Texts in Applied Mathematics 65

David G. Schaeffer
John W. Cain

Ordinary Differential Equations: Basics and Beyond

Springer

Texts in Applied Mathematics

Volume 65

More information about this series at http://www.springer.com/series/1214

David G. Schaeffer • John W. Cain

Ordinary Differential Equations: Basics and Beyond

Springer

David G. Schaeffer
James B. Duke Professor
of Mathematics, Emeritus
Department of Mathematics
Duke University
Durham, NC, USA

John W. Cain
Mathematics Department
Harvard University
Cambridge, MA, USA

ISSN 0939-2475 ISSN 2196-9949 (electronic)
Texts in Applied Mathematics
ISBN 978-1-4939-6387-4 ISBN 978-1-4939-6389-8 (eBook)
DOI 10.1007/978-1-4939-6389-8

Library of Congress Control Number: 2016945045

Mathematics Subject Classification (2010): 34-01, 37-01, 00-01

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media LLC
The registered company address is 233 Spring Street, New York, NY 10013, U.S.A

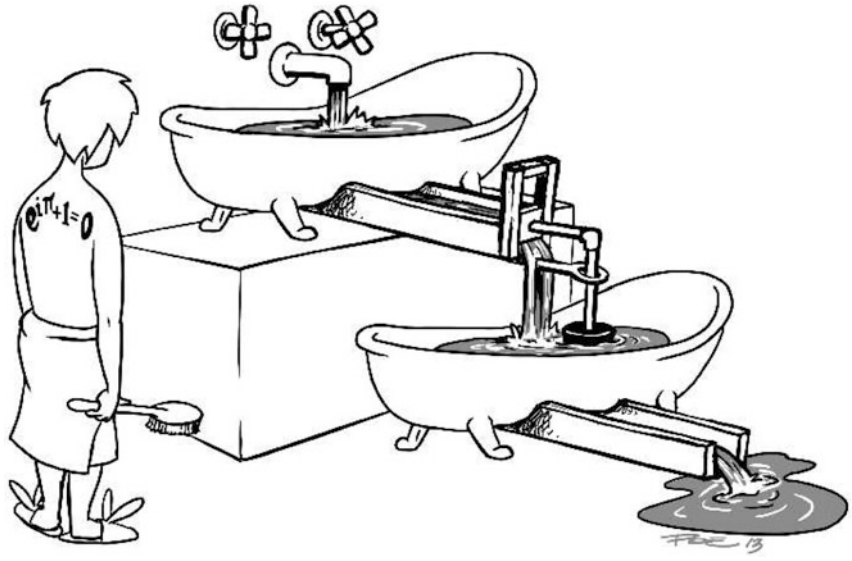

Figure 1: *Being literal about the phrase "bathtub model": A whimsical representation of the positive feedback in Sel'kov's model for glycolysis, equation (5.2). Water is added to the top tub at a constant rate, from which it drains into the lower tub. As the water level in the lower tub rises, the float in the tub also rises, which opens the gate wider and lets water drain faster from the upper tub. (Drawn by Jeff Poe, 2013. Copyright 2014 by D.G. Schaeffer and J.W. Cain.)*

To our wives and daughters...

...and to the many generations of students who struggled along with us as we taught the course and learned the subject (more or less in that order).

Preface

What's special about this book? In writing it we had four goals in mind:

1. To guide students through the theory of ordinary differential equations (ODEs) from an introductory level up through fairly advanced material at the graduate level, including bifurcation theory but not chaos, in a one-semester course;

2. to support the theory with meaningful examples of ODEs drawn from diverse other fields;

3. to enable students to read it profitably on their own,[1] outside of a course, with no more frustration than is inherent in learning any mathematics;

4. to make the book *enjoyable* to read.[2]

★ Regarding our first goal, you do not need any previous exposure to ODEs to read this book. The only prerequisites are some experience in the analytical foundations of calculus and a course in linear algebra. Moreover, you need not be completely comfortable with either of these subjects—we believe that applying them to ODEs, a task that demonstrates their utility, will help you to master them.[3]

While we succeeded in making the transition from introductory theory to advanced topics, we failed spectacularly in writing a book that can be covered in a one-semester course—we couldn't shut off the spigot. (In the "Notes to the Professor" below, we offer guidance about what topics may be cut to fit a one-semester course.)

★ Regarding our second goal, we apply the theory of ODEs to help the student understand equations from the following fields.

[1] Professors, please take note: we believe that because of this feature the book can be comfortably used in an *inverted classroom*. (See "Notes for the Professor" below.)

[2] Hey, we had fun writing it . . . mostly.

[3] And to facilitate your task, in Appendices B and C we review the information from analysis and linear algebra that we actually use.

- Biology: interacting-species population models, activator–inhibitor systems including a toy model for the Turing instability, the chemostat, Sel'kov's model for glycolysis, simple models for biological switches and clocks, some neurological models including the Morris–Lecar equations.
- Physics: various mechanical systems including systems based on the pendulum, Duffing's equation, the Lorenz equations.
- Chemistry: the Michaelis–Menten reaction rate for enzymatic reactions.
- Chemical engineering: the continuous stirred tank reactor.
- Electrical engineering: van der Pol's equation.

A knowledge of high-school science is sufficient preparation for these applications. Indeed, we believe that math students can learn much of the science through ODEs, transferring intuition from one application to another.[4]

★ Regarding our third goal: We draw on our years of teaching courses on this material to try to anticipate as many confusions of the beginning student as possible.[5] We supplement many of the exercises with long discussions that explain their significance. We include in the written text the kind of parenthetical remarks that typically are part of the classroom experience. We distinguish between dull tasks that really are necessary (yes, learning the subject does require some of these) and those that can be omitted or at least postponed.[6] This readability comes at some cost: it makes the book more wordy than the usual mathematician's minimalist style, and this fact might put off some readers.

★ Regarding our fourth goal, we write in an informal style, as though we were speaking to a student in the room with us. If being understandable and being rigorous are in conflict, we lean toward the former. We strive to give counterexamples that

[4]Please forgive us a folksy story in support of this point. A graduate student in biology was explaining his research to one of us. The student began by asking what he thought was a rhetorical question: "Should I give you the equations, or should I explain the biology?" He had already launched into the latter and was astounded when we answered, "Just give us the equations, we'll make up our own story to go along with it." And after having seen and discussed the equations, we found that our story was pretty much on target.

[5]Dear Reader, please let us know at

`http://www.math.duke.edu/ode-book/contact-us`

where we have failed in this attempt. We'll put a better explanation on our web site and, so twentieth-century, fix it in the second edition.

[6]We use the euphemism "a task for the dedicated reader" to describe the latter.

are as dramatic as possible.[7] We make connections to interesting scientific problems related to ODEs, such as the surprising fact that a pendulum can be stabilized in an inverted, straight-up, position by rapidly vibrating its pivot point. We even try for the occasional joke.[8] For example, we call the more advanced problems "PHD exercises," not because they relate in any way to the doctorate, but as an acronym for "piled higher and deeper." (If you are not familiar with the mildly vulgar misinterpretation of the usual sequence of degrees, B.S., M.S., Ph.D., from which this phrase derives, ask an older colleague.) Similarly, we call the notes at the end of the chapters "Pearls of wisdom." (Trust us, we don't *really* take ourselves that seriously.)

Here are some other pedagogical features of the book:

- A dedicated website

  `http://www.math.duke.edu/ode-book`

  containing software templates for solving many ODEs in this book, some problem solutions, supplementary readings, a list of known errors, and a link to contact us.

- Very detailed references within this book and to other sources, including specific sections or even specific pages, which can be easily followed.

- Consistent color conventions in figures[9] for nullclines, periodic orbits, stable and unstable manifolds, which enhance their value greatly.

- Great attention to consistent notation, including a detailed overview of our conventions in Appendix A.[10]

- An extensive collection of exercises, some of which we are vain enough to regard as creative, along with explanations of why they matter.

---

[7] For instance, the absolute value function is typically cited as an example of a Lipschitz continuous function on the line that is not differentiable (at the one point $x = 0$). However, in Section 3.6.1 we construct an example of a Lipschitz continuous function that is *not differentiable on any open interval in* $\mathbb{R}$. Besides being more interesting, the construction exercises the "analysis muscle."

[8] Attempted humor nipped in the bud: the publishers vetoed the title "Come romp with Dave and John across ODE land."

[9] If you are reading a black-and-white version of this book, color figures are available on the web site.

[10] We recommend that readers skim this appendix before starting the book and review it from time to time.

- An epilogue (Chapter 10) that gives an overview and references to ODE topics that continue the work begun in this book.

This book contains a great deal of information about properties of specific ODE models drawn from applications. Typically, after introducing an analytical technique, we illustrate it on one or more particular ODEs. The same equation may be used repeatedly in support of different techniques. For example, regarding activator–inhibitor systems, global existence is derived in Chapter 4 for such equations, scaling is used in Chapter 5 to reduce the equations to the simplest possible form, the stabilities of equilibria are determined in Chapter 6, the bifurcation of periodic solutions is analyzed in Chapter 8, and global bifurcation of solutions is considered in Chapter 9. This style of organization makes the theory less dry, but it also creates a problem: *how to learn what's known about any specific equation without paging through the whole book.* Well, here's a workaround: in the index under each specific ODE we list the pages where that equation is studied, and you can follow these references to focus on the behavior of a particular equation.

Notes for the Student

We are delighted that you are joining us for a journey through the study of ODEs, a subject that your authors love dearly. Part of the allure of ODEs is the "something for everyone" aspect of the field—whether you are theory-oriented or application-oriented, whether you prefer geometric reasoning or analytical reasoning, and whether you prefer computer-based or pen-and-paper-based calculations, your preferences will have a useful role in understanding ODEs.

At the same time, we urge you not to neglect the techniques you may find less congenial. You will achieve a deeper, more flexible, more portable understanding of the subject if you can synthesize theory, application, and numerics/asymptotics. Building on your current strengths, you can use this subject to enlarge your mathematical tool kit.

Let's talk about exercises, beginning with an inconvenient truth: *you don't understand the material in a chapter until you can do some of the exercises!* To help you derive maximal benefit from the exercises, we often include generous commentary on the significance of the problem. We tried to hold routine "apply the definition" exercises to a minimum. This makes the problems that remain more interesting, but it also means that lengthy hints may be needed (and are given).

In case you feel discouraged by the many pages of exercises following each chapter, let us point out that besides the obvious fact that you needn't do every problem, the problems are actually a lot shorter than they may appear—only a fraction of the exercise (in normal-sized type) actually asks you to do anything. Much of the physical space on the page is devoted to hints and commentary (in smaller type).

To help you focus your efforts, the exercises are divided into subsections as follows:

- core exercises,
- a group of exercises focused on a common theme,
- PHD exercises.

The core exercises give you a chance to practice using the basic ideas in the chapter; their specific purposes are listed at the start of the subsection. Focused subsections contain several related, not-to-be-neglected, exercises that may probe an area more thoroughly or may introduce ideas to be studied later. PHD exercises may be more difficult or develop some related topic not studied elsewhere in the text. (We won't refer to PHD exercises later in the text, except in other PHD exercises or in the Pearls, so these may be skipped without bad consequences.)

Besides the explicit exercises, every statement in the text is a sort of implied exercise—you should be able to prove these. (Sometimes we include a specific parenthetical remark like *"Prove this!"* or *"Why?"* at places where we think a reader might be tempted to move on without sufficient reflection, to the detriment of understanding.)

Notes for the Professor

This book is suited to a variety of courses. Because of its length, the most obvious fit is a two-semester graduate course covering the entire book, along with a supplemental foray into chaos.[11] However, to accommodate other options, large portions of the text could be cut without significant loss of continuity, such as:

→ Some longer proofs, like Sections 4.6.4–4.6.7 and 8.5.2.
→ The longer applications, like Sections 6.4, 8.6, and 9.8.
→ Perturbation-theory calculations, like Sections 7.5–7.7.
→ All appendices.
→ The discussion of scaling (Chapter 5).
→ The epilogue (Chapter 10).

[11] Appropriate chaos references are given in Section 10.6.

For additional savings, one could:

→ Assign Section 1.6 as independent reading.[12]

→ Cover fewer examples (of trapping regions in Sections 4.3 and 4.4 and of bifurcation in Chapters 8 and 9).

To make this quantitative: the book contains just over 300 pages of regular text.[13] What's left after the above cuts—a little over 150 pages of material[14]—can be covered in a reasonably paced one-semester course, perhaps with time left over for end-of-term projects or to reinstate some of the cut topics.

The book is also suited to a variety of students, not just the obvious audience of math graduate students. We believe that the level of exposition will be challenging, but still accessible, for strong upper-level undergrads in math,[15] all the more so if the primary focus is restricted to either theory or applications. The wealth of applications in the book should make it appealing to theoretically inclined graduate students in science or engineering.

In some semesters we used the book in an *inverted* classroom in which students were assigned reading before class and class time was freed up for more active learning. We liked the result.

To conclude, if you are weighing adoption of this text for a course at your institution, feel free to contact us (via the website) with questions regarding how, or whether, our book might serve the needs of your students.

Acknowledgments

We are greatly indebted to Tom Beale, Jim Nolen, and Steve Schecter, who tested early drafts of this book in their ODE courses. Please know that their careful reading and attention to detail has shielded you, dear reader, from *many* errors (not

[12]This section illustrates what the qualitative theory of ODEs, the focus of Chapters 6–9 of this book, can say about one specific ODE model. While students may benefit from a quick look at things to come, detailed understanding of the phenomena in this section will not be needed until Chapter 6.

[13]Plus roughly 65 pages of appendices, 35 pages of Pearls, and 110 pages of exercises.

[14]Moreover, the first four items on the cut list contain some of the densest writing in the book, so omitting them lightens the text more than a mere page count indicates.

[15]We are inordinately proud of our treatment of bifurcation theory, which leads us to make the following suggestion: a seminar, to follow a first course in ODEs, in bifurcation phenomena based on Chapters 8 and 9.

to mention patches of clunky exposition) that your authors may not have caught otherwise.

We are grateful to John Guckenheimer, whose feedback helped reshape our thinking about the book, and to Marty Golubitsky and Tom Witelski, who were most helpful consultants on multiple occasions. Thanks also to Michael Peper, the math librarian at Duke, for tracking down various obscure references we sent his way.

We are privileged to be acquainted with two talented artists who granted us permission to print/reprint several of their works. Jeff Poe sketched the frontispiece as well as three of the figures in Chapter 8. Fiona Ross allowed us to reprint her intricate example of a Jordan curve, appearing in Appendix B.

Bard Ermentrout facilitated creation of this book in ways he may not realize. His freely available XPPAUT software frequently saved the day when we needed to plot stable/unstable manifolds or bifurcation diagrams. Without XPPAUT, we shudder to contemplate how many hours would have been squandered writing our own (highly inferior) computer code to perform such computations.

Durham, NC, USA — David G. Schaeffer
Cambridge, MA, USA — John W. Cain

Contents

Chapter 1

Introduction

This chapter summarizes various ideas that an introductory course in ordinary differential equations (ODEs) usually covers. Because of this wide focus, the exposition may seem somewhat diffuse—a small price to be paid for our not requiring previous training in ODEs.

1.1 Some Simple ODEs

1.1.1 Examples

An *ordinary differential equation* (ODE) is an equation involving an unknown function of one variable and some of its derivatives. We hasten to assure you that this bland phrase has meaning for us primarily through examples, so let's proceed to these immediately.

Most simply,[1] we have the equation for exponential growth or decay,

$$x' = ax, \tag{1.1}$$

where the growth rate a is a constant (let's say real), $x(t)$ is the unknown function, and x' denotes the derivative of x with respect to t. The *logistic equation* modifies this equation, in case $a > 0$, by scaling the growth rate to unity and adding a negative term that limits growth as x becomes large:

$$x' = x - x^2. \tag{1.2}$$

The equation

$$x'' + x = 0 \tag{1.3}$$

[1] Well, $x' = f(t)$ is simpler, but this equation belongs to calculus, not ODEs.

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8_1

arises in what is called *simple harmonic motion*. In physical terms, which will be introduced in Section 1.4, equation (1.3) describes the motion of a mass pulled back toward equilibrium by a frictionless spring. Here, of course, x'' denotes the second derivative. A useful point of comparison for (1.3) is

$$x'' + \sin x = 0, \tag{1.4}$$

which describes the motion of a frictionless pendulum under gravity, given some simplifying assumptions about units (cf. Section 1.4). Two other noteworthy modifications of (1.3) are

$$\text{(a) } x'' - tx = 0 \quad \text{and} \quad \text{(b) } x'' + (\kappa + 2\varepsilon \cos t)x = 0, \tag{1.5}$$

known as *Airy's equation* and *Mathieu's equation*, respectively.

We conclude our first round of examples with a Riccati equation

$$x' = x^2 - t \tag{1.6}$$

and a purely pedagogical example

$$1 + (x')^2 = x^2. \tag{1.7}$$

1.1.2 Descriptive Concepts

The most basic concept used to describe ODEs is *order*, which refers to the order of the highest derivative that appears in the equation. Thus equations (1.1), (1.2), (1.6), and (1.7) are of first order, while (1.3), (1.4), and (1.5a,b) are of second order. Here is an example of a third-order equation:[2]

$$\frac{d^3y}{dx^3} = \frac{\alpha y + \beta}{y^3}, \tag{1.8}$$

where $y(x)$ is the unknown function of the variable x. This example also illustrates the following three points: (i) Usually, the independent variable in the ODEs we study is time, but other choices also occur—in this equation, x represents a spatial coordinate. (The dependent variable $y(x)$ represents the thickness of a thin film as a function of position.) (ii) We have written the derivative using the d/dx-notation rather than with primes; no mathematical significance should be attached to this choice; it is only a matter of taste as to which notation seems more attractive to us in a given situation (at the time of writing). (iii) An ODE need not be defined for

[2]We resist the temptation to dazzle you with interesting higher-order ODEs drawn from numerous fields. Truth to tell, higher-order equations play a smaller role in the theory than *systems* of several coupled ODEs, which we introduce in Section 1.5.

all values of either the dependent or independent variable. For example, (1.8) is not defined for $y = 0$. Incidentally, much information can be gained by focusing on such exceptional points of an equation—called *singularities* in the usual terminology.

Normally we will solve for the highest derivative of the dependent variable as a function of lower-order derivatives and of t: thus, for an equation of order n, we obtain

$$x^{(n)} = f(x, x', \ldots, x^{(n-1)}, t). \tag{1.9}$$

The value of this convention is illustrated by equation (1.7), which may be rewritten

$$x' = \pm\sqrt{x^2 - 1}. \tag{1.10}$$

Several problems in (1.7) become evident from rewriting the equation in this way: (i) Two different ODEs are in fact hidden in (1.7); to get an unambiguous ODE, we need to choose between the plus and minus signs in (1.10). (ii) For some values of x—specifically for $|x| < 1$—equation (1.7) has no real-valued solutions. (iii) The right-hand side (RHS) of the equation, $\sqrt{x^2 - 1}$, is not a differentiable function. (Item (iii) has unpleasant consequences—see Exercise 4(b).)

