - ambient temperature, 41
 - ansatz, 153
 - asymptotically stable equilibrium solution, 64
 - asymptotically stable fixed point, 64
 - asymptotically stable improper node, 380
 - asymptotically stable node, 378
 - asymptotically stable spiral point, 380
 - Atlantic cod, 18
 - autonomous, 61
 - auxiliary equation, 152
 - basis, 155
 - basis functions, 155
 - bifurcation, 68
 - transcritical, 69
 - bifurcation diagram, 69
 - bifurcation, pitchfork, 71
 - Bolt, Usain, 11
 - bolus, 227
 - Bromwich Integral, 243
 - butadiene, 83
 - carrying capacity, 19
 - center, 380
 - characteristic equation, 152
 - characteristic variable scales, 202
 - circuits
 - RC, 44
 - closed-loop control, 297, 298
 - closed-loop transfer function, 299
 - cochlea, 15
 - cohort, 132
 - commutative diagram, 238
 - compartmental model, 16, 43, 319
 - compartmental model, lsd, 359
 - competing species model, 368
 - complex conjugate, 169
 - compliance, 147
 - conjugate
 - complex, 169
 - conservation law, 18
 - constant coefficient, 38
 - constant coefficient first order system, 322
 - constant coefficient linear system, 328
 - control
 - closed-loop, 297, 298
 - feedback, 297, 298
 - open-loop, 296
 - PI, 301
 - PID, 303
 - proportional, 297
 - proportional-integral, 301
 - proportional-integral-derivative, 303
 - tuning, 302, 304

- control function, 293
 control theory, 232
 conventional current, 421
 convolution, 282, 283
 Convolution Theorem, 284
 Coulomb damping, 148
 critical point, 62
 critically damped, 165
 damped harmonic oscillator, 144
 damped pendulum, 226, 315
 damping
 - viscous, 142
 - dashpot, 142
 - deconvolution, 290
 - defective matrix, 335
 - deformation
 - elastic, 142
 - plastic, 142
 - derivative gain, 303
 - diagonalization, 353
 - difference equation, 79
 - differential equation
 - linear, 23
 - nonlinear, 23
 - differential equation, ordinary, 14
 - dimension, 26
 - dimensionless, 27
 - dinitrogen pentoxide, 84
 - Dirac mass, 272
 - direction field, 61, 372
 - disturbances, 297
 - driven harmonic oscillator, 143
 - eigenvalue, 328
 - eigenvector, 328
 - elastic deformation, 142
 - envelope, 197
 - equation
 - harvested logistic, 206
 - logistic, 204
 - equilibrium position, 142
 - equilibrium solution, 62, 369
 - asymptotically stable, 64
 - stable, 64
 - unstable, 64
 - error control, 104
 - error, Euler's Method, 93
 - Euler's Method, 90
 - Improved, 98
 - exchange of stability, 69
 - Existence-Uniqueness Theorem, 75, 94
 - exponential growth, 19
 - exponential order, 235
 - feedback control, 297, 298
 - filter
 - low-pass, 47
 - Final Value Theorem, 247
 - finite difference, 130
 - first order, 23
 - first order accuracy, 93
 - first order system, 321, 367
 - First Shifting Theorem, 242
 - fish, 18
 - fixed point, 62, 369
 - asymptotically stable, 64
 - stable, 64
 - unstable, 64
 - forced harmonic oscillator, 143
 - frequency domain, 244
 - fundamental matrix solution, 351
 - fundamental set of solutions, 155
 - gain, 193, 297
 - derivative, 303
 - integral, 301, 303
 - proportional, 301, 303
 - gain function, 193
 - general solution, 14, 24, 153, 155, 330
 - real-value, 332
 - global error, 109
 - growth rate, 19
 - half-life, 46, 227
 - hang time, 31
 - hard spring, 200
 - harmonic oscillator
 - damped, 144
 - driven, 143
 - forced, 143
 - pure, 143
 - undamped, 143
 - unforced, 143
 - Hartman-Grobman Theorem, 399, 400
 - harvesting, 20, 206
 - Heaviside function, 249, 253
 - henry, 147
 - Heun's Method, 100
 - Hill-Keller, 12

- homelessness, 361, 407
homogeneous, 38, 143, 328
homogeneous linear system, 328
Hook's law, 200
Hooke's Law, 142, 200
hydrogen peroxide, 82
hyperbolic equilibrium point, 400

impedance, 432
Improved Euler Method, 98
impulse, 267
impulse response, 286
impulsive, 229
incubator, 293
inductance, 147, 426
inductor, 147
initial condition, 14
initial value problem, 14
Initial Value Theorem, 246
input-output system, 280
integral equation, 311
integral gain, 301, 303
integrating factor, 38, 40
integrodifferential equation, 317
Intermediate Value Theorem, 72
inverse Laplace transform, 243
inverted pendulum, 316
investing, 33
irreducible, 242