Next we define the vitally important notion of linearity. We shall call an ODE *linear*[3] if it may be written in the form

$$x^{(n)} = a_1(t)x^{(n-1)} + a_2(t)x^{(n-2)} + \ldots + a_{n-1}(t)x' + a_n(t)x + g(t); \tag{1.11}$$

i.e., the unknown function x and its derivatives appear only raised to the first power. Thus equations (1.1), (1.3), and (1.5a,b) are linear. Equations (1.2), (1.6), and (1.7) are nonlinear; perhaps less obviously, (1.4) is also nonlinear, because

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \ldots$$

has higher powers of x hidden in it. Note that $x'' = x'x$ is also nonlinear, because of the product on the RHS of the equation.

The linear equation (1.11) is called *homogeneous* if $g(t) \equiv 0$, and it is said to have *constant coefficients* if the functions $a_j(t)$, $j = 1, \ldots, n$, actually do not depend on t. A closely related concept for nonlinear equations: (1.9) is called *autonomous* if the function f does not depend on t. Solutions of an autonomous equation are translationally invariant: if $x(t)$ is a solution of $x^{(n)} = f(x, x', \ldots, x^{(n-1)})$ for some time interval, then for every constant t_0, the shifted function $\tilde{x}(t) = x(t - t_0)$ is also a solution.

[3]More formally, we may say that equation (1.9) is linear if the function f is linear in its first n arguments; no restriction on the t-dependence is implied.

1.2 Solutions of ODEs

1.2.1 Examples and Discussion

Being mathematicians, we must define our terms, even when the definition is more tautological than informative, as in the following: A function $x(t)$ is called a *solution* of (1.9) if the two sides of the equation become equal when this function is substituted into the equation. An understanding of this term will emerge from the examples.

As you may readily check, for every constant C, $x(t) = Ce^{at}$ is a solution of the first equation (1.1). (In Exercise 1(a) we show you how to prove that this is the most general solution of (1.1).) The general solution of the second equation (1.2) will be determined in Section 1.3.3 below, using a technique introduced there.

Regarding the third equation (1.3), for constants $C_1, C_2 \in \mathbb{R}$, the formula

$$x(t) = C_1 \cos t + C_2 \sin t \tag{1.12}$$

provides a solution. This solution provides an instance of the principle of *linear superposition* for a linear homogeneous ODE. Specifically, if $x_1(t)$ and $x_2(t)$ are solutions of (1.11) (with $g(t) \equiv 0$), then for constants C_1, C_2, the linear combination

$$C_1 x_1(t) + C_2 x_2(t)$$

is also a solution. In the language of linear algebra, the set of solutions of a homogeneous linear ODE forms a vector space. As we shall see in Chapter 2, *every* solution of (1.3) can be written in the form (1.12); this means that the set of solutions of (1.3) is a two-dimensional vector space for which $\{\cos t, \sin t\}$ is a basis.

The above examples of solutions illustrate one of the most fundamental points in the whole subject: *ODEs have infinitely many solutions.* Thus, some auxiliary information must be given to pick out exactly one solution from the infinite set of solutions. The most common such auxiliary information is an *initial condition* (IC). For example, if one seeks a solution of (1.1) subject to the auxiliary condition

$$x(0) = b,$$

where b is a real constant, then $x(t) = be^{at}$ is the *unique* solution of this more specific problem. (To *show* that it is unique, we need to know that Ce^{at} is the general solution of (1.1), as you prove in Exercise 1(a).) Similarly, regarding (1.3), given real constants b_0, b_1, there is a unique function of the form (1.12) that satisfies the initial condition

$$x(0) = b_0, \quad x'(0) = b_1.$$

For a general equation, given constants $b_0, b_1, \ldots, b_{n-1}$, the *initial value problem* (IVP) seeks a solution of (1.9) such that

$$x(0) = b_0,\ x'(0) = b_1,\ \ldots\ ,\ x^{(n-1)}(0) = b_{n-1}. \tag{1.13}$$

The initial conditions may be imposed at any point $t = t_0$, but we will usually impose these conditions at $t = 0$ as in (1.13). Of course the general case is easily reduced to (1.13) without loss of generality.

We can exhibit the general solutions of equations (1.1)–(1.3). However, it is a rare and pleasant occurrence when one is able to find *any* explicit solution of an ODE. In Section 1.3 we shall describe three classes of equations that can be solved explicitly.

Warning: The symbol x may refer simply to a generic real variable, or it may refer to a function $x(t)$ that satisfies an ODE. Both usages may occur in the same sentence! For example, the first usage is intended in the italicized, initial part of the sentence, "*If $f(x)$ is not a differentiable function*, then uniqueness issues for the equation $x' = f(x)$ may be quite subtle," while the second usage is intended in the remainder of the sentence. Also, in connection with writing arguments of a function, please read Note 1 in Section A.2.

1.2.2 Geometric Interpretation of Solutions

The geometric interpretation of ODEs is an essential part of the subject. The interpretation is clearest for first-order equations. We illustrate this with the help of Figure 1.1, which shows the direction field for the Riccati equation (1.6). Imagine that at every point in the t, x plane a line segment with slope $x^2 - t$ is drawn. A function $x(t)$ is a solution of (1.6) iff at every point $(t, x(t))$, the graph of this function is tangent to the line segment at that point. This interpretation makes it seem natural that ODEs have many solutions and that a unique solution may be selected by specifying a starting point for the curve at $t = 0$.

Besides the direction field, Figure 1.1 also shows several solutions of the ODE (1.6). Although we have used the computer to draw these solutions, useful approximations may be sketched by hand by choosing a starting point $t = 0$, $x(0) = b$ and drawing a curve that is everywhere tangent to the direction field. *(Try this yourself!)*

From studying the curves shown in Figure 1.1, we make the following conjecture:

Conjecture 1.2.1. *(i) If the initial condition $x(0)$ is large and positive, a solution of (1.6) grows without bound as t increases. (ii) If $x(0)$ is negative, or positive but not too large, the solution is asymptotic to a curve in the half-plane $\{x < 0\}$.*

See Section 1.8.3 to pursue this conjecture analytically.

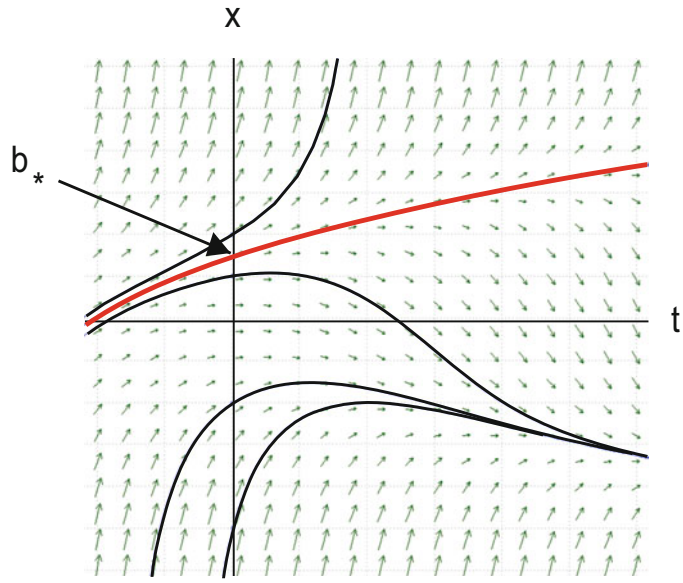

Figure 1.1: *Direction field for the Riccati equation* (1.6)*, with several solution trajectories corresponding to different choices of initial condition* $x(0)$*.*

1.3 Snippets of Solution Techniques

1.3.1 Computer Solutions

If your primary interest in ODEs is in applications, numerical solutions are more important than any analytical solution technique. Even if your primary interest is theory, it is foolish not to take advantage of the numerical solutions that are readily available with the computer. On frequent occasions in this book we will expect you to compute numerical solutions. Any software, commercial or free, that you wish to employ will probably be adequate. Moreover, on the book's website we introduce you to one specific (free) software package—please visit

`http://www.math.duke.edu/ode-book/computing/intro` .

In particular, we include code for solving many of the equations in this chapter. We also include some words of caution about what can go wrong in numerical work.

Get started on the computer right away! For example, in Exercise 9 we invite you to verify Conjecture 1.2.1 and to locate the special initial condition $x(0) = b_*$ at the boundary between the two different asymptotic behaviors.

In the remainder of this section we introduce three analytical methods for finding explicit solutions of certain ODEs. These methods are part of the elementary theory of ODEs, and we shall assume that you can use them when needed. No other ideas from the elementary theory will be assumed.

1.3.2 Linear Equations with Constant Coefficients

If the coefficients a_j in (1.11) are independent of t and $g(t) \equiv 0$, then one may find explicit solutions of this equation using exponentials. This method will be developed extensively in Chapter 2, so here we limit ourselves to applying it in a simple example,

$$x'' + \beta x' + x = 0. \tag{1.14}$$

(This equation differs from (1.3) by the first-order term $\beta x'$. As discussed in Section 1.4, this new term represents friction, and normally $\beta > 0$.) Let's look for solutions of (1.14) of the form $x(t) = e^{\lambda t}$. Substituting into (1.14), we see that $e^{\lambda t}$ satisfies (1.14) if

$$(\lambda^2 + \beta\lambda + 1)e^{\lambda t} = 0. \tag{1.15}$$

In other words, because the derivative of the exponential is a multiple of itself, finding an exponential solution of (1.14) reduces to solving the algebraic equation $\lambda^2 + \beta\lambda + 1 = 0$. This polynomial, which is called the *characteristic polynomial* of (1.14), has roots

$$\lambda_\pm = \frac{-\beta \pm \sqrt{\beta^2 - 4}}{2}.$$

Thus, $e^{\lambda_\pm t}$ is a solution of (1.14), and by linear superposition, for any constants C_+, C_-,

$$C_+ e^{\lambda_+ t} + C_- e^{\lambda_- t} \tag{1.16}$$

is also a solution. (If $\beta = 2$, then $\lambda_+ = \lambda_-$, so these two terms are multiples of each other; see Exercise 2 for how to generalize (1.16) in this case.) As follows from our work in Chapter 2, apart from the exceptional case $\beta = 2$, (1.16) is in fact the general solution of (1.14).

If the friction coefficient β is positive, then the solutions $e^{\lambda_\pm t}$ decay as t increases. If $0 < \beta < 2$, then the roots $\lambda_\pm$ are complex. In this case, we may separate the real and imaginary parts of the roots,

$$\lambda_\pm = -\beta/2 \pm i\sqrt{1 - \beta^2/4},$$

and use Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$ to rewrite the solution

$$e^{\lambda_+ t} = e^{-\beta t/2}\left(\cos\sqrt{1 - \beta^2/4}\, t + i\sin\sqrt{1 - \beta^2/4}\, t\right)$$

and similarly for $e^{\lambda_- t}$. Alternatively, we may form linear combinations of the complex-valued solutions $e^{\lambda_\pm t}$ to produce *real*-valued functions that constitute a different basis for the set of solutions of (1.14):

$$e^{-\beta t/2}\cos\sqrt{1 - \beta^2/4}\, t, \qquad e^{-\beta t/2}\sin\sqrt{1 - \beta^2/4}\, t.$$

This calculation illustrates a tension that exists in this text: usually, we are interested in real-valued solutions of an ODE, but often it is convenient to consider complex-valued solutions in order to take advantage of the complex exponential. In general, as here, a complex exponent indicates oscillatory behavior of real-valued solutions of an ODE.

1.3.3 First-Order Linear Equations

For a first-order linear ODE, say

$$x' + a(t)x = g(t), \tag{1.17}$$

one can find explicit solutions even if the coefficients are variable. In Exercise 1(b) we ask you to verify the following claim: Let $\bar{a}(t) = \int_0^t a(s)\,ds$; then for every constant C,

$$x(t) = Ce^{-\bar{a}(t)} + \int_0^t e^{\bar{a}(s)-\bar{a}(t)}\, g(s)\, ds \tag{1.18}$$

satisfies (1.17). Moreover, since $x(0) = C$, (1.18) also provides a solution to the initial value problem. (This solution is in fact unique, which will fall out from the general theory below.)

For interested readers: the derivation of (1.18) is based on asking whether you can multiply (1.17) by some function $\phi(t)$ such that the left-hand side (LHS) becomes an exact derivative. Well, yes, choose $\phi(t) = e^{\bar{a}(t)}$ and observe that

$$e^{\bar{a}(t)}\,[x' + a(t)x] = \frac{d}{dt}[e^{\bar{a}(t)}\,x].$$

The function $\phi(t)$ is called an integrating factor (cf. Section 1.2 of [10].). The same idea is the basis of the proof in Exercise 1(a) that Ce^{at} is the general solution of (1.1).

1.3.4 Separable Equations

A first-order ODE is called *separable* if the RHS may be factored as

$$x' = f(x)g(t). \tag{1.19}$$

For example, equations (1.2) and (1.7) are separable,[4] where in both cases, the factor $g(t)$ is trivial. On the other hand, (1.6) is not separable. Let's illustrate how to exploit separability by solving (1.2). In the following derivation, we temporarily suspend concerns of rigor—we shall freely perform manipulations that might seem

[4]Equation (1.1) is also separable, but usually this term is reserved for nonlinear equations.

problematic in order to obtain a formula for the solution. After it has been derived, we may verify that the formula actually does provide a solution. Given such a verification, there is no need to justify questionable intermediate steps, even though this may be possible.

We write the LHS of (1.2) using the notation $x' = dx/dt$, and we treat dx and dt as separate factors. Let's bring all x-dependence in the equation to the LHS and all t-dependence to the RHS, obtaining

$$\frac{dx}{x - x^2} = dt. \tag{1.20}$$

The LHS of (1.20) may be expanded in partial fractions:

$$\frac{1}{x - x^2} = \frac{1}{x} + \frac{1}{1 - x}.$$

Integrating (1.20), we derive

$$\ln|x| - \ln|1 - x| = t + C,$$

where C is an arbitrary constant of integration. Exponentiation of this equation yields

$$\left|\frac{x}{1 - x}\right| = C'e^t, \tag{1.21}$$

where $C' = e^C$. According to (1.21), $x/(1-x)$ cannot pass through zero or infinity, so it cannot change sign. We may remove the annoying absolute value signs by defining $C'' = +C'$ if $x/(1 - x)$ is positive and $C'' = -C'$ if this quantity is negative. Then (1.21), thus simplified, may be solved for $x(t)$:

$$x(t) = \frac{C''}{C'' + e^{-t}}. \tag{1.22}$$

In Exercise 1(c) we ask you to verify that the above manipulations actually produce solutions of (1.2).

To solve the IVP, we seek a value of C'' in (1.22) that will satisfy the initial condition $x(0) = b$. Please check that apart from the problematic value $b = 1$, the initial condition is satisfied if and only if

$$C'' = \frac{b}{1 - b}. \tag{1.23}$$

It is ironic that (1.22) runs into trouble precisely in the case that the original ODE has the trivial solution $x(t) \equiv 1$. This behavior arises from one of the gaps in rigor

in the above derivation: if $x(t) \equiv 1$, then the term $dx/(1-x)$ is undefined and hence cannot be integrated.[5]

Note that if $C'' < 0$, the solution (1.22) is not defined for all t. This is a warning: an IVP may have a solution *only for a finite time.* (We explore this issue in a more serious way beginning in Chapter 3.)

Despite the above nonexistence problem, (1.22) gives acceptable predictions regarding the future evolution of a population, for which $x \geq 0$. Specifically, in Exercise 1(c) we ask the you to show that if $b \geq 0$, then the solution of the IVP—obtained from (1.22), (1.23) if $b \neq 1$ and from $x(t) \equiv 1$ if $b = 1$—exists for all $t \geq 0$. You may also verify that the solution $x(t)$ tends to 1 as $t \to \infty$. (In Exercise 6, we ask you to extract this latter conclusion geometrically from the direction field.)

1.4 Physically Based Second-Order ODEs

Many phenomena in ODEs can be seen already in certain second-order ODEs that arise in physical applications. In this section we explain the derivation of a few of these historically important equations, which are a rich source of intuition about ODEs. Some math students resist the introduction of such material into an ODE course with rationalizations like, "I never could understand physics." If you buy into such sentiments, we urge you to get beyond these self-limiting preconceptions. Although this material may seem like a distraction, the gain in insight from a modest investment is enormous.

1.4.1 Linear Spring–Mass Systems

The motion of spring–mass systems is governed by Newton's second law of motion,

$$\text{mass} \times \text{acceleration} = \text{sum of all forces},$$

or more compactly and more famously, $F = ma$. Consider for example the system illustrated in Figure 1.2. The mass is constrained to move along a single axis. If we let x measure the displacement of the mass from a reference position, then the acceleration is simply the second derivative d^2x/dt^2.

There are two forces acting on the mass: (i) the restoring force F_{spring} from the spring and (ii) the drag F_{friction}. The restoring force opposes any displacement from the equilibrium position. The simplest assumption, called *Hooke's law*, is that the force is proportional to the displacement from equilibrium. If we measure

[5]Observe that $x \equiv 0$ also satisfies (1.2). Thus, in the derivation, dx/x is likewise meaningless for this solution, but the final answer nevertheless captures the solution. This behavior reminds us that solutions obtained using separability always need to be examined carefully.

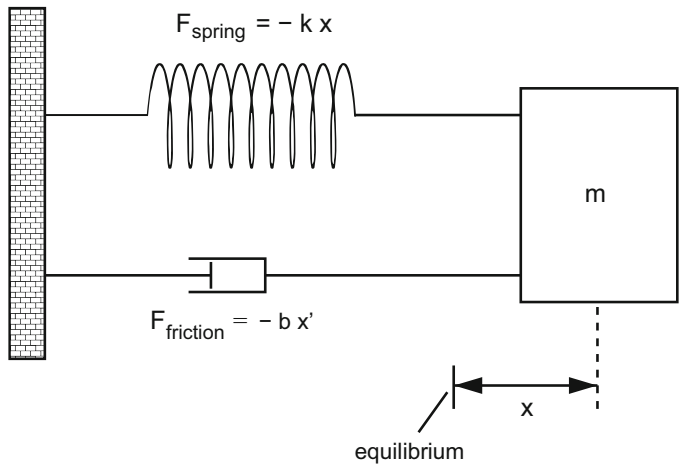

Figure 1.2: *Schematic diagram of the mass–spring–dashpot system corresponding to* (1.24)*. The spring pulls the mass back toward its equilibrium at* $x = 0$*, i.e., to the left for the displacement shown in the figure. Friction acts in the direction opposite the current velocity of the mass, which is not indicated in the figure.*

displacements relative to the equilibrium position, we have the simple formula for the restoring force $F_{\text{spring}} = -kx$, where k is the constant of proportionality.

While the restoring force depends on the *displacement*, the drag depends on the *velocity*, the derivative of the displacement. The simplest assumption is that this force opposes the motion of the mass with a strength proportional[6] to its speed: in symbols, $F_{\text{friction}} = -b\,dx/dt$, where $b \geq 0$. Truth to tell, this formula is a rather poor approximation of dry friction,[7] i.e., friction of a mass sliding over a dry surface. Despite this inaccuracy, friction is widely approximated by such a term, because of the appealing fact that this leads to linear ODEs.

Combining the above forces in Newton's equation, we get the ODE for the motion of the mass

$$mx'' = -bx' - kx. \tag{1.24}$$

Equation (1.14) is a special case of this equation.[8] As for (1.14), the general solution of (1.24) is a linear combination of exponentials (apart from the exceptional

[6]You may groan that we use the same letter b for the constant of proportionality as for the initial conditions. Please trust us—in the larger scheme of things, this conflict of notation will not cause confusion.

[7]This formula is more typical of the drag from a viscous fluid at low to moderate velocities—see Section 5.4 in [72] or look online. Reflecting this situation, in Figure 1.2 friction is represented by a "dashpot": i.e., a piston sliding through a viscous fluid.

[8]In fact, the more general equation (1.24) can be reduced to (1.14) by scaling the variables appropriately. The uses of scaling will be developed systematically in Chapter 5.

case $b^2 = 4mk$). Substituting into the equation, we find that $e^{\lambda t}$ is a solution of (1.24) if

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4mk}}{2m}. \tag{1.25}$$

This formula for the roots contains interesting information about how friction changes the behavior of the system. If there is no friction (i.e., $b = 0$), the roots (1.25) are pure imaginary, and the solutions $e^{\lambda t}$ of (1.24) are trigonometric functions with (angular) frequency $\omega = \sqrt{k/m}$; thus, oscillations continue forever. As b increases from zero, these exponential solutions retain their oscillatory character, with some decrease in the frequency, but are confined within a decaying exponential envelope. This behavior continues as b increases, the decay becoming more rapid, until $b^2 = 4mk$. After this point, both roots (1.25) are real, and the solution $x(t)$ will cross $x = 0$ at most once in the course of its decay. The cases $b^2 < 4mk$, $b^2 = 4mk$, and $b^2 > 4mk$ are called *underdamped*, *critically damped*, and *overdamped*, respectively.

These ideas have a practical consequence in the automotive world. The shock absorbers of a car can be crudely modeled by (1.24). As the name suggests, one wants shock absorbers to have a lot of damping, i.e., to be overdamped. This gives rise to the following quick test for whether shock absorbers are worn out. Depress the car and release it from rest. If the car returns monotonically to equilibrium (overdamped), then the shock absorbers are OK. If, on the other hand, the car oscillates up and down during its return to equilibrium (underdamped), then the shock absorbers need to be replaced.

A few words of reassurance in case you are feeling uncomfortable with this level of mechanics. You have a perfectly adequate grasp of the physics when you can combine the following three themes in your head:

- Intuition about the response of the system on the level of, "If I pull on the mass and release it, it will drift back toward equilibrium, possibly with oscillations."
- The derivation of (1.24) from Newton's law, including the formulas for the forces.
- Finding exponential solutions of (1.24) and seeing that the behavior of these solutions conforms with your intuition.

1.4.2 Nonlinearity, Part I: The Restoring Force

It is easy to imagine springs in which the restoring force is not exactly proportional to the displacement; indeed, exact linearity is the unlikely behavior. For example, although the restoring force is gravity rather than a spring, consider a pendulum as illustrated in Figure 1.3(a). The mass is confined to move on a circle by a rigid (massless) arm, say of length ℓ, and its position is specified by a single coordinate,

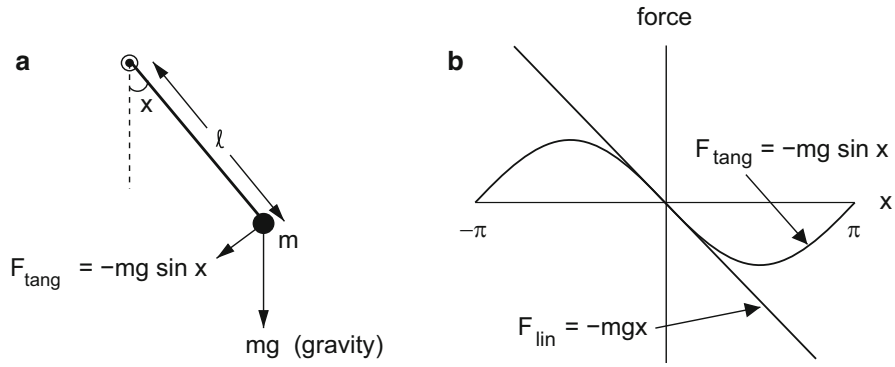

Figure 1.3: *(a) Schematic diagram of the pendulum corresponding to* (1.26)*. (b) Comparison of the tangential component of gravity acting on a pendulum,* $F_{\text{tang}} = -mg \sin x$*, with its linear approximation* $F_{\text{lin}} = -mgx$*.*

let's say an angle rather than a displacement. If x, the angle that the pendulum makes with the vertical, is measured in radians, then ℓx equals the displacement of the mass along the circumference, $\ell\, dx/dt$ equals its velocity, and $\ell\, d^2x/dt^2$ equals its tangential acceleration.[9] The tangential component of gravity is $F_{\text{tang}} = -mg \sin x$. Thus, if there is no friction, Newton's equation of motion may be written

$$x'' + (g/\ell) \sin x = 0, \tag{1.26}$$

where we have divided both terms by $m\ell$. This derivation explains the origin of (1.4), but our purpose here is simply to illustrate how this restoring force deviates from linearity. If x is small, then $\sin x \approx x$, so in this range, the force is approximately linear, but as x increases, the force falls behind this linear growth (see Figure 1.3(b)). Of course, it is also possible for a force to grow more *rapidly* than linearly.

A more extreme deviation from Hooke's law occurs in the cantilever beam (i.e., supported at one end only) placed between two magnets indicated in Figure 1.4. If the beam bends so that its tip is displaced slightly to the right, the beam is closer to the magnet on the right and hence more strongly attracted to it than to the magnet on the left. If the magnetic forces dominate the bending resistance of the beam, then following such a small displacement, the net force on the beam will be to the right; i.e., the beam is pulled *away* from the equilibrium at the centerline rather than toward it. However, if the beam moves beyond the magnet to the right, then both magnetic forces and the bending resistance all pull the beam back toward the center line. Suppose we naively describe the bending of the beam by a single

[9]The mass also experiences a radial acceleration, $-\ell(dx/dt)^2$, from which the tension in the arm may be calculated. However, the motion of the pendulum is determined by the tangential equation (1.26), without consideration of radial forces.

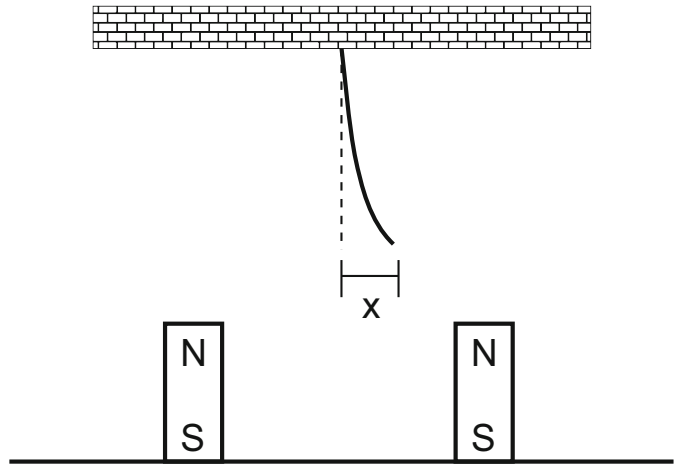

Figure 1.4: *Schematic diagram of a cantilever beam between two magnets.*

variable x, say the displacement of the tip. The simplest force law reproducing the above behavior is

$$F(x) = +k_1 x - k_2 x^3, \tag{1.27}$$

where both k_1, k_2 are positive. Despite its crudeness as a physical model, (1.27) is often used in applications, and mathematically it provides a useful illustrative example; indeed, the ODE

$$x'' + \beta x' - x + x^3 = 0, \tag{1.28}$$

called *Duffing's equation*,[10] has been extensively studied. The force law (1.27) is called a *double-well potential*, which leads us to the concept of potential energy, which we need to define.

1.4.3 Energy

The potential energy $V(x)$ associated with a force law $F(x)$ is defined as the work that must be done against the force to move the mass from a reference position, typically equilibrium, to the position specified by x; in symbols,

$$V(x) = -\int_0^x F(s)\, ds. \tag{1.29}$$

[10]For now, you may regard (1.28) as a particular case of the force law (1.27), but in fact, the general case may be reduced to (1.28) through appropriate scaling (see Chapter 5).

You may wonder why we have used different letters—β in (1.28) and b in (1.24)—for analogous coefficients. As we will explore in Chapter 5, in ODEs arising in applications, most parameters have units (such as length, mass, inverse time) associated with them. Although we are not completely consistent, we try to use Latin letters for parameters with nontrivial dimensions and reserve Greek letters for dimensionless parameters. The latter usually are derived as composites of several dimensional parameters.

The potential functions for Hooke's law, for the pendulum, and for (1.27) are graphed in Figure 1.5; the figure explains the name double-well potential for (1.27). Of course, (1.29) is equivalent, apart from an additive constant, to $F(x) = -\partial V/\partial x(x)$, so the equation of motion of a particle moving in a force field[11] with potential energy $V(x)$ is

$$mx'' + bx' + \frac{\partial V}{\partial x}(x) = 0, \tag{1.30}$$

where we have assumed linear friction. Note that this equation is nonlinear except in the special case that $V(x)$ is quadratic.

As we will discuss in Section 1.8.1, one may loosely visualize solutions of (1.30) as the motion of a marble rolling in the x, z-plane along a curve given by the equation $z = V(x)$, with gravity pointing in the negative z-direction. Although this analogy is quantitatively inaccurate, it nonetheless provides useful qualitative understanding.[12] Moreover, the analogy may help resolve the inevitable confusion arising from the minus sign in (1.29). For example, is the potential energy of a linear attractive restoring force $F = -kx$ given by $V(x) = +kx^2/2$ or $V(x) = -kx^2/2$? It has to be the former, whose graph may be described as a bowl that would *confine* a rolling marble.

Let us also define the *total energy* of a mass moving in a force field. This quantity is the sum of the potential energy of the particle and its *kinetic* energy, in symbols

$$E = \frac{m}{2}(x')^2 + V(x). \tag{1.31}$$

The key property of E is that if $x(t)$ satisfies (1.30), then energy is dissipated at the rate

$$\frac{dE}{dt} = -b(x')^2. \tag{1.32}$$

In particular, if there is no friction, then energy is conserved (i.e., it remains constant as time evolves).

[11] Although we introduced forces in connection with spring–mass systems, we want to consider more general force laws than can reasonably be associated with any spring. For that reason, we adopt the physicists' phrase, "a particle in a force field." If you are interested in mechanics, you should probably learn about the Lagrangian approach to this subject—see Chapter 7 in [88]. With this formalism, it is easy to derive equations of motion when there are constraints, e.g., a particle sliding along a curve in the plane.

[12] One is reminded of the aphorism, "A simple lie may be more useful than a complicated truth" (adapted from de Tocqueville).

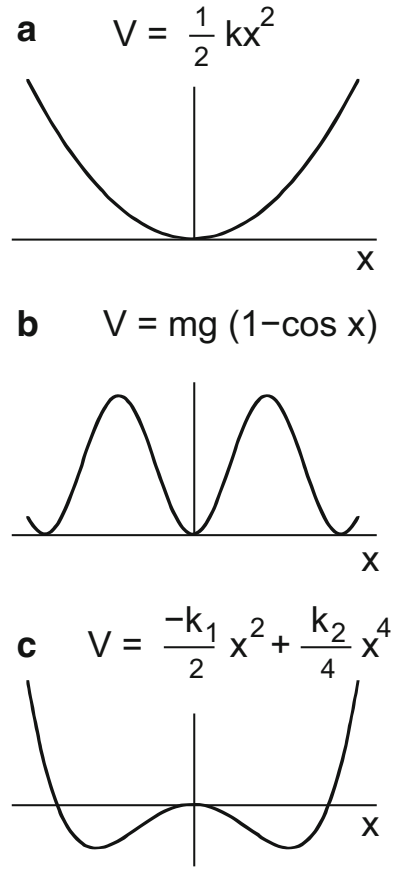

Figure 1.5: *The potential functions for (a) Hooke's law; (b) the pendulum; and (c) the double-well potential.*

1.4.4 Nonlinearity, Part II: The Frictional Force

We have assumed $b \geq 0$ in (1.24), since friction normally dissipates energy. However, in certain electrical circuits, what amounts to negative friction can arise over a limited region of state space. The most famous ODE exhibiting such behavior is van der Pol's equation

$$x'' + \beta(x^2 - 1)x' + x = 0, \tag{1.33}$$

where x measures a voltage in a circuit. If x is small (specifically, $|x| < 1$), the coefficient of x' is negative, leading to growth, while if x is large, this coefficient has the usual positive sign. Historically, van der Pol's equation arose in modeling circuits with vacuum tubes,[13] but of much greater current interest, it also arises in modeling semiconductor circuits.

By way of background, a second-order linear ODE arises from the description of an electrical circuit containing linear elements: an inductor (L), a resistor (R), and a capacitor (C). Specifically, the voltage $x(t)$ across the capacitor in Figure 1.6 at time t satisfies[14]

$$Lx'' + \frac{L}{RC}x' + \frac{1}{C}x = 0. \tag{1.34}$$

(Note that apart from the interpretation of the coefficients, this is exactly the same ODE (1.24) that describes a spring–mass system.) The van der Pol equation arises if the linear resistor in the figure is replaced by an appropriate nonlinear element, including a battery, in which current depends nonmonotonically on voltage. It may appear that with "negative friction," energy is being created out of nowhere, but the derivation explains how (1.33) is consistent with conservation of energy.

1.5 Systems of ODEs

All of the examples of ODEs considered above contained a single unknown function. The advanced theory of ODEs is most efficiently formulated in terms of *systems*—i.e., several simultaneous equations—of ODEs, which involve several unknown functions. Moreover, many physical and biological systems are most naturally modeled by systems of ODEs.[15]

---

[13] Reference [90] is one of the original papers. Modern derivations of van der Pol's equation are available online.

[14] Equation (1.34) may be derived from Kirchhoff's laws—see [72] or look online. Incidentally, a circuit with the elements in series, rather than in parallel, is probably more familiar to most readers. We consider the parallel circuit because it relates more directly to van der Pol's equation.

[15] There is an unfortunate conflict between different fields in the use of the word *system*. At its first occurrence in this sentence, "system" is used in its biological sense "a group of interacting, interrelated, or interdependent elements forming a complex whole." At its second occurrence, "system" is used in its more restricted mathematical sense, "several simultaneous equations."

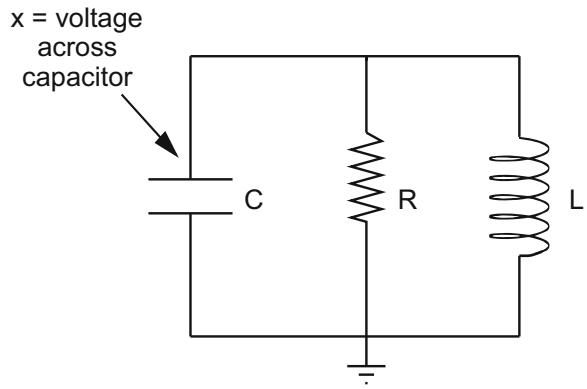

Figure 1.6: *An inductor, a resistor, and a capacitor in parallel. The voltage x across each of the elements satisfies* (1.34)*. If the linear resistor is replaced by a suitably chosen "nonlinear resistor," then x satisfies van der Pol's equation* (1.33)*.*

One of the best-known biological models is the Lotka–Volterra system

$$\begin{aligned} x' &= ax - bxy, \\ y' &= cxy - dy, \end{aligned} \tag{1.35}$$

where the parameters a, b, c, d are all positive. Let us describe the physical assumptions underlying (1.35), since in our view, such understanding is an essential part of acquiring facility with ODEs. This system describes the evolution of two interacting populations, a predator (say sharks, represented by y) and a prey (say regular fish or teleosts,[16] represented by x). In the absence of predators (i.e., $y = 0$), the prey population satisfies $x' = ax$, the equation for exponential growth. However, their growth rate is reduced by predation, which is assumed to occur at a rate proportional to each population.[17] Similarly, the predator equation for y represents a balance between two effects: the predator population increases by a term proportional to the amount of food the predators consume and decreases by a "death" term proportional to their population. Perhaps surprisingly, in the full equation for the evolution of y, *these two effects are simply added*! In general, when several effects occur in a physical system, typically the ODE describing its evolution is obtained simply by

[16] *Teleost* is the biologists' term for what one normally thinks of as "fish." Teleosts have a bony skeleton, in contrast to sharks, whose skeleton is made of cartilage. Teleosts appeared later in evolution, so they are sometimes called "modern" fish or "bony" fish. For purposes of the Lotka–Volterra model, teleost means "fish that are good to eat"—the model was developed to understand perplexing changes in fish harvests during World War I. (See Exercise 22.)

[17] For greater realism, the underlying process should be modeled probabilistically. An ODE model provides a useful approximation for the evolution of average populations, *provided the populations are large*. The rate term proportional to xy may be derived from the probability that members of the two species encounter one another. For more detail, see Section 10.2 and the references therein. In chemical kinetics, the corresponding approximation is called the *law of mass action*.

adding the contributions of each individual effect in the ODE. Of course, although the equations may be simple to formulate, solving them is anything but simple.

Besides arising naturally, systems of ODEs also arise as a mathematical convenience. For example, we claim that van der Pol's equation (1.33) is equivalent to the 2×2 first-order system

$$\begin{array}{rcl} y_1' &=& y_2, \\ y_2' &=& -\beta(y_1^2 - 1)y_2 - y_1. \end{array} \tag{1.36}$$

To see this, suppose $x(t)$ is a solution of (1.33). Then let $y_1(t) = x(t)$ and $y_2(t) = x'(t)$; it is easily seen that the two-component vector $(y_1(t), y_2(t))$ satisfies (1.36). Conversely, if $(y_1(t), y_2(t))$ satisfies (1.36), then a trivial calculation shows that $x(t) = y_1(t)$ satisfies (1.33).

This construction is quite general. Specifically, the nth-order ODE (1.9) is equivalent to the $n \times n$ system for functions $y_1(t), \ldots, y_n(t)$,

$$\begin{array}{rcl} y_1' &=& y_2, \\ y_2' &=& y_3, \\ \vdots & \vdots & \vdots \\ y_{n-1}' &=& y_n, \\ y_n' &=& f(y_1, y_2, \ldots, y_n, t). \end{array} \tag{1.37}$$

The proof of this statement is completely analogous to the above calculation with van der Pol's equation. The presentation of the theory is simplified using vector notation. Thus, for example, if we write $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, then (1.37) can be written compactly as $\mathbf{y}' = \mathbf{F}(\mathbf{y}, t)$, where the vector-valued function $\mathbf{F}(\mathbf{y}, t)$ has the components on the RHS of (1.37).

The theory of first-order systems of ODEs is enriched by the geometric interpretation of such equations. For instance, Figure 1.7 illustrates the geometric interpretation[18] of van der Pol's equation (1.36). The figure shows the vector field

$$\mathbf{F}(\mathbf{y}) = \begin{bmatrix} y_2 \\ -\beta(y_1^2 - 1)y_2 - y_1 \end{bmatrix} \tag{1.38}$$

defined by the RHS of (1.36). A curve $\mathbf{y}(t)$ is a solution of (1.36) iff for every t, the tangent to the curve at the point $\mathbf{y}(t)$ equals $\mathbf{F}(\mathbf{y})$. A few typical solution curves

[18] This figure is our first instance of a *phase plane plot*: i.e., graphs of a few well-chosen trajectories for a two-dimensional equation that indicate the behavior of the general solution.