Jacobian matrix, 400
jump discontinuity, 235

KISS philosophy, 129

Laplace s -domain, 244
Laplace transform, 339
 linearity, 235
Laplace transform, inverse, 243
Law of Mass Action, 81
least-squares, 112
least-squares estimation, 112
Leibniz notation, 51
linear, 23
linear first order system, 322
linear ordinary differential equation, 37
linear system
 constant coefficient, 328
linearization, 88, 397
linearly independent, 156, 353
loans, 78

logistic equation, 20, 204
logistic equation with harvesting, 20
Lotka-Volterra, 408
Lotka-Volterra competing species model, 368
low-pass filter, 47
LSD, 319

matrix
 defective, 335
matrix exponential, 349
maximum domain, 74
Mean Value Theorem, 72
method of undetermined coefficients, 175
money
 investing, 33
mortgage, 78

natural frequency, 162
Newton's Law of Cooling, 41, 42
Newton's Second Law, 12, 32
node
 asymptotically stable, 378
 asymptotically stable improper, 380
 unstable, 378
 unstable improper, 380
nondimensionalization, 204
nonhomogeneous, 38, 143, 328
nonhomogeneous general solution, 174
nonhomogeneous linear system, 328, 339
nonlinear, 23
nonlinear pendulum, 370
nullcline, 383

ODE, 14
open-loop control, 296
optimization, 120
order, 23, 93
order, second, 23
order,first, 23
ordinary differential equation, 14
 autonomous, 61
 constant coefficient, 38
 homogeneous, 38
 linear, 37
 nonhomogeneous, 38
 separable, 49
 time-invariant, 61
 variable coefficient, 38

parameter estimation, 21
parsec, 202

- partial differential equations, 41
 pendulum, 222, 224
 damped, 226, 315
 inverted, 316
 nonlinear, 370
 pharmacokinetics, 227, 319
 phase plane, 371
 phase portrait, 62, 385, 392
 phase space, 371
 PI control, 301
 PID control, 303
 piecewise continuous, 235
 pitchfork bifurcation, 71
 plant, 295
 plastic deformation, 142
 poles, 246
 Post Inversion Formula, 243
 predator-prey, 408
 principle of superposition, 153
 process variable, 294
 proportional control, 297
 proportional control gain, 297
 proportional gain, 301, 303
 proportional-integral control, 301
 proportional-integral-derivative control, 303
 pure harmonic oscillator, 143
 pure resonance, 194

 Q-factor, 199

 radian, 27
 rational function, 244
 RC circuits, 44
 RC time constant, 47
 reaction
 first-order, 82
 second-order, 83
 zeroth-order, 81
 reaction order, 81
 reaction rates, 81
 real-value general solution, 332
 reduction of order, 163
 reference signal, 294
 rescaling, 204
 residual, 115, 118
 residual sum of squares, 115, 118
 resonance, 186, 192
 pure, 194
 Runge-Kutta 4th order method, 102
 Runge-Kutta Methods, 102

 saddle point, 378
 salt tank, 339
 salt tank model, 43
 scaling, 204
 Schwarzschild radius, 30
 second order, 23
 Second Shifting Theorem, 255
 second-order reactions, 83
 Segway scooter, 315
 semi-stable, 65
 separable ODE, 49
 setpoint, 294
 Shifting Theorem, First, 242
 Shifting Theorem, Second, 255
 sifting property, 270
 sink, 65, 378
 SIR model, 370
 slope field, 61
 solution
 general, 153, 330
 solution, general, 14
 source, 65, 378
 spiral sink, 380
 spiral source, 380
 spring constant, 142
 stable equilibrium solution, 64
 stable fixed point, 64
 stable star point, 378
 star point
 stable, 378
 unstable, 378
 step response, 290
 sublimation, 127
 sum of squares, 119
 residual, 115, 118
 sum of squares function, 113
 superposition, 153
 system identification, 280, 287

 tangent line approximation, 88
 temperature, 42
 Theorem
 Convolution, 284
 theorem
 Existence-Uniqueness, 75, 94
 Existence-Uniqueness for systems, 324
 Final Value, 247
 Initial Value, 246
 time domain, 243
 time-invariant ODE, 61

- transcritical bifurcation, 69
- transfer function, 281
 - closed-loop, 299
- tuned mass damper, 362
- tuning, 302, 304
- two-compartment model, 320
- undamped harmonic oscillator, 143, 165
- underdamped, 159
- undetermined coefficients, 175, 341
- unforced harmonic oscillator, 143
- unit impulse response, 286
- unit step function, 249, 253
- unit-free, 29
- unstable equilibrium solution, 64
- unstable fixed point, 64
- unstable improper node, 380
- unstable node, 378
- unstable spiral point, 380
- unstable star point, 378
- variable coefficient, 38
- vector field, 61
- viscosity, 56
- viscous damping, 142