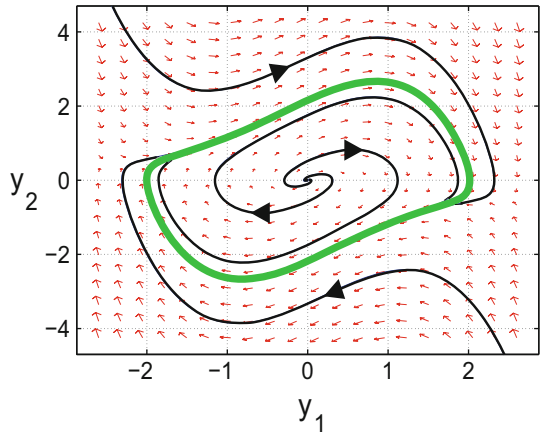

Figure 1.7: *The vector field associated with van der Pol's system* (1.36)*, with* $\beta = 1$*, and several sample solution trajectories. All nonzero solutions of this equation converge to the periodic solution (green curve).*

are also shown in the figure. As $t \to \infty$, every nonzero solution converges to the periodic trajectory, shown in green. (A nonconstant solution is called *periodic* if there exists a time $T > 0$ such that $\mathbf{y}(t+T) = \mathbf{y}(t)$ for all t.) We invite you to use the computer to verify this claim numerically. Considerable theory, which we will develop in Chapter 7, is needed in order to verify it analytically.

We conclude this section with some terminology regarding systems of ODEs. A system $\mathbf{y}' = \mathbf{F}(\mathbf{y}, t)$ is called *linear* if $\mathbf{F}$ has the special form $\mathbf{F}(\mathbf{y}, t) = A(t)\mathbf{y} + \mathbf{g}(t)$, where $A(t)$ and $\mathbf{g}(t)$ are matrix-valued and vector-valued functions of time, respectively. It has *constant coefficients* if the matrix A does not depend on t, and it is *homogeneous* if $\mathbf{g}(t) \equiv \mathbf{0}$. A general system $\mathbf{y}' = \mathbf{F}(\mathbf{y}, t)$ is called *autonomous* if $\mathbf{F}$ is independent of t.

Given an *autonomous* system $\mathbf{y}' = \mathbf{F}(\mathbf{y})$, we call a point $\mathbf{b}_*$ an *equilibrium* of this system if $\mathbf{F}(\mathbf{b}_*) = 0$. For such a point, the constant function $\mathbf{y}(t) \equiv \mathbf{b}_*$ is a solution of this system.

1.6 Topics Covered in This Book

1.6.1 General Overview

In a first course in ODEs, finding explicit solutions of equations is primary. This is a fascinating subject that offers boundless opportunities for ingenuity. However, the

sad fact is that for most equations, explicit solutions cannot be found,[19] and we are forced to develop understanding in the absence of explicit solutions.

The book divides naturally into three parts. The first part, Chapters 2–4, provides the theoretical underpinning for the initial value problem. Chapter 2 prepares the way for this with a careful study of linear systems of ODEs with constant coefficients. In Chapters 3–4, we address three fundamental issues regarding the IVP for nonlinear systems of ODEs:

- existence of solutions (local in Section 3.2, global in Section 4.2);
- uniqueness of solutions (Section 3.3);
- continuous dependence on the initial data (Section 4.5).

The first two phrases are probably self-explanatory; we shall wait till Chapter 4 to flesh out the third.

The middle part, Chapter 5, explores how ODEs may be rendered more transparent by clever scalings of the variables. This chapter is less theoretical than those before or after it, and it connects strongly with ODEs from applications.

In the third part, Chapters 6–9, we develop what is called the *qualitative theory of ODEs*. The goal of this subject is to predict the behavior of solutions of ODEs without knowing explicit solutions. You may well wonder what kind of information can be obtained under these circumstances. We think this question is best answered through example. Therefore, to conclude this introductory chapter, in the next subsection we identify a class of ODEs, an elaboration of the Lotka–Volterra equation (1.35), and illustrate what the qualitative theory can say about such equations. It is not necessary to understand the specific results in detail—for present purposes the spirit of these results is more important than the results themselves.

In a final, supplemental, chapter we indicate some directions for further study, including references.

1.6.2 A Case Study in the Qualitative Theory of ODEs

A central question in the qualitative theory of ODEs is to characterize the asymptotic behavior of solutions as $t \to \infty$. For the Lotka–Volterra equation (1.35), the large-time behavior of solutions is easily described. Let's consider the simplified equation[20]

$$\begin{aligned} x' &= x - xy, \\ y' &= \rho(xy - y), \end{aligned} \tag{1.39}$$

[19] Approximate solutions, obtained either from numerical computations or asymptotic analysis, sometimes provide an adequate substitute. We will touch briefly on both kinds of approximate solutions, but they are not the main focus in this book.

[20] We show in Chapter 5 that by scaling the variables, (1.35) can be reduced to (1.39). In the meantime, you may simply regard (1.39) as a special case of (1.35).

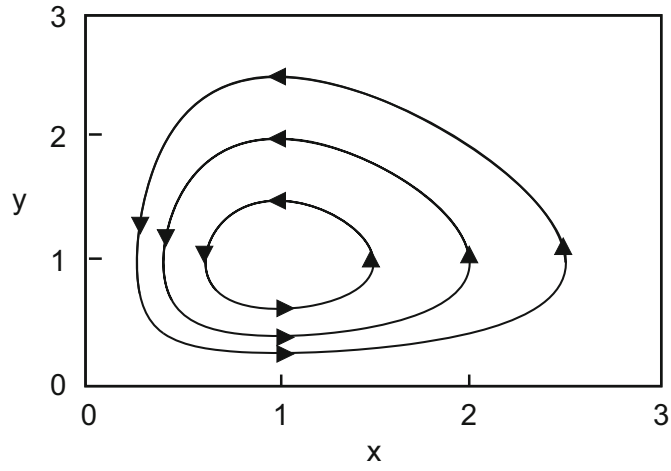

Figure 1.8: *Several solution curves of the Lotka–Volterra system* (1.39)*, with* $\rho = 1$*. The solutions are periodic and encircle the equilibrium at* $(1, 1)$*.*

where ρ is a positive constant. One particular solution of (1.39) is the constant (equilibrium) solution $x(t) \equiv 1$, $y(t) \equiv 1$, which describes a steady balance between the two species, or coexistence. In Exercise 3, we guide you through a proof that every other solution in the open first quadrant[21] $\{x > 0, y > 0\}$ circles this equilibrium point in a periodic fashion, as indicated in Figure 1.8. Indeed, the orbits are level sets of the function

$$L(x, y) = \rho(x - \ln x) + y - \ln y, \tag{1.40}$$

which has a global minimum at $(1, 1)$. (Don't forget that you can use the computer to check this claim numerically!)

However, the Lotka–Volterra model is too simplistic for accurate modeling of predator–prey systems. Without unleashing all the complexities of a realistic model, let's correct two unsatisfactory consequences of the linear growth rate of the prey in (1.39):

- *Solutions of the prey-only equation* $x' = x$ *grow indefinitely as time evolves.* As in (1.2), logistic growth—say $x' = x - x^2/K = x(1 - x/K)$, where K is a constant—would be a more realistic prey-only equation. We may interpret K as the carrying capacity of the environment.[22]

[21] *Food for thought:* How do solutions on the boundary of the first quadrant, $\{x \equiv 0\}$ or $\{y \equiv 0\}$, behave?

[22] You may wonder why we should not choose K equal to unity, as in (1.2). We could in fact do this, but only at the expense of either losing some generality or making the y-equation in the system less transparent. The mystery surrounding this and other applications of scaling arguments should be dispelled by Chapter 5.

- *No matter how small $x(0)$ may be, the prey never become extinct.* This defect may be corrected in an ad hoc manner by replacing the linear growth term[23] by $x(x-\varepsilon)/(x+\varepsilon)$. Then for $x < \varepsilon$, the growth rate is negative; thus if $x(0) < \varepsilon$, the prey will die out. On the other hand, for large x, the growth rate is close to x, as in the original equations.

Inserting both of these modifications of the prey growth rate into the system (1.39) gives us the equations

$$\begin{aligned} (a) \quad x' &= x\left(\frac{x-\varepsilon}{x+\varepsilon}\right)(1-\frac{x}{K}) - xy, \\ (b) \quad y' &= \rho(xy-y), \end{aligned} \tag{1.41}$$

which we call the *augmented Lotka–Volterra* equations. If $\varepsilon = 0$ and $K = \infty$, then we obtain the unmodified Lotka–Volterra equations (1.39). In studying (1.41), we assume that[24]

$$0 < \varepsilon < \min\{K, 1\}. \tag{1.42}$$

In contrast to the continuing oscillations of (1.39), solutions of (1.41) typically converge to a steady state with populations independent of time. The solution may converge to a coexistence equilibrium, to a prey-only equilibrium, or to total extinction. Not surprisingly, which asymptotic behavior is selected depends on the initial conditions and on the parameters ε, K, ρ. However, as we will see shortly, some surprising behavior is hidden in the details.

To elaborate, let us calculate the equilibrium solutions of this system[25]—i.e., points where $x' = y' = 0$. From (1.41b), we find that $\rho(xy - y) = 0$ if either $y = 0$ or $x = 1$. Substituting $y = 0$ into (1.41a) gives the three equilibria

$$\begin{array}{lll} \text{(i)} & (0,0) & \text{Extinction,} \\ \text{(ii)} & (\varepsilon,0) & \text{Extinction threshold,} \\ \text{(iii)} & (K,0) & \text{Prey-only equilibrium.} \end{array} \tag{1.43}$$

Substituting $x = 1$ gives a fourth,

$$\text{(iv)} \quad (1, y_{\text{coeq}}) \quad \text{Coexistence equilibrium} \tag{1.44}$$

where

$$y_{\text{coeq}} = (1 - 1/K)(1-\varepsilon)/(1+\varepsilon). \tag{1.45}$$

Note that if $K < 1$, then $y_{\text{coeq}} < 0$; i.e., coexistence is physically possible only if $K > 1$.

[23] A growth rate that depends on the population size is called the *Allee effect*.

[24] We want $\varepsilon < K$, so that the carrying capacity exceeds the threshold for extinction, and when $K > 1$, we want $\varepsilon < 1$, so that the prey population at the coexistence equilibrium (1.44) exceeds the threshold for extinction.

[25] Note that there are *multiple* equilibria—we shall see that this is typical for nonlinear systems.

Figure 1.9 encapsulates information about solutions of (1.41). In the lower-right panel, the parameter set (1.42) is divided into three regions by the curves

$$\begin{array}{ll} K = 1 & \text{(between Regions I and II)} \\ K = (1 + 2\varepsilon - \varepsilon^2)/2\varepsilon & \text{(between Regions II and III).} \end{array} \tag{1.46}$$

In Chapter 8, we will invoke the qualitative theory to predict that solutions of (1.41) exhibit qualitatively different asymptotic behavior for (ε, K) belonging to different regions, behavior that we can characterize without solving the equations, either analytically or numerically. At this point, without understanding the basis for (1.46), we choose one set of parameter values within each region and solve (numerically) the IVP for two choices of initial conditions. Specifically, in each of the other three panels in Figure 1.9, for $\rho = 1$ and $\epsilon = 1/5$, trajectories are shown for initial conditions $(x(0), y(0)) = (1.5, 0.5)$ and $(2, 2)$. We may summarize the observed asymptotic behavior as follows: In all regions, if there are too many predators at the start, everybody dies. If there are fewer predators initially, different behavior is possible, depending on the region. In Region I, the carrying capacity K is too small to sustain both populations, but solutions may converge to a prey-only equilibrium. If resources are somewhat more abundant (Region II), solutions may converge to a coexistence equilibrium. But here's the shocker: if the carrying capacity is too large (Region III), *both species are driven to extinction*, even if initially the number of predators is very small (but positive)! Although increasing the carrying capacity seems like it should promote the overall health of the system, a worse fate results—greater resources lead to growing oscillations that spiral to disaster.

While this behavior is interesting in its own right, we remind you that our main point here is to illustrate the power of the qualitative theory.

1.7 Exercises

Preamble: You may feel discouraged to find yourself staring at ten pages of exercises, but we all know that they are an essential part of learning the subject. Moreover, the situation is not as bad as it may seem. The problems are actually a lot shorter than they may appear—only the text in normal-sized type actually asks you to do anything, the text in smaller type consists of ample hints and commentary. Believing that a "divide and conquer" strategy may reduce the intimidation factor, we have grouped these exercises into several categories according to their primary purposes, as follows: (1) core exercises to give you practice with the ideas in the chapter, (2) computational exercises, (3) anticipatory exercises that introduce themes that will be important later in the book (not to be skipped), and (4) "PHD exercises" that may be more difficult or develop some related topic.

You may think that there is a large number of computational exercises, but you will be doing yourself a favor if you do enough of them (in this and later chapters) so that you become as comfortable using the computer as using pencil and paper.[26] In addition to the explicit computational

[26]Perhaps, for a computer-savvy generation, we should be giving pencil and paper a plug.

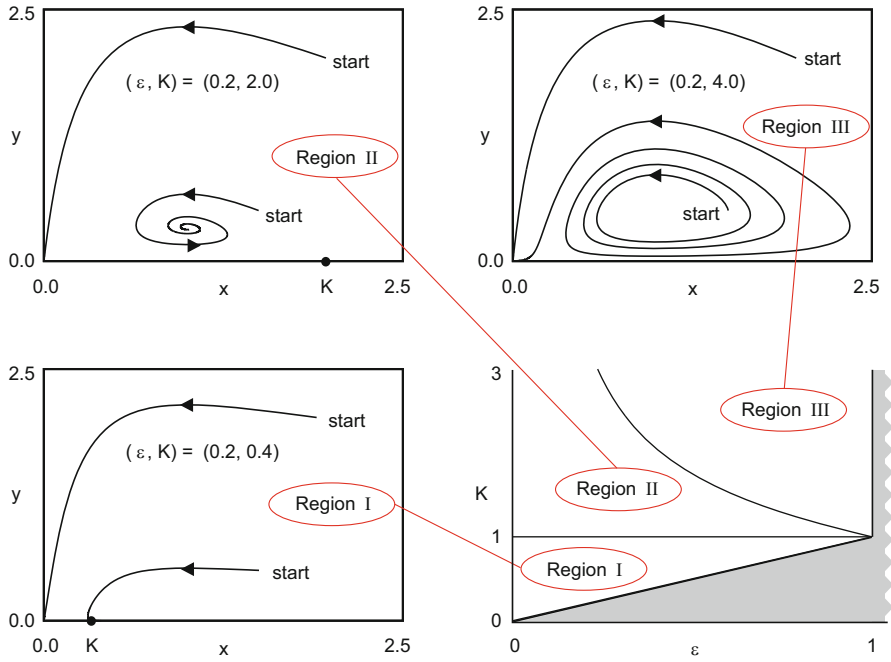

Figure 1.9: *The lower-right panel shows three regions in the ε-K parameter space* (1.42) *for which solutions of* (1.41) *have different asymptotic behavior. The boundaries between regions are formed by the curves* (1.46)*. The other three panels illustrate typical (or "generic"—see Section 1.8.2) behaviors of solution trajectories for each region, using parameters and initial conditions as described in the text. The periodic trajectories of the unperturbed system* (1.35)*, which are shown in Figure 1.8, are completely changed by the seemingly minor perturbation in* (1.41)*; moreover, the change depends in a nonobvious way on parameters.*

exercises below, we invite you to use the software to check any statements made in the text. For example, although Exercise 3 gives an analytical proof that every solution of (1.39) (in the first quadrant) is periodic, it may be reassuring to see this behavior in computed solutions.

1.7.1 Core Exercises

The goal of the Core exercises is as follows:

To deal with unfinished business	1–3
To practice solving separable ODEs	4
To increase your general facility with ODEs	5–7

1. (a) Show that every solution of $x' = ax$ equals Ce^{at} for some constant C.

   Hint: Show that for a solution $x(t)$, the derivative of $e^{-at}x$ is zero.

   (b) Verify that formula (1.18) solves the first-order linear equation (1.17).

   Hint: Write $e^{\bar{a}(s)-\bar{a}(t)} = e^{\bar{a}(s)}e^{-\bar{a}(t)}$ and move $e^{-\bar{a}(t)}$ outside the integrand. Then differentiate using the product rule.

   (c) Prove the following claims about the logistic equation made in the text.

   - The function (1.22) satisfies (1.2).
   - Provided $b \neq 1$, equation (1.22) with C'' given by (1.23) satisfies the initial condition $x(0) = b$.
   - If $b \geq 0$, the explicit solution and its exceptional case provide a solution of the IVP for all positive time.

   (d) Derive (1.32), the equation for energy dissipation in a spring–mass system.

2. Show that if $\beta = 2$ in (1.14), then for every choice of constants C_1, C_2,

$$x(t) = C_1e^{-t} + C_2te^{-t} \tag{1.47}$$

   is a solution of this equation.

   Discussion: This problem represents a minor special case of what we will prove in the next chapter. Formula (1.47) is in fact the general solution of (1.14) when $\beta = 2$. Incidentally, the second independent solution te^{-t} can be obtained as a limit as $\beta \to 2$ of carefully chosen solutions of (1.14) when $\beta \neq 2$, specifically, the limit of

$$\frac{e^{\lambda_+ t} - e^{\lambda_- t}}{\lambda_+ - \lambda_-}.$$

3. *Introduction:* Although it is not possible to solve the Lotka–Volterra (1.39) equations for x and y as functions of t, it is possible to eliminate time and derive an implicit relation between x and y along solution curves. To this end, we derive an ODE for y as a function of x from the chain rule

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{\rho(xy - y)}{x - xy}. \tag{1.48}$$

   (a) If $L(x, y)$ is defined by (1.40), derive

$$L(x, y) = \text{const} \tag{1.49}$$

   as an implicit solution of (1.48), using the fact that this equation is separable.

   (b) Verify that the level sets of $L(x, y)$ are closed curves.

   Discussion: Combining (a) and (b), we see that the trajectories of (1.39) are *contained* in closed curves. To complete the proof that every nonconstant trajectory is periodic, we would have to rule out the possibility that a trajectory might not complete the entire circuit around the closed curve. You may either fill this gap on your own, or you may look ahead to Corollary 7.1.2.

   Incidentally, in the context of this exercise let us formalize some geometric language that will be useful later. By a *trajectory* we mean a specific parametrized curve $\{(x(t), y(t)\}$, where $x(t)$ and $y(t)$ solve the ODE (1.39). By contrast, the term *orbit* refers to the curve as a geometric object, independent of any parametrization. Thus, in this problem we find orbits of (1.39), but not trajectories.

4. *Introduction:* Using separability, find the general solutions for the following equations, and if initial conditions are given, solve the specific IVPs. (Except for (a), each of these problems raises issues beyond mere practice in finding explicit solutions.)

   (a) The Gompertz model for tumor growth, in which the center is starved for oxygen (see p. 217 of Edelstein-Keshet [19]):

$$dN/dt = \mu e^{-\alpha t} N.$$

   (b) An equation based on the pedagogical example (1.7):

$$x' = \sqrt{|x^2 - 1|}, \qquad x(0) = 1. \tag{1.50}$$

   Discussion: We have inserted absolute values inside the radical so that the RHS of the equation is defined for all x. Carry through the usual recipe, find a solution, and check it. Now, here is another, equally valid, solution: $x(t) \equiv 1$; i.e., *the solution of* (1.50) *is not unique!* In Chapter 3, we will give conditions that guarantee that an initial value problem has a unique solution. In the meantime, you may want to ponder what misbehavior of $\sqrt{|x^2 - 1|}$ leads to this nonuniqueness.

(c) An academic example to stimulate your thinking:

$$x' = -1/x, \qquad x(0) = 1.$$

Puzzle: The formula you obtain from separability will not be usable for large positive t. What went wrong?

(d) A system that will be used frequently as an illustration in later chapters:

$$\begin{aligned} x' &= x - y - (x^2 + y^2)x \qquad x(0) = b_1, \\ y' &= x + y - (x^2 + y^2)y \qquad y(0) = b_2. \end{aligned}$$

Hint: Solving this system would be hopeless except for the fact that it may be rewritten in polar coordinates

$$r' = r(1 - r^2), \qquad \theta' = 1,$$

in which the two equations are uncoupled. *(Derive these equations!)*

5. *Directions:* Make up your own ODEs with the following properties. The ODEs should be in the standard form where the highest derivative has been "solved for."

(a) A third-order scalar ODE that is nonlinear and nonautonomous.

(b) A fifth-order, linear, homogeneous scalar equation with constant coefficients such that every solution tends to zero as $t \to \infty$.

Hint: It will proved below, but you may assume for now, that the general solution of $x^{(5)} + a_1 x^{(4)} + \ldots + a_5 x = 0$ is a linear combination of the exponentials $e^{\lambda_k t}$, $k = 1, \ldots, 5$, where $\lambda_1, \ldots, \lambda_5$ are the roots of the characteristic polynomial $\lambda^5 + a_1\lambda^4 + \ldots + a_5 = 0$, *provided* these roots are distinct. To construct your example, make up a fifth-order polynomial with distinct roots all lying in the left half-plane.

(c) A first-order nonlinear autonomous three-dimensional system.

Remark: After you have constructed your example, we invite you to check out what is probably the most famous such system in existence, the Lorenz equations, (8.5).

(d) A first-order 2×2 linear system that is autonomous but not homogeneous.

Remark: This exercise shows that "homogeneous linear system with constant coefficients" and "autonomous linear system" are slightly different concepts.

(e) A first-order inhomogeneous 3×3 linear system with variable coefficients.

6. (a) Sketch the direction field for the logistic equation (1.2).

   (b) Trace curves tangent to the direction field to argue that for every initial condition $x(0) > 0$, the solution of the IVP tends to 1 as $t \to \infty$.

7. *Introduction:* The following easy exercise illustrates other representations of solutions of (1.3).

   (a) Show that if $D_1, D_2 \in \mathbb{C}$, then

   $$x(t) = D_1 e^{it} + D_2 e^{-it} \tag{1.51}$$

   is a complex-valued solution of (1.3). Also show that if $D_1 = \bar{D}_2$, where bar indicates complex conjugation, then (1.51) is real-valued, and conversely.

   (b) Show that for every solution $x(t)$ of the form (1.12), there exist real constants C, δ such that

   $$x(t) = C \sin(t + \delta), \tag{1.52}$$

   and conversely.

1.7.2 Computational Exercises

8. Compare numerical solutions of the logistic equation (1.2) with the analytical solution (1.22).

   Remark: This exercise is more for practice in using software than for any interesting math.

9. (a) For the Riccati equation (1.6), compute numerical solutions to show that for negative and for small positive values of the initial condition b, the solution is asymptotic to the parabola $x^2 - t = 0$; and for large positive values of b, the solution blows up in finite time.

   Advice: Blowup may be seen in better detail if you plot x on a log scale.

   (b) Estimate b_*, the initial datum that separates the two behaviors.

   Remark: In the language of Section 1.8.2, solutions of the Riccati equation generically either blow up in finite time or are asymptotic to the parabola $x = -\sqrt{t}$. The dividing case with $x(0) = b_*$ is nongeneric (and hard to compute).

10. (a) In the van der Pol equation (1.36), set $\beta = 1$ and solve the initial value problem for several choices of initial conditions.

    (b) Choose some other values of β—don't mess around, make a big change, like $\beta = 10$ or 0.1—and repeat the above computation.

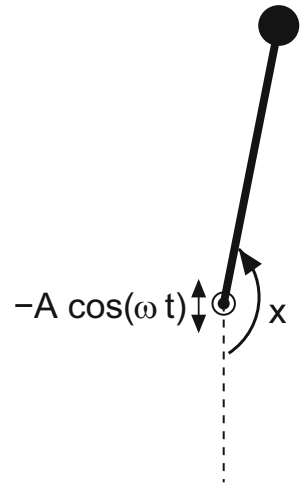

Figure 1.10: *Schematic of the vertically vibrated pendulum of Exercise 12.*

Discussion: This exercise in intended to show that typical solutions of the van der Pol equation tend to periodic behavior as $t \to \infty$. As long as the initial conditions are nonzero, you will see a periodic solution emerge. In Part (b), changing the value of β will give a different periodic orbit from the one shown in Figure 1.7.

11. For the augmented Lotka–Volterra equation (1.41), choose various initial conditions to verify the behavior summarized in Figure 1.9.

12. *Introduction:* This is a fun problem that demonstrates an amazing fact: if the supporting pin of a pendulum is vibrated vertically sufficiently rapidly (cf. Figure 1.10), the "straight up" position of the pendulum may be *stable*! (Look online for a movie.) If the position of the pin is $-A\cos\omega t$ and if friction is proportional to (angular) speed, then the (angular) displacement x of the pendulum approximately satisfies an equation of the form (1.53) below, where α is proportional to A. Note the similarity to Mathieu's equation (1.5b). In Chapter 7, we will give an analytical derivation that vibrations can stabilize the inverted equilibrium.

    (a) Write the equation

    $$x'' + \beta x' + [1 - \alpha\omega^2 \cos\omega t] \sin x = 0 \tag{1.53}$$

    as a first-order system.

    (b) Start with the pendulum at rest and nearly vertical, say $x(0) = 3.1$, and let $\alpha = \beta = 0.1$. Solve the IVP for various values of ω, say starting with $\omega = 1$ and doubling it repeatedly until you reach an ω such that the pendulum comes to rest in the straight-up position.

1.7.3 Anticipatory Exercises

13. *Introduction:* Consider a second-order inhomogeneous linear scalar ODE

$$x'' + a_1(t)x' + a_2(t)x = g(t). \tag{1.54}$$

Let $x_{\text{partic}}(t)$ be some solution of (1.54). Such a solution is called a *particular* solution, which provides the mnemonic for the subscript.

(a) Show that for *every* solution $x(t)$ of (1.54), there is a solution $x_{\text{homog}}(t)$ of the homogeneous equation (i.e., (1.54) with the inhomogeneous term $g(t)$ set equal to zero) such that

$$x(t) = x_{\text{partic}}(t) + x_{\text{homog}}(t).$$

Remark: It is worth retaining this idea—i.e., you can solve an inhomogeneous linear equation with a particular solution plus a homogeneous solution. This technique is taught in all elementary courses on ODEs.

(b) Consider periodic forcing of a spring–mass system

$$mx'' + bx' + kx = \Gamma \cos \omega t. \tag{1.55}$$

Find a particular solution of this equation by looking for a solution in the form $x_{\text{partic}}(t) = A \cos \omega t + B \sin \omega t$.

(c) Graph the amplitude $\sqrt{A^2 + B^2}$ in $x_{\text{partic}}(t)$ as a function of ω, paying particular attention to the limit of small $b > 0$.

Discussion: You will see that if b is small, the amplitude has a large spike when ω is close to the frequency $\omega_0 = \sqrt{k/m}$ of the undamped oscillator.

To generalize: consider a weakly damped linear ODE that has slowly decaying oscillatory solutions of the form $e^{-\varepsilon t} \cos \omega_0 t$. If such a system is subjected to an oscillatory force whose frequency is close to ω_0, the response may be very large. This phenomenon is known as *resonance*. It may be seen over a vast range of scales, ranging from a child pumping a swing to the collapse[27] of the Tacoma Narrows Bridge (Braun [10] p. 173). Other forms of resonance, including in a nonlinear equation, will be studied below.

(d) Show that assuming $b > 0$, every solution of (1.55) tends to $x_{\text{partic}}(t)$ as $t \to \infty$.

Hint: If $b^2 = 4mk$, you will need to use ideas from Exercise 2 to complete this part of the exercise.

[27] See also [82] for a discussion of a less dramatic resonance-induced bridge episode.

14. (a) Solve the IVP

$$\varepsilon x'' + x' + x = 0, \qquad x(0) = a, \qquad x'(0) = b,$$

where ε is a small positive parameter.

Hint: Although this problem can be solved exactly, it's messy, and an approximation is perfectly adequate. The two roots of the characteristic equation are approximately $-1, -1/\varepsilon$. Look for an approximate solution of the IVP in the form

$$Ae^{-t} + Be^{-t/\varepsilon}$$

that uses these approximate roots in the exponents.

(b) *Introduction:* Given that $\varepsilon \ll 1$, it is tempting to consider an approximation setting $\varepsilon = 0$ in the ODE. This approximation has the alarming effect of transforming a second-order ODE into a first-order equation! In particular, it is no longer possible to satisfy both initial conditions. Nevertheless, in the next part of the exercise we invite you to throw caution to the wind.

Solve the IVP

$$x' + x = 0, \qquad x(0) = a.$$

(c) Set $\varepsilon = 0.1$, $a = 1$, $b = 1$, and plot the two solutions.

Discussion: Observe that apart from an initial short-lived transient, the two solutions closely track one another; i.e., in this instance we didn't get burned by the brutal approximation. However, an approximation that changes the order of an ODE has to be approached with caution. To drive this point home, you might repeat the exercise for $\varepsilon x'' - x' - x = 0$. We will encounter similar issues on multiple occasions below.

15. (a) Suppose $x = b_*$ is an equilibrium of a scalar ODE $x' = f(x)$. Make an educated guess (don't bother with a "proof") as to which of the following two statements is true if $f'(b_*) > 0$ and which if $f'(b_*) < 0$:

- If the initial datum $x(0)$ is sufficiently close to b_*, then $x(t)$ tends to b_* as $t \to \infty$.
- No matter how close the initial datum $x(0) \neq b_*$ may be to b_*, the solution $x(t)$ moves further away from b_* as t increases.

(b) Check your hunch against the two equilibria $x = 0$ and $x = 1$ of the logistic equation (1.2).

Remark: This problem anticipates the concept of stability of an equilibrium, a truly fundamental idea in the qualitative theory of ODE. It will be studied in earnest beginning in Chapter 6.

1.7.4 PHD Exercises

16. *Introduction:* In this exercise you show that a solution of a scalar ODE $x' = f(x)$ cannot cross a zero of f, provided $f \in \mathcal{C}^1$. In fact, this conclusion follows simply from the general uniqueness result, Theorem 3.3.4 below, but it may be useful preparation for later work to derive this particular case now. The following statement concerns behavior of a solution $x(t) > 0$ near a zero of f located at the origin, but the general case can be reduced to this special case by translation and reflection.

Let $x(t)$ be a solution of a scalar ODE, $x' = f(x)$, where f is $\mathcal{C}^1$ and $f(0) = 0$. Show that if $x(0) > 0$, then $x(t) > 0$ for as long as the solution exists.

Hint: Since $f(0) = 0$ and $f \in \mathcal{C}^1$, you may bound f as follows: there is an interval $[0, a]$ where $a > 0$, and a constant C such that

$$f(x) \geq -Cx \qquad \text{for } 0 \leq x \leq a.$$

Now suppose the claim is false, say $x(t) > 0$ for $0 \leq t < t_*$ but $x(t_*) = 0$. There may be a long interval in which $x(t) > a$, but you may ignore this. Isolate an interval $[t_* - \varepsilon, t_*)$ such that

$$0 < x(t) \leq a \qquad \text{if } t_* - \varepsilon \leq t < t_*.$$

Then show that $(d/dt)\,[e^{Ct}x(t)] \geq 0$ on this interval and hence conclude that

$$x(t) \geq x(t_* - \varepsilon)e^{C[(t_*-\varepsilon)-t]} \qquad \text{for } t \in [t_* - \varepsilon, t_*],$$

which contradicts the assumption that $x(t_*) = 0$.

17. *Directions:* Use separability to solve the following ODEs:

(a) The logistic equation with "constant harvesting":

$$x' = x(1 - x) - \mu,$$

where μ is a positive constant.

Advice: This problem is technically complicated; don't pursue it if it ceases to be enjoyable. The cases $0 < \mu < 1/4$, $\mu = 1/4$, and $1/4 < \mu$ must be treated separately. To understand why the behavior of the equation changes at $\mu = 1/4$, think about the equilibrium equation $x(1 - x) - \mu = 0$.

In this ODE, it is assumed that constant harvesting continues even if $x \to 0$, which of course is unsustainable. A consequence of this faulty assumption: the equation can predict negative populations.

(b) The equation $x'' = x'x$, which was offered to illustrate nonlinearity.

Hint: Integrate the equation once to obtain $x' = x^2/2 + C$, where C is a constant, and then apply separability. The cases $C > 0$, $C = 0$, and $C < 0$ lead to different formulas.

18. (a) Apply Theorem 1.8.1 below to show that solutions of the IVP

$$x' = x^2 - t, \qquad x(0) = b, \tag{1.56}$$

blow up in finite time if $b \geq 1$.

Hint: Suppose (1.56) is solvable on the interval $0 \leq t < t_*$. Let $u(t) = (1 - t/2)^{-1}$ be the solution of

$$u' = u^2/2, \qquad u(0) = 1. \tag{1.57}$$

Since $b \geq 1$, we have the initial bound $u(0) \leq x(0)$. Show that for $0 \leq t < 2$,

$$\frac{1}{2(1-t/2)^2} = \frac{du}{dt}(t) \leq f(u(t), t) = \frac{1}{(1-t/2)^2} - t$$

by clearing the denominator and applying calculus. Derive a contradiction from the theorem by assuming $t_* > 2$.

(b) Apply Theorem 1.8.1 to show that if $b \leq 0$, then the solution $x(t)$ of (1.56) satisfies

$$-\sqrt{b^2 + t} \leq x(t) \leq 0 \qquad (t \geq 0) \tag{1.58}$$

for as long as the solution exists.

Hint: To bound $x(t)$ from below, let $u(t) = -\sqrt{b^2 + t}$ and verify the hypotheses of Theorem 1.8.1. To bound $x(t)$ from above, first formulate an analogue of Theorem 1.8.1 in which the directions of all inequalities are reversed, and then verify the hypotheses of the reformulated theorem for the function $u(t) \equiv 0$.

Remarks: In Part (a), the solution of (1.57) is a useful basis for comparison, because (1.57) is simple enough that it can be solved explicitly, while it still retains the quadratic growth of (1.56) as $x \to \infty$. In Part (b), the comparison function $u(t) = -\sqrt{b^2 + t}$ for the lower bound was the first function we thought of such that $u(t) \approx -\sqrt{t}$ as $t \to \infty$ and $u(0) \leq b$ when $b < 0$, and this choice worked.

Incidentally, the estimate (1.58) actually implies that the solution of the IVP exists for all positive time if $b \leq 0$; this will follow from our results in Chapter 4.

19. *Introduction:* Exercise 9 applied computation to study the IVP for Riccati's equation (1.56); Exercise 18 applied rigorous analysis; the present exercise applies asymptotics to the same end.

Show that there are formal series solutions of the Riccati equation (1.6),

$$x_+(t) = \sqrt{t}\left(1 + \frac{a_0}{t^{3/2}} + \frac{a_1}{t^3} + \dots\right) \quad \text{and} \quad x_-(t) = \sqrt{t}\left(-1 + \frac{b_0}{t^{3/2}} + \frac{b_1}{t^3} + \dots\right),$$

series in *inverse* powers of $t^{3/2}$.

Discussion: When we say *formal* series, we are allowing the possibility that the series may not converge. Thus to show that a formal series solution exists, you need only derive a recurrence relation for successive coefficients in the series. Although we hesitate to endorse mathematical sloppiness, we admit that in actual practice, one often just calculates the first couple of coefficients in the series and gets a general sense of how subsequent terms will work out.

The series x_- approximates the asymptotic behavior of all those solutions that remain in the lower half-plane $\{x < 0\}$ as $t \to \infty$. Since the series are based on inverse powers, they are useful in the limit $t \to \infty$. We invite you to compare, for large t, say $t > 5$, the sum of the first few terms of these series with your numerical solutions from Exercise 9. By contrast, the series x_+ approximates (when t is large) the unique solution that separates the two typical asymptotic behaviors of solutions of (1.6). The asymptotic solution is particularly useful in the latter case, since some effort is required to locate this solution numerically.

20. *Introduction:* The following problem derives an amusing, totally unsuspected, connection between two, very different, ODEs that appeared in this chapter.

Show that if $x(t)$ satisfies Airy's equation (1.5a), then over an interval where $x(t) \neq 0$, the function $y(t) = -x'(t)/x(t)$ satisfies the Riccati equation (1.6).

21. Derive the ODE that reparametrizes solutions of (1.59), in Section 1.8 below, by arc length.

Remark: Your equation will be similar to (1.60), but with a significant difference: parametrization by arc length breaks down at a point where $\mathbf{F}(\mathbf{x})$ vanishes, while (1.60) remains nonsingular.

22. *Introduction:* In the modification of (1.35),

$$\begin{aligned} x' &= ax - bxy - fx, \\ y' &= cxy - dy - fy, \end{aligned}$$

the terms proportional to f are added to model the effect of fishing on the populations.

Verify that these equations have a coexistence equilibrium

$$x = \frac{d+f}{c}, \qquad y = \frac{a-f}{b}.$$

Discussion: If there is no fishing ($f = 0$), then the equilibrium is $(x, y) = (d/c, a/b)$. The astounding fact is that if f is positive and not too large, then the equilibrium value of x is increased. That is, if predation is the main cause of prey death, then *moderate fishing boosts the prey population*, both in absolute numbers and in fraction of the catch.

During World War I, the reverse phenomenon occurred: because of the war, there was less fishing, and a higher fraction of sharks appeared in the catch. The model (1.35) was developed to help understand this observation. See Section 4.10 of [10] for more particulars.

1.8 Pearls of Wisdom

1.8.1 Miscellaneous

In the first lesson in ODEs, one learns that extra conditions must be imposed to single out one solution of an equation from among infinitely many. In this book we apply initial conditions to select a specific solution. However, you should know that in many applications, other conditions are more appropriate. Chief among these are *boundary conditions*, in which constraints on the solution of an ODE defined on an interval $0 \le t \le T$ are imposed at *both* ends of the interval. We briefly discuss such problems in Section 10.1, along with references.

There is a general construction to reformulate a nonautonomous system of ODE in n dimensions, say

$$\mathbf{y}' = \mathbf{F}(\mathbf{y}, t),$$

where $\mathbf{y}(t) = (y_1(t), \ldots, y_n(t))$ and $\mathbf{F} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, as an autonomous system in one higher dimension. Specifically, form a new unknown $\mathbf{z}(t) = (z_1(t), \ldots, z_n(t), z_{n+1}(t))$ by adding one component and require that $\mathbf{z}(t)$ satisfy $\mathbf{z}' = \mathbf{G}(\mathbf{z})$, where the RHS $\mathbf{G} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is defined by

$$\mathbf{G}(\mathbf{z}) = \begin{bmatrix} \mathbf{F}(\tilde{\mathbf{z}}, z_{n+1}) \\ 1 \end{bmatrix},$$

$\tilde{\mathbf{z}}$ being shorthand for the first n components of $\mathbf{z}$. The two systems are connected through the observation that $z_{n+1}(t)$ is essentially equivalent to time.

In Section 1.4 we suggested that one might visualize solutions of (1.30), a particle moving in a potential $V(x)$, as the motion of a marble rolling in the x, z-plane along a curve given by the equation $z = V(x)$. This analogy is qualitatively useful; for example, as we will prove in Chapter 6, a particle moving according to (1.28) will indeed come to rest at the bottom of one of the wells of the potential in Figure 1.5(c), just as a rolling marble would do. On the other hand, it is quantitatively inaccurate. In the first place, rolling introduces a whole new level of complexity—one needs to distinguish between rolling with and without slipping, which requires examining friction between the marble and the surface. Even if we ignore rolling and imagine a particle *sliding* (with negligible friction) along the curve $z = V(x)$, the analogy is still flawed. In an extreme case, the particle may move so rapidly that it lifts off the curve. Less dramatically, in sliding along a curve even at slow speed, motion in the z-direction influences the x-component of the motion.[28]

[28] You can check this statement using the Lagrangian formulation of mechanics (cf. footnote number 11) to derive the equations for such constrained motion.

Sometimes it is convenient to reparametrize time in an ODE. One useful reparametrization of an autonomous system

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}) \tag{1.59}$$

comes from solving the ODE

$$\frac{d\mathbf{x}}{d\tau} = \frac{\mathbf{F}(\mathbf{x})}{\sqrt{1+|\mathbf{F}(\mathbf{x})|^2}}. \tag{1.60}$$

Although (1.59) and (1.60) have different solutions, both equations have the same *orbits*; this follows from the uniqueness theorem of Chapter 3. Moreover, while solutions of (1.59) may blow up in finite time, even if $\mathbf{F}$ is defined on all of $\mathbb{R}^d$, solutions of (1.60) exist for all τ; this follows from existence results in Chapter 4. The reparametrization in (1.60) is subtly different from reparametrizing orbits by arc length; see Exercise 21.

1.8.2 The Concept "Generic"

Let us begin to develop the important concept of *generic* behavior, which is a more technical synonym for "typical" behavior, with an implication of robustness. To be specific, consider the trajectories illustrated in the upper left panel of Figure 1.9, which are solutions of the augmented Lotka–Volterra equation (1.41) with the parameter values $\varepsilon = 0.2$, $K = 2$, $\rho = 1$. We claim that for "generic" initial data, the solution of the IVP for this problem

converges to either the coexistence or the extinction equilibrium. (1.61)

Note that not *every* solution has this long-time behavior. For example, the threshold-equilibrium solution $(x(t), y(t)) \equiv (\varepsilon, 0)$, which is constant in time, isn't covered by (1.61). You may object that this solution does not belong to the open first quadrant. Well, point taken, but here is an example that counters that objection. As illustrated in Figure 1.11, consider a one-parameter family of initial conditions lying on the line $\{x = 2\}$, say $(x(0), y(0)) = (2, b)$ where $0 < b < 2$. On the one hand, if b is close to 0, the solution will converge to the coexistence equilibrium. On the other hand, if b is large, the solution will converge to extinction. By continuity, somewhere in between these extremes is an initial condition $(2, b_*)$ that separates these behaviors. As one might guess, the solution with initial condition $(2, b_*)$ in fact converges to the threshold equilibrium $(\varepsilon, 0)$ as $t \to \infty$, which is not covered by (1.61).

So what are we claiming in (1.61)? Solutions with initial conditions $(x(0), y(0)) = (2, b)$ where $b \neq b_*$ are robust in the sense that every sufficiently small perturbation of the initial data produces a solution with the same asymptotic behavior. By contrast, the solution with initial conditions $(x(0), y(0)) = (2, b_*)$ is far from robust—the

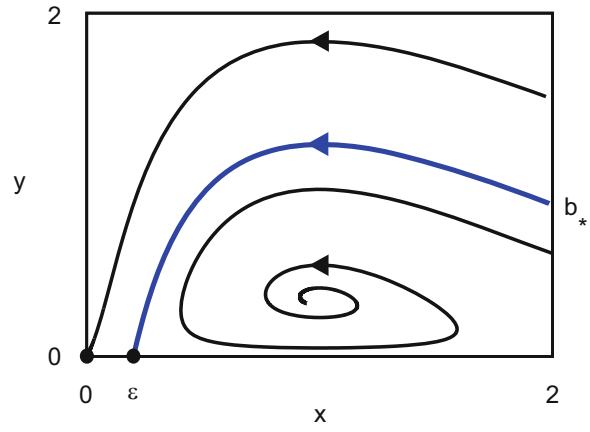

Figure 1.11: *An elaboration of the flow of* (1.41) *in Figure 1.9, the case* $\varepsilon = 0.2, K = 2$ *(Region II). Most trajectories converge to either the coexistence equilibrium or the extinction equilibrium. One distinguished trajectory, with initial conditions* $(x(0), y(0)) = (2, b_*)$*, converges to the extinction threshold,* $(\varepsilon, 0)$*.*

slightest perturbation of the initial conditions can change the asymptotic behavior. Therefore, we dismiss the initial condition $(2, b_*)$ as "nongeneric." The dichotomy (1.61) holds for robust solutions.

The example helps us articulate the mathematical value of this concept. We wanted to classify the possible behavior of solutions of (1.41), and it is easier to exclude nonrobust behavior and enumerate just robust cases. This point becomes more emphatic if we allow parameters in (1.41) to vary: claim (1.61) remains valid for all parameters (ε, K) that lie in Region II of Figure 1.9, even though there are substantial (and interesting, to be sure) changes in nongeneric behavior, which we will explore in later chapters.

Using the computer to test the above claims provides further insight. You will find that it is impossible to locate b_* exactly. If b is close to b_*, your solution will linger near the equilibrium $(\varepsilon, 0)$, but eventually it will fall away, either to extinction or to coexistence. To generalize: *It is virtually impossible to reproduce nongeneric behavior on the computer.*

We shall continue to develop the numerous mathematical associations of the term "generic" as we proceed in the book.

1.8.3 A Comparison Theorem

In this book, in focusing on differential *equations*, we neglect differential *inequalities*, except for the following token result. This neglect is not for want of significant theory—whole books (e.g., [84]) have been written about the subject.

Theorem 1.8.1. *Let $x(t)$ be a solution of the scalar ODE $x' = f(x,t)$, where f is $\mathcal{C}^1$, on the interval $0 \le t < T$. If the differentiable function $u(t)$ satisfies*

$$\frac{du}{dt}(t) \le f(u(t),t), \qquad 0 \le t < T, \tag{1.62}$$

and the initial bound $u(0) \le x(0)$, then $u(t) \le x(t)$ for all $t \in [0,T)$.

Although the proof of this result is not hard, we do not give it here. It drops out as a consequence of the existence theory for systems of ODEs; see Exercise 4.15. You may want to try to prove the theorem on your own, or see page 29 of [9].

In Exercise 18 we guide you in using the result to verify the conjectured asymptotic behavior of solutions of the Riccati equation from Section 1.2.2.

Chapter 2

Linear Systems with Constant Coefficients

2.1 Preview

The bulk of this chapter is devoted to homogeneous linear systems of ODEs with real constant coefficients. This means systems of the form

$$\begin{array}{rcl} x_1' &=& a_{11}x_1 + a_{12}x_2 + \ldots + a_{1d}x_d, \\ x_2' &=& a_{21}x_1 + a_{22}x_2 + \ldots + a_{2d}x_d, \\ \vdots & \vdots & \qquad\qquad \vdots \\ x_d' &=& a_{d1}x_1 + a_{d2}x_2 + \ldots + a_{dd}x_d. \end{array} \tag{2.1}$$

(From now on, we shall let d be the dimension of our systems, so that the index n is available for other uses.) The written-out system (2.1) is awkward to read or write, and we shall normally use the vastly more compact linear-algebra notation

$$\mathbf{x}' = A\mathbf{x}, \tag{2.2}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ is a d-dimensional vector of unknown functions, A is a $d \times d$ matrix with real entries, and matrix multiplication is understood in writing $A\mathbf{x}$. In vector notation, an appropriate initial condition for (2.2) is

$$\mathbf{x}(0) = \mathbf{b}, \tag{2.3}$$

where $\mathbf{b} \in \mathbb{R}^d$.

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8_2

In this chapter we will show that the initial value problem (2.2), (2.3) has the unique solution given by the formula[1]

$$\mathbf{x}(t) = e^{At}\mathbf{b}, \tag{2.4}$$

where the exponential of a matrix is defined in analogy with the exponential of a scalar,

$$e^{At} = I + At + \frac{1}{2!}(At)^2 + \frac{1}{3!}(At)^3 + \dots \quad . \tag{2.5}$$

Note that each term in the series is a square matrix, so adding terms is dimensionally consistent. Our first task, the goal of Section 2.2, is to prove that the series (2.5) converges and to derive the basic properties of the matrix exponential. Once we have established the formalism of "norms," the proofs for matrix exponentials are completely analogous to proofs for scalar exponentials.

It turns out that using the series (2.5) is generally not the most convenient way to compute e^{At}. In Section 2.3 we discuss how to use the eigenvalues and eigenvectors of a matrix to compute its exponential.

The following simple calculation motivates the appearance of the eigenvalue problem in solving linear systems of ODEs. We ask whether, in analogy with a scalar linear ODE, there might be solutions of the vector equation (2.2) of the form $e^{\lambda t}$ times a constant. Of course the "constant" has to be a vector for dimensions to be consistent. Thus we refine our question to this: are there any scalars λ and any vectors $\mathbf{v} \in \mathbb{R}^d$ such that

$$\mathbf{x}(t) = e^{\lambda t}\mathbf{v} \tag{2.6}$$

is a solution of (2.2)? Since $\mathbf{x}'(t) = \lambda e^{\lambda t}\mathbf{v}$, the two sides of (2.2) are equal iff

$$A\mathbf{v} = \lambda\mathbf{v},$$

where we have canceled the (nonzero) exponential factor $e^{\lambda t}$. In other words, (2.6) *is a solution of* (2.2) *if and only if* $\mathbf{v}$ *is an eigenvector of* A *with eigenvalue* λ.

To complete the overview of the chapter: in three short additional sections, we discuss the asymptotic behavior of solutions of (2.2) as $t \to \infty$ (Section 2.4), we draw phase portraits for typical two-dimensional linear systems (Section 2.5), and we give formulas for solving an inhomogeneous equation (Section 2.6).

2.2 Definition and Properties of the Matrix Exponential

2.2.1 Preliminaries About Norms

To proceed, we need to define what it means for a series of matrices like (2.5) to converge. We could define convergence of a series of matrices in terms of the convergence

[1]The two factors on the RHS of (2.4) must be written in the order in which they appear.

of each entry, considered as a series of real numbers. However, proofs are smoother if we introduce a metric on the set of matrices and use it to define convergence. The metric is based on how matrices act when multiplying vectors, so we begin with some concepts related to vectors.

For a vector $\mathbf{x} \in \mathbb{R}^d$, the d-dimensional generalization of the Pythagorean theorem suggests that we define the length of $\mathbf{x}$, written $|\mathbf{x}|$, by

$$|\mathbf{x}| = \sqrt{\sum_{j=1}^{d} x_j^2}. \tag{2.7}$$

In analysis it is more common to call the expression $|\mathbf{x}|$ the *norm* of $\mathbf{x}$ rather than its length, and we shall follow that usage. Note that the norm may be expressed as

$$|\mathbf{x}| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}, \tag{2.8}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product on $\mathbb{R}^d$,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{j=1}^{d} x_j y_j. \tag{2.9}$$

In two and three dimensions it is known that

$$\langle \mathbf{x}, \mathbf{y} \rangle = |\mathbf{x}|\,|\mathbf{y}| \cos\theta, \tag{2.10}$$

where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$. In d dimensions, (2.10) is used to *define* the angle between two vectors. The following lemma, the Cauchy–Schwarz inequality, supports this definition by guaranteeing that $\cos\theta$ computed from (2.10) is at most unity in absolute value.

Lemma 2.2.1. *For all vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \le |\mathbf{x}|\,|\mathbf{y}|. \tag{2.11}$$

The inequality is strict unless one of the vectors is a multiple of the other.

Proof. If $\mathbf{y} = \mathbf{0}$, both sides of (2.11) vanish and the inequality is trivial, so we assume $\mathbf{y} \neq \mathbf{0}$. Consider choosing a constant c to minimize

$$|\mathbf{x} + c\mathbf{y}|^2 = |\mathbf{x}|^2 + 2c\langle \mathbf{x}, \mathbf{y} \rangle + c^2 |\mathbf{y}|^2. \tag{2.12}$$

To find the minimum, we differentiate (2.12) with respect to c, set the derivative equal to zero, and solve the resulting trivial linear equation for c, obtaining

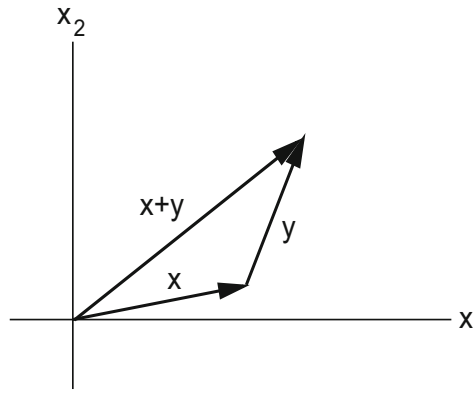

Figure 2.1: *Illustrating the triangle inequality.*

$$c_* = -\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{|\mathbf{y}|^2}.$$

On substituting into (2.12) and combining terms, we obtain

$$|\mathbf{x} + c_*\mathbf{y}|^2 = |\mathbf{x}|^2 - \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{|\mathbf{y}|^2}.$$

Since $|\mathbf{x} + c_*\mathbf{y}|^2 \geq 0$, (2.11) follows. By reexamining the above argument, we may conclude that the inequality is strict unless one vector is a multiple of the other. *(Do this!)* □

In the next lemma we collect several simple but fundamental properties of the norm function.

Lemma 2.2.2. *For any vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ *and scalar* $c \in \mathbb{R}$,

(i) $|\mathbf{x}| \geq 0$, *and* $|\mathbf{x}| = 0$ *iff* $\mathbf{x} = \mathbf{0}$,

(ii) $|c\mathbf{x}| = |c|\,|\mathbf{x}|$,

(iii) $|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}|$.

Inequality (iii) is called the triangle inequality, for reasons suggested by Figure 2.1.

Proof. We leave the derivation of properties (i) and (ii) as an exercise; we merely call your attention to the fact that the same vertical-bar notation is used with slightly different meanings in $|c|$ and in $|\mathbf{x}|$—the absolute value of the scalar and the norm of the vector. Regarding the proof of (iii), observe that

$$|\mathbf{x} + \mathbf{y}|^2 = |\mathbf{x}|^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle + |\mathbf{y}|^2.$$

Using Lemma 2.2.1 to bound the middle term, we compute that

$$|\mathbf{x}+\mathbf{y}|^2 \le |\mathbf{x}|^2 + 2|\mathbf{x}|\,|\mathbf{y}| + |\mathbf{y}|^2 = (|\mathbf{x}| + |\mathbf{y}|)^2.$$

The result follows on taking a square root. □

The norm of a vector specifies its "size." The analogous quantity measuring size for matrices, written with double bars $\|A\|$ and also called the norm, is defined in terms of the operation of a matrix on vectors; specifically, if A is a $d_1 \times d_2$ matrix, we define

$$\|A\| = \max_{|\mathbf{x}|\le 1} |A\mathbf{x}|. \tag{2.13}$$

In this expression, for $A\mathbf{x}$ to be defined, $\mathbf{x}$ must be a d_2-dimensional vector, while $A\mathbf{x}$ has dimension d_1. Thus, if $d_1 \ne d_2$, $|\mathbf{x}|$ and $|A\mathbf{x}|$ are computed with respect to different spaces, even though the notation does not indicate this explicitly.

It follows from compactness that the maximum (2.13) actually exists, as we ask you to show in Exercise 1(a). (Compactness is discussed in Section B.1 of Appendix B. If you have scant experience with this concept, it may seem rather nebulous. We hope that the practice of using compactness in this and other applications below will demystify it.)

Let us collect several useful properties of the matrix norm.

Lemma 2.2.3. *For all matrices A, B of appropriate dimensions, for every vector $\mathbf{x} \in \mathbb{R}^d$, and for every scalar $c \in \mathbb{R}$,*

(i) $\|A\| \ge 0$, *and* $\|A\| = 0$ *iff* $A = 0$,

(ii) $\|cA\| = |c|\,\|A\|$,

(iii) $\|A + B\| \le \|A\| + \|B\|$,

(iv) $|A\mathbf{x}| \le \|A\|\,|\mathbf{x}|$,

(v) $\|AB\| \le \|A\|\,\|B\|$,

(vi) $\|A^2\| \le \|A\|^2$.

In the exercises we ask you to verify these properties. Although this task probably seems less than exciting, we urge you not to skip over it too lightly, because it helps develop proficiency with the use of norms, proficiency you will need later.

The next result relates the norm of a vector or matrix to information about the size of its entries.

Lemma 2.2.4. *If $\mathbf{x} \in \mathbb{R}^d$, then*

$$\max_{1\le j\le d} |x_j| \le |\mathbf{x}| \le \sum_{j=1}^{d} |x_j|,$$

and if A is a $d_1 \times d_2$ matrix with entries $\{a_{jk}\}$, then

$$\max_{1\le j\le d_1}\max_{1\le k\le d_2}|a_{jk}| \le \|A\| \le \sum_{j=1}^{d_1}\sum_{k=1}^{d_2}|a_{jk}|.$$

The vector part of this lemma shouldn't cause you any difficulty; we refer you to Exercise 2 for hints on how to prove the matrix part.

Note that properties (i–iii) in Lemmas 2.2.2 and 2.2.3 are the same. Indeed, every function $\|\cdot\|$ from a vector space to the nonnegative reals satisfying these three properties is called a norm.[2] Given a norm, one may define the distance between two vectors in the space. Thus, for vectors in $\mathbb{R}^d$ and for matrices we define

$$\operatorname{dist}(\mathbf{x},\mathbf{y}) = |\mathbf{x}-\mathbf{y}|, \qquad \operatorname{dist}(A,B) = \|A-B\|, \tag{2.14}$$

respectively.

2.2.2 Convergence

We use the notion of distance to define convergence. Specifically, we say that a sequence $\{A_n\}$ of matrices converges to a limit matrix L if

$$\lim_{n\to\infty}\|A_n - L\| = 0,$$

and we say that an infinite series of matrices $\sum_0^\infty A_n$ converges if the sequence of partial sums $\sum_0^N A_n$ converges as $N\to\infty$. That is, $\sum_0^\infty A_n$ converges to a limit matrix L if for every $\varepsilon > 0$, there is an integer N_0 such that

$$N > N_0 \quad\Longrightarrow\quad \left\|\sum_{n=0}^{N} A_n - L\right\| < \varepsilon.$$

Lemma 2.2.5. *A sequence of matrices $\{A_n\}$ converges if and only each sequence of entries converges, and likewise for an infinite series $\sum_0^\infty A_n$.*

Proof. This result is easily proved using Lemma 2.2.4; we leave the details as a practice exercise. □

We shall say that a series $\sum_0^\infty A_n$ of matrices is *absolutely convergent* if $\sum_0^\infty \|A_n\| < \infty$. Note that the terms of the latter series are nonnegative real numbers.

[2] The vector space need not be finite-dimensional. Indeed, in the next chapter, we shall encounter a norm on an infinite-dimensional space.

Lemma 2.2.6. *If $\sum_0^\infty A_n$ is absolutely convergent, then the series converges; i.e., there exists a matrix L such that*

$$\lim_{N\to\infty} \sum_{n=0}^{N} A_n = L.$$

Moreover, for every integer M,

$$\left\| \sum_{n=0}^{M} A_n - L \right\| \le \sum_{n=M+1}^{\infty} \|A_n\|. \tag{2.15}$$

Proof. In Exercise 2 we ask you to derive the lemma from the analogous result for scalars. The inequality (2.15), a seemingly innocuous generalization of the triangle inequality to an infinite sum, actually requires a limiting argument that we outline in the hints. □

Proposition 2.2.7. *If A is a square matrix, the series $\sum_0^\infty (At)^n/n!$ in (2.5) converges absolutely.*

Proof. For the general term in (2.5) we have the estimate

$$\left\| \frac{(At)^n}{n!} \right\| \le \frac{(|t|\,\|A\|)^n}{n!}, \tag{2.16}$$

and by comparison with the series representation of the scalar exponential $e^{|t|\,\|A\|}$, we see that $\sum_0^\infty \|(At)^n/n!\| < \infty$. □

We will write e^{At} for the sum (2.5), which is guaranteed to exist by Lemma 2.2.6. To apply calculus to the series, we need it to converge uniformly, for which the range of t must be restricted.

Corollary 2.2.8. *For every real number T, the series $\sum_0^\infty (At)^n/n!$ converges uniformly on $\{t : |t| \le T\}$; i.e., for every T and $\varepsilon > 0$, there is an integer N_0 such that if $N > N_0$ and if $|t| \le T$, then*

$$\left\| \sum_{n=0}^{N} \frac{(At)^n}{n!} - e^{At} \right\| < \varepsilon.$$

Proof. By (2.15) and (2.16),

$$\left\| \sum_{n=0}^{N} \frac{(At)^n}{n!} - e^{At} \right\| \le \sum_{n=N+1}^{\infty} \frac{(|t|\,\|A\|)^n}{n!} \le \sum_{n=N+1}^{\infty} \frac{(T\,\|A\|)^n}{n!}.$$

Since the series for $e^{T\|A\|}$ converges, by taking N sufficiently large, the RHS of this inequality can be made arbitrarily small. □

2.2.3 The Main Theorem and Its Proof

Having slogged through a lot of rather dry material, and with more of the same ahead of us, let us reward ourselves by jumping ahead to the main result of Section 2. We suppose A is a square matrix.

Theorem 2.2.9. *For every* $\mathbf{b} \in \mathbb{R}^d$*, the initial value problem*

$$\mathbf{x}' = A\mathbf{x}, \qquad \mathbf{x}(0) = \mathbf{b}$$

has the unique solution that is given by the formula

$$\mathbf{x}(t) = e^{At}\mathbf{b}. \tag{2.17}$$

To prove the theorem, it's back to the salt mines. First we must study the dependence of e^{At} on t. If $\phi(t)$ is a matrix-valued function, we shall say that ϕ is *continuous* at t or that ϕ is *differentiable* at t with derivative L if

$$\lim_{\Delta t \to 0} \phi(t + \Delta t) = \phi(t) \qquad \text{or} \qquad \lim_{\Delta t \to 0} \frac{\phi(t + \Delta t) - \phi(t)}{\Delta t} = L,$$

respectively. It is natural to interpret these limits using norms, but of course, by Lemma 2.2.4, the matrix function ϕ is continuous or differentiable if and only if each entry of ϕ is continuous or differentiable. In the obvious notation, we shall write $\phi'(t)$ for the derivative of ϕ at t, and we shall call ϕ *continuously differentiable* on an interval if ϕ is differentiable at every point in the interval and $\phi'(t)$ is continuous with respect to t there.

Proposition 2.2.10. *The exponential* e^{At} *is a continuously differentiable function of* t *and*

$$\frac{d}{dt} e^{At} = Ae^{At} = e^{At}A.$$

Discussion: To prove this result we need to differentiate an infinite series of functions term by term; in symbols,

$$\frac{d}{dt} \sum_{n=0}^{\infty} f_n(t) = \sum_{n=0}^{\infty} \frac{df_n}{dt}(t). \tag{2.18}$$

Such a result is trivial for *finite* sums, but without additional conditions, it fails badly for infinite sums. This issue is discussed, along with key examples, in Section B.2 of Appendix B. (See also Section 2.8.2 below.) According to Corollary B.2.6, to invoke (2.18) we need for the series on the RHS to converge *uniformly*.

Proof. Each term $(At)^n/n! = t^n(A^n/n!)$ in the series for e^{At} is continuously differentiable; indeed,

$$\frac{d}{dt}\frac{(At)^n}{n!} = nA\frac{(At)^{n-1}}{n!} = n\frac{(At)^{n-1}}{n!}A.$$

Taking a finite sum and observing that $n/n! = 1/(n-1)!$ provided $n \geq 1$, we calculate

$$\frac{d}{dt}\sum_{n=0}^{N}\frac{(At)^n}{n!} = A\left\{\sum_{n=1}^{N}\frac{(At)^{n-1}}{(n-1)!}\right\} = \left\{\sum_{n=1}^{N}\frac{(At)^{n-1}}{(n-1)!}\right\}A.$$

Shifting the summation index in the series on the RHS, we see that

$$\sum_{n=1}^{N}\frac{(At)^{n-1}}{(n-1)!} = \sum_{m=0}^{N-1}\frac{(At)^m}{m!}. \tag{2.19}$$

As $N \to \infty$, the series (2.19) converges to e^{At}; indeed, the convergence is uniform for bounded t. Hence the proposition follows by applying Corollary B.2.6. □

In the next two results we suppress t in e^{At} for brevity. Since A is an arbitrary matrix, we lose no generality by doing this.

Proposition 2.2.11. *The exponential of the zero matrix is the identity; in symbols, $e^0 = I$. For every matrix A, the exponential e^A is invertible and*

$$(e^A)^{-1} = e^{-A}.$$

Proof. It is readily seen from the series expansion (2.5) that $e^0 = I$. Regarding the claim about inverses, for real s let $\phi(s) = e^{As}e^{-As}$. According to the previous proposition, each factor in ϕ is continuously differentiable. In Exercise 1(b) we ask you to prove that the product of two continuously differentiable matrix-valued functions is continuously differentiable and its derivative is given by Leibniz's rule for differentiation of a product. Thus,

$$\frac{d}{ds}\phi(s) = \left(\frac{d}{ds}e^{As}\right)e^{-As} + e^{As}\left(\frac{d}{ds}e^{-As}\right) = e^{As}(+A)e^{-As} + e^{As}(-A)e^{-As} = 0,$$

where we have applied Proposition 2.2.10. Since the derivative vanishes, $\phi(s) = \phi(0) = I$ for all s, in particular for $s = 1$, and the result is proved. □

An incidental consequence of Proposition 2.2.10 is that A and e^{At} commute. More generally, we have the following:

Proposition 2.2.12. *If $AB = BA$, then $Ae^B = e^BA$, $e^Ae^B = e^Be^A$, and*

$$e^{A+B} = e^Ae^B.$$

Proof. We prove only the displayed formula; the other two results are left as exercises. Let

$$\phi(s) = e^{-s(A+B)}e^{sA}e^{sB}.$$

By Leibniz's rule

$$\frac{d}{ds}\phi(s) = e^{-s(A+B)}(-A-B)e^{sA}e^{sB} + e^{-s(A+B)}(A)e^{sA}e^{sB} + e^{-s(A+B)}e^{sA}(B)e^{sB}.$$

In the third term we may commute the middle two factors, e^{sA} and B, and then all three terms add up to zero. Thus $\phi(s) = \phi(0) = I$ for all s. □

It is now a simple matter to prove the main result of this section:

Proof of Theorem 2.2.9. It is obvious that $\mathbf{x}(t) = e^{At}\mathbf{b}$ satisfies the initial condition. To show that it satisfies the equation, just differentiate and apply Proposition 2.2.10. To show uniqueness, suppose $\mathbf{x}(t)$ is one solution and let

$$\mathbf{y}(t) = e^{-At}\mathbf{x}(t). \tag{2.20}$$

Differentiate (2.20) to show that

$$\frac{d}{dt}\mathbf{y}(t) = -e^{-At}A\mathbf{x}(t) + e^{-At}\frac{d}{dt}\mathbf{x}(t).$$

Since $\mathbf{x}(t)$ satisfies the ODE, the two terms in this equation cancel, yielding $d\mathbf{y}(t)/dt = \mathbf{0}$. Thus

$$\mathbf{y}(t) = \mathbf{y}(0) = \mathbf{x}(0) = \mathbf{b},$$

and the uniqueness result follows on multiplying (2.20) by e^{At}. □

2.3 Calculation of the Matrix Exponential

2.3.1 The Role of Similarity

Suppose $\mathbf{x}(t)$ is a solution of the homogeneous linear system

$$\mathbf{x}' = A\mathbf{x}. \tag{2.21}$$

Let us consider a linear change of coordinates for the unknown functions $\{x_j(t)\}$; i.e., let S be a nonsingular matrix and define a new vector of unknown functions by $\mathbf{y}(t) = S\mathbf{x}(t)$. Then we may derive an ODE for $\mathbf{y}(t)$ as follows:

$$\mathbf{y}' = S\mathbf{x}'(t) = SA\mathbf{x} = SAS^{-1}\mathbf{y}.$$

In other words, $\mathbf{y}$ also satisfies a linear homogeneous system of ODEs, and the coefficient matrix in the system for $\mathbf{y}$ is SAS^{-1}, a matrix similar to A in the technical sense of linear algebra. This is an idea worth remembering:

> *Two systems of ODEs with similar coefficient matrices, say* $\mathbf{x}' = A\mathbf{x}$ *and* $\mathbf{y}' = SAS^{-1}\mathbf{y}$*, differ only by the choice of coordinates on* $\mathbb{R}^d$*. In particular, their solutions exhibit exactly the same phenomena.*

Expanding on this theme, the following proposition guarantees that the exponentials of similar matrices are themselves similar.

Proposition 2.3.1. *If* $B = SAS^{-1}$*, then* $e^{Bt} = Se^{At}S^{-1}$.

You are asked to prove this result in Exercise 3.

To illustrate the value of this proposition, let's suppose that the matrix A in (2.21) is diagonalizable over $\mathbb{R}$. Specifically, suppose that $A = S\Lambda S^{-1}$, where

$$\Lambda = \operatorname{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_d) = \begin{bmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ 0 & 0 & \lambda_3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \lambda_d \end{bmatrix}$$

is a diagonal matrix with the (real) eigenvalues of A along its diagonal. By Proposition 2.3.1,

$$e^{At} = Se^{\Lambda t}S^{-1}.$$

Now, for every power, Λ^n is also diagonal, simply $\operatorname{Diag}(\lambda_1^n, \lambda_2^n, \ldots, \lambda_d^n)$. Thus, the series for $e^{\Lambda t}$ converges to the diagonal matrix

$$e^{\Lambda t} = \operatorname{Diag}(e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_d t}),$$

and hence

$$e^{At} = S \operatorname{Diag}(e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_d t})\, S^{-1}. \tag{2.22}$$

In words, we have reduced calculating a matrix exponential to calculating several scalar exponentials and a couple of matrix multiplications.

To take advantage of (2.22), we need to be able to calculate the similarity matrix that diagonalizes A, and the next result tells us how to do this. In this proposition, the notation $\operatorname{Col}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d)$ denotes the matrix whose columns are the specified vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d$. If A is diagonalizable over $\mathbb{R}$, then there is a basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d\}$ for $\mathbb{R}^d$ consisting of eigenvectors of A.

Proposition 2.3.2. *If A is diagonalizable over $\mathbb{R}$ with eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d\}$ that form a basis for $\mathbb{R}^d$ and if $S = \mathrm{Col}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d)$, then*[3]

$$S^{-1}AS = \Lambda,$$

where Λ is the diagonal matrix whose j, j-entry is the eigenvalue λ_j of A associated with the eigenvector $\mathbf{v}_j$.

Proof. For every matrix A, the product $A\mathbf{e}_j$ with the jth unit vector of the standard basis equals the jth column of A. Therefore,

$$\text{the } j\text{th column of } AS = (AS)\mathbf{e}_j = A(S\mathbf{e}_j) = A\mathbf{v}_j = \lambda_j\mathbf{v}_j.$$

On the other hand,

$$\text{the } j\text{th column of } S\Lambda = (S\Lambda)\mathbf{e}_j = S(\Lambda\mathbf{e}_j) = \lambda_j S\mathbf{e}_j = \lambda_j\mathbf{v}_j.$$

In words, we have shown, column by column, that $AS = S\Lambda$, and we multiply this equation by S^{-1} to complete the proof. □

Suppose A is diagonalizable over $\mathbb{R}$. Let's compare the ODE $\mathbf{x}' = A\mathbf{x}$ with the ODE $\mathbf{y}' = \Lambda\mathbf{y}$ obtained from the change of variable $\mathbf{y} = S^{-1}\mathbf{x}$, where S diagonalizes A, i.e., $S^{-1}AS = \Lambda$. The rate of change of x_j depends on all the components of $\mathbf{x}$, while the rate of change of y_j, which equals $\lambda_j y_j$, depends only on the same component y_j. In other words, by diagonalizing A we are performing a change of coordinates on $\mathbb{R}^d$ such that *the new coordinates y_j evolve uncoupled from one another!* Pretty clever, huh?

2.3.2 Two Problematic Cases

The hypothesis in Proposition 2.3.2 may fail in two ways (and both failures may occur together):

- A has repeated eigenvalues but not enough eigenvectors, or
- A has complex eigenvalues.

Let's consider simple special cases before dealing with the general situation.

[3] Up to this point, it does not matter whether we consider the basic equation expressing similarity to be $B = S^{-1}AS$ or $B = SAS^{-1}$. Here, however, there is a difference: the columns of the matrix S such that $S^{-1}AS$ is diagonal are the eigenvectors of A, while the characterization of S^{-1} is less transparent.

The following is the simplest example of a matrix that has a deficiency of eigenvectors:

$$A = \begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix}.$$

It is readily seen that $\lambda = a$ is the only possible eigenvalue of A, but the eigenspace associated with this eigenvalue, $\ker(A - aI)$, is only one-dimensional. However, let us write

$$A = aI + N, \qquad \text{where } N = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Since I and N commute, we have from Proposition 2.2.12 that

$$e^{(aI+N)t} = e^{aIt}e^{Nt} = e^{at}e^{Nt}.$$

Moreover,[4] since $N^2 = 0$, the exponential series for e^{Nt} truncates to just two terms,

$$e^{Nt} = I + Nt = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}, \qquad \text{so} \qquad e^{At} = e^{at}\begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}. \tag{2.23}$$

Regarding the second problematic case, complex eigenvalues, $\lambda = a \pm bi$, occur for the matrix

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}, \tag{2.24}$$

which we write as

$$A = aI + bJ, \qquad \text{where } J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Since I and J commute,

$$e^{(aI+bJ)t} = e^{at}e^{bJt}. \tag{2.25}$$

As with nilpotent matrices, the exponential of J can be computed conveniently using the power-series definition, because of the fact that $J^2 = -I$; thus,

$$J^n = \begin{cases} I & \text{if } n = 0 \pmod 4, \\ J & \text{if } n = 1 \pmod 4, \\ -I & \text{if } n = 2 \pmod 4, \\ -J & \text{if } n = 3 \pmod 4. \end{cases}$$

[4] Recall that a square matrix N is called *nilpotent* if there exists a positive integer k such that $N^k = 0$.

Grouping odd and even powers (this rearrangement of terms will be justified in Exercise 3(c)), we see that

$$e^{bJt} = \left(1 - \frac{1}{2!}(bt)^2 + \frac{1}{4!}(bt)^4 + \dots\right) I + \left(bt - \frac{1}{3!}(bt)^3 + \frac{1}{5!}(bt)^5 + \dots\right) J, \quad (2.26)$$

in which the power series for $\cos bt$ and $\sin bt$ can be recognized. On substituting this formula into (2.25), we obtain

$$e^{At} = e^{at} \begin{bmatrix} \cos bt & -\sin bt \\ \sin bt & \cos bt \end{bmatrix}. \quad (2.27)$$

Incidentally, in Exercise 10 we ask you to prove, with hints, that a 2×2 matrix with nonreal eigenvalues $a \pm bi$ is similar to (2.24). Equation (2.24) is called the *real canonical form* for 2×2 matrices with complex eigenvalues.

The exponential of (2.24) may also be calculated by diagonalization. Indeed, the calculation *looks* pretty simple. Let

$$S = \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix}, \qquad S^{-1} = \frac{1}{2} \begin{bmatrix} 1 & i \\ 1 & -i \end{bmatrix}.$$

The columns of S are eigenvectors of A, so as in Proposition 2.3.2, we have $S^{-1}AS = \Lambda$, where

$$\Lambda = \begin{bmatrix} a + bi & 0 \\ 0 & a - bi \end{bmatrix}.$$

Therefore,

$$e^{At} = S\, e^{\Lambda t}\, S^{-1} = S \begin{bmatrix} e^{(a+bi)t} & 0 \\ 0 & e^{(a-bi)t} \end{bmatrix} S^{-1}.$$

Recalling Euler's formula $e^{ibt} = \cos bt + i \sin bt$ and multiplying out the product, we rederive (2.27).

Unfortunately, there is a gap in this argument; i.e., *we are applying results of the previous section about real matrices to complex matrices.* In fact, all these results *do* carry over to the complex case, but this must be proved. Even the basic definitions are subtly different. Temporarily, for a real vector $\mathbf{x}$ or a real matrix A we shall write $|\mathbf{x}|_{\mathbb{R}}$ or $\|A\|_{\mathbb{R}}$ for the norms defined above. Generalizing to complex vectors, if $\mathbf{z} \in \mathbb{C}^d$, we let

$$|\mathbf{z}|_{\mathbb{C}} = \sqrt{\sum_{j=1}^{d} |z_j|^2}. \quad (2.28)$$

This norm may be calculated from the complex inner product $|\mathbf{z}|_{\mathbb{C}} = \sqrt{\langle \mathbf{z}, \mathbf{z} \rangle_{\mathbb{C}}}$, where

$$\langle \mathbf{z}, \mathbf{w} \rangle_{\mathbb{C}} = \sum_{j=1}^{d} \bar{z}_j w_j, \tag{2.29}$$

with $\bar{z}_j$ denoting the complex conjugate of z_j. If A is a matrix with complex entries, then let

$$\|A\|_{\mathbb{C}} = \max_{|\mathbf{z}|_{\mathbb{C}} \leq 1} |A\mathbf{z}|_{\mathbb{C}}. \tag{2.30}$$

If $\mathbf{x} \in \mathbb{R}^d$, then $|\mathbf{x}|_{\mathbb{R}} = |\mathbf{x}|_{\mathbb{C}}$. If A has real entries, it is not obvious but still true that

$$\|A\|_{\mathbb{R}} = \|A\|_{\mathbb{C}}, \tag{2.31}$$

which we ask you to verify in Exercise 2. Because of (2.31), we may omit the subscript $\mathbb{R}$ or $\mathbb{C}$ in writing norms—if A has complex entries, we understand $\|\cdot\|_{\mathbb{C}}$, and if A has real entries, it doesn't matter which norm we choose. Moreover, in Exercise 2 we ask the dedicated reader to check that properties of norms and the various results about the exponential of matrices in Section 2.2 all carry over to the complex case. In this way, calculating the exponential of (2.24) by diagonalization may be justified.

2.3.3 Use of the Jordan Form

By a *Jordan block* we mean a square matrix of the form

$$B = \begin{bmatrix} \lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda & 1 & \dots & 0 & 0 \\ 0 & 0 & \lambda & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda \end{bmatrix}. \tag{2.32}$$

In words, B has entries λ on the diagonal, 1 on the "superdiagonal," and zeros elsewhere. The matrix B may be of any dimension, including 1×1, in which case B is simply the scalar λ. The diagonal entry λ is the only eigenvalue of B. No matter how large the dimension of B may be, there is only one linearly independent eigenvector.

The Jordan normal-form theorem (see Appendix B of Strang [79]) asserts that every square matrix A is similar to a diagonal array of Jordan blocks; in symbols, $S^{-1}AS = J$, where

$$J = \begin{bmatrix} B_1 & 0 & 0 & \dots & 0 \\ 0 & B_2 & 0 & \dots & 0 \\ 0 & 0 & B_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & B_M \end{bmatrix}. \tag{2.33}$$

Here, for $m = 1, 2, \dots, M$, the matrix B_m is a $d_m \times d_m$ Jordan block, and $\sum_1^M d_m = d$, the dimension of A. To shorten the formulas we shall generalize the notation for diagonal matrices and write J as

$$J = \operatorname{Diag}(B_1, B_2, \dots, B_M).$$

The Jordan canonical-form theorem is spot-on for computing the exponential of a matrix. First observe that

$$\text{if} \quad A = S \operatorname{Diag}(B_1, \dots, B_M)\, S^{-1}, \quad \text{then} \quad e^{At} = S \operatorname{Diag}(e^{B_1 t}, \dots, e^{B_M t})\, S^{-1}. \tag{2.34}$$

Moreover, each Jordan block B_m may be exponentiated by the same method as was used for the 2×2 case above. Specifically to exponentiate a $d \times d$ Jordan block,

- Write $B = \lambda I + N$ where N is the $d \times d$ nilpotent matrix with ones on the superdiagonal.
- Observe that by Proposition 2.2.12, $e^{Bt} = e^{\lambda t} e^{tN}$.
- Calculate e^{tN} with the truncated power series

$$I + tN + \frac{1}{2!}(tN)^2 + \dots + \frac{1}{(d-1)!}(tN)^{d-1}.$$

Painful though it may be, let's write out the result:

$$e^{Bt} = e^{\lambda t} \begin{bmatrix} 1 & t & t^2/2 & \dots & t^{d-2}/(d-2)! & t^{d-1}/(d-1)! \\ 0 & 1 & t & \dots & t^{d-3}/(d-3)! & t^{d-2}/(d-2)! \\ 0 & 0 & 1 & \dots & t^{d-4}/(d-4)! & t^{d-3}/(d-3)! \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & t \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}. \tag{2.35}$$

Despite the messiness of (2.35), this matrix has a simple structure: 1's along the main diagonal, t's along the superdiagonal, $t^2/2$ above that, etc. Thus, the j,k-entry of e^{Bt} equals $e^{\lambda t} t^{k-j}/(k-j)!$ if $k \geq j$ and vanishes otherwise.

In the exercises we ask you to use this information to calculate the exponentials of various matrices. In order to do so, you need to be able to find the Jordan normal form of a matrix. A method for this is explained in Section C.1 of Appendix C. *We urge you to read this section now.* We think you will find our approach to this topic refreshing. In particular, normal forms are determined with natural calculations finding generalized eigenvectors, with no mention of the minimal polynomial.[5] Of course in general, both the Jordan normal form J and the similarity matrix S have complex entries.

As we have seen, the Jordan normal form is perfect for theoretical analysis of e^{At} because it exhibits the structure of the matrix so clearly. However, this normal form is problematic in numerical computation because it is so sensitive to round-off errors. For example, consider the matrices

$$\begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} a & 1 \\ 0 & a+\varepsilon \end{bmatrix}. \tag{2.36}$$

No matter how small $\varepsilon > 0$ may be, the structure of the Jordan normal forms of these two matrices are completely different—a single 2×2 block *vs.* two 1×1 blocks. (In Exercise 5(b) we ask you to compare the exponentials of these matrices.)

2.4 Large-Time Behavior of Solutions of Linear Systems

2.4.1 The Main Results

We shall say that the origin in $\mathbb{R}^d$ is a *sink* (or *attractor*) for a linear homogeneous system $\mathbf{x}' = A\mathbf{x}$ if for every initial condition $\mathbf{b} \in \mathbb{R}^d$,

$$\lim_{t\to\infty} e^{At}\mathbf{b} = \mathbf{0}.$$

The eigenvalues of A provide an elegant test for such behavior.

Theorem 2.4.1. *The origin is a sink for* $\mathbf{x}' = A\mathbf{x}$ *iff*

$$\max_{1\le j\le d} \Re\lambda_j < 0. \tag{2.37}$$

The theorem is an immediate consequence of the following, more quantitative, result.

[5] Although the minimal polynomial is useful in *proving* that transformation to the Jordan form is possible, it is a distraction in *calculating* the Jordan form.

Proposition 2.4.2. *For every $\varepsilon > 0$, there is a constant K such that*

$$e^{\mu t} \le \|e^{At}\| \le K e^{(\mu+\varepsilon)t}, \tag{2.38}$$

where

$$\mu = \max_{1\le j\le d} \Re\lambda_j. \tag{2.39}$$

Remark. Note that the imaginary part $\Im\lambda_j$ has no influence on $\|e^{At}\|$. This behavior may already be seen in the scalar equality $|e^{(\mu+i\nu)t}| = e^{\mu t}$.

Proof. To prove the upper bound, we recall (2.33), the Jordan normal form of A. By (2.34),

$$\|e^{At}\|_{\mathbb{R}} = \|e^{At}\|_{\mathbb{C}} \le \|S\|_{\mathbb{C}}\, \|e^{\mathrm{Diag}(B_1,\dots,B_M)t}\|_{\mathbb{C}}\, \|S^{-1}\|_{\mathbb{C}},$$

and it is easily shown that

$$\|e^{\mathrm{Diag}(B_1,\dots,B_M)t}\|_{\mathbb{C}} = \max_{1\le m\le M} \|e^{B_m t}\|_{\mathbb{C}}. \tag{2.40}$$

Applying (the complex version of) Lemma 2.2.4 to (2.35), we deduce that

$$\|e^{B_m t}\|_{\mathbb{C}} \le d_m^2\, e^{\Re\lambda_m t}\ \max\{1, t^{d_m-1}\} \qquad (t \ge 0). \tag{2.41}$$

Of course $d_m \le d$, so we may estimate (rather extravagantly)

$$\max\{1, t^{d_m-1}\} \le \max\{1, t^{d-1}\}.$$

In Exercise 1(c) we ask you to show that for every $\varepsilon > 0$ there is a constant C such that

$$\max\{1, t^{d-1}\} \le C e^{\varepsilon t}. \tag{2.42}$$

Combining these, we obtain the upper bound in (2.38) with $K = d^2 C\, \|S\|_{\mathbb{C}}\, \|S^{-1}\|_{\mathbb{C}}$.

If the maximum in (2.39) is achieved by a real eigenvalue of A, we may obtain the lower bound in (2.38) by applying e^{At} to an eigenvector of A with eigenvalue μ, say $\mathbf{v}$, where $|\mathbf{v}| = 1$:

$$\|e^{At}\| \ge |e^{At}\mathbf{v}| = e^{\mu t}|\mathbf{v}| = e^{\mu t}.$$

On the other hand, if only complex eigenvalues of A achieve the maximum, we may estimate $\|e^{At}\|_{\mathbb{C}}$ with the same argument and invoke (2.31). □

In Exercise 3(d) we ask you to prove the following two extensions of Proposition 2.4.2, the first trivial and the second more subtle. *Read the hint for the second result.* It introduces a specific similarity transformation you need to know about.

Corollary 2.4.3. *If A is diagonalizable, then (2.38) is satisfied with $\varepsilon = 0$. Indeed, if $A = S\,\mathrm{Diag}(\lambda_1, \ldots, \lambda_d)\,S^{-1}$, then*

$$\|e^{At}\| \leq Ke^{\mu t} \tag{2.43}$$

where $K = \|S\|\,\|S^{-1}\|$.

Corollary 2.4.4. *There is a matrix B that is similar to A for which $\|e^{Bt}\| \leq e^{(\mu+\varepsilon)t}$; i.e., B satisfies (2.38) with the constant $K = 1$.*

2.4.2 Tests for Eigenvalues in the Left Half-Plane

Because of Theorem 2.4.1, it is useful to be able to test whether a matrix has all its eigenvalues in the left half-plane without actually having to find the eigenvalues. For 2×2 and 3×3 matrices, the following two results give simple tests. Please forgive us a homily:

> *Use these results! To our frustration, generations of students have ignored them, wasting their time by calculating eigenvalues when it was not actually necessary.*

Proposition 2.4.5. *If A is a 2×2 matrix with real entries, then $\Re\lambda_j < 0$ iff*

(i) $\mathrm{tr}A < 0$,

(ii) $\det A > 0$.

Proposition 2.4.6. *If A is a 3×3 matrix with real entries, then $\Re\lambda_j < 0$ iff*

(i) $\mathrm{tr}A < 0$,

(ii) $\frac{1}{2}\,\mathrm{tr}A\,[(\mathrm{tr}A)^2 - \mathrm{tr}(A^2)] < \det A$,

(iii) $\det A < 0$.

The proof of Proposition 2.4.5 is left as an exercise. Proposition 2.4.6 is proved in Section C.4 of Appendix C. To motivate Condition (ii) in Proposition 2.4.6, see Exercise 15.

As practice, let's apply Proposition 2.4.6 to determine for what values of the parameter b the eigenvalues of the matrix

$$A(b) = \begin{bmatrix} -1 & -16 & 0 \\ 1 & 0 & b \\ 0 & 1 & -3 \end{bmatrix} \tag{2.44}$$

lie in the left half-plane. A simple calculation shows that

$$\mathrm{tr}A = -4, \qquad \det A = b - 48, \qquad \mathrm{tr}(A^2) = 2b - 22.$$

Condition (i) is always satisfied, and the remaining two require that

$$\text{(ii) } 3b < 28, \qquad \text{(iii) } b < 48.$$

Thus, by the proposition, the eigenvalues of $A(b)$ satisfy $\Re\lambda_j < 0$ if and only if $b < 28/3$. (What happens to the eigenvalues of $A(b)$ as b passes through $28/3$? Another reason to do Exercise 15.)

The following result is sometimes useful to test for oscillatory behavior in a 2×2 system. It may be derived by examining the quadratic formula for the eigenvalues of A.

Proposition 2.4.7. *A real 2×2 matrix has complex eigenvalues if and only if*

$$(\operatorname{tr} A)^2 < 4 \det A.$$

2.5 Classification and Phase Portraits for 2×2 Systems

In two dimensions there is an extensive vocabulary classifying linear systems[6]

$$\mathbf{x}' = A\mathbf{x} \tag{2.45}$$

based on the eigenvalues of A. Suppose that A is nonsingular, so that $\mathbf{x} = \mathbf{0}$ is an isolated equilibrium of (2.45). In Table 2.1, we have listed several descriptive terms for the equilibrium of such an equation. (Note that these terms are not mutually exclusive.) It will be useful below to have these terms available, and we recommend that during some "captive time," such as on a long flight, you commit them to memory.

Each type of equilibrium has a characteristic phase-plane portrait, as we now explore.

2.5.1 Saddles and Nodes

If A has real nonzero eigenvalues, then (2.45) has either a saddle or a node. Because equations with similar coefficient matrices exhibit the same phenomena, we assume that A is in Jordan normal form. Supposing for the moment that A is diagonalizable, we take

$$A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tag{2.46}$$

where $\lambda_k \neq 0$. The general solution of $\mathbf{x}' = A\mathbf{x}$ is given by

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} x_1(0)e^{\lambda_1 t} \\ x_2(0)e^{\lambda_2 t} \end{bmatrix}.$$

[6] In Chapter 6, we will extend this classification to equilibria of nonlinear systems.

Name	Eigenvalues of coefficient matrix	Characterizing inequalities
Sink	Both eigenvalues in LHP, $\{\Re\lambda < 0\}$	$\det A > 0$, $\mathrm{tr}A < 0$
Source	Both eigenvalues in RHP, $\{\Re\lambda > 0\}$	$\det A > 0$, $\mathrm{tr}A > 0$
Saddle	Both eigenvalues real, of opposite sign	$\det A < 0$
Node	Both eigenvalues real, of same sign	$0 < 4\det A \leq (\mathrm{tr}A)^2$
Focus	Eigenvalues complex conjugates, $\Re\lambda \neq 0$	$0 < (\mathrm{tr}A)^2 < 4\det A$
Center	Eigenvalues complex conjugates, $\Re\lambda = 0$	$0 = \mathrm{tr}A < \det A$

Table 2.1: *Descriptive terms for the isolated equilibria of two-dimensional linear systems. A sink may be either a node or a focus, and likewise for a source. A node may be either a sink or a source, and likewise for a focus.*

If $x_1(0) = 0$, then the trajectory follows the x_2-axis. Otherwise, by eliminating t, we find that the trajectory lies along the graph of a power function,

$$x_2 = C\,|x_1|^p, \tag{2.47}$$

where $p = \lambda_2/\lambda_1$ and C is a constant. Recalling our earlier terminology, we see that (2.47) specifies *orbits*.

If (2.45) has a saddle, then $p < 0$ in (2.47). Reindexing the eigenvalues if necessary, we may assume that $\lambda_1 < 0 < \lambda_2$. A phase portrait for such a case is shown in Figure 2.2(a). (To the eye, all cases with $\lambda_1 < 0 < \lambda_2$ tend to look qualitatively similar, so we show the phase portrait only for one specific choice of eigenvalues.) The coordinate axes are special orbits; these trajectories converge to the origin in one of the limits $t \to -\infty$, $t \to \infty$. Other trajectories, which are curved, tend to infinity in both limits, approaching one of the coordinate axes asymptotically.

If (2.45) has a node, then $p > 0$. If both eigenvalues of A are negative, then all trajectories converge to zero, as shown in Figure 2.2(b), where for definiteness we have assumed $\lambda_2 < \lambda_1 < 0$. Equal eigenvalues lead to phase portraits qualitatively different from Figure 2.2(b); this includes both the case $A = \lambda I$ and that in which A is similar to a Jordan block. In Exercise 17 we ask you to draw phase portraits for such cases. *Do this exercise!* Being fluent in visualizing flows enriches the subject enormously.

If at a node the eigenvalues of A are negative, as in Figure 2.2(b), then the origin is a sink, and we say that (2.45) has a *stable* node. By contrast, if the eigenvalues are positive, then trajectories follow the same orbits—given by (2.47) if A is diagonalizable—but move away from the origin; i.e., the origin is a source, and we say that (2.45) has an *unstable* node.

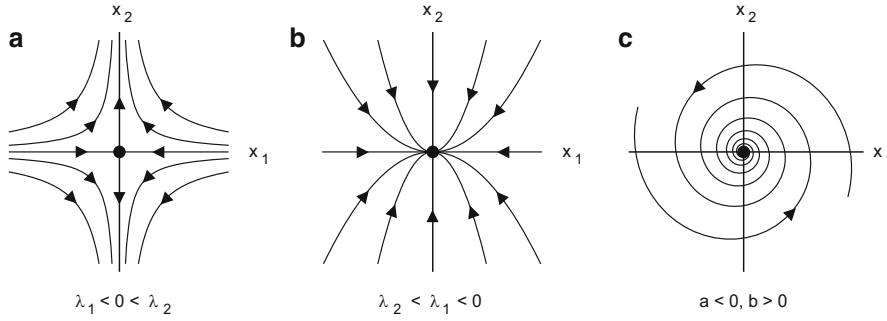

Figure 2.2: *Representative phase portraits for* (2.45). *(a) A saddle: A given by* (2.46) *with* $\lambda_1 < 0 < \lambda_2$. *(b) A stable node: A given by* (2.46) *with* $\lambda_2 < \lambda_1 < 0$. *(c) A stable focus: A given by* (2.48) *with* $a < 0$ *and* $b > 0$.

2.5.2 Foci and Centers

If A has complex eigenvalues, i.e., if (2.45) has a focus or a center, then it is more convenient to study the real canonical form,

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \tag{2.48}$$

where $b > 0$, than the Jordan normal form. If $a < 0$, it may be seen from (2.27) that trajectories of (2.45) spiral inward toward the origin (cf. Figure 2.2c), while if $a > 0$, they spiral outward. The equilibria in these two cases are called a *stable focus* and an *unstable focus*, respectively. If $a = 0$, the trajectories are concentric circles, and the equilibrium is called a *center*.

2.5.3 Additional Remarks

(a) For a general matrix, the phase portrait of (2.45) could be determined by applying an appropriate linear transformation to the phase portrait associated with the corresponding Jordan or real canonical form. In practice, however, much simpler calculations often suffice to visualize the phase portrait.

For example, suppose $\det A < 0$; i.e., suppose the equation has a saddle point. For definiteness assume $\lambda_1 < 0 < \lambda_2$, with associated eigenvectors $\mathbf{v}_1, \mathbf{v}_2$, respectively. The general solution of (2.45) is

$$C_1 e^{\lambda_1 t}\mathbf{v}_1 + C_2 e^{\lambda_2 t}\mathbf{v}_2. \tag{2.49}$$

Trajectories with $C_2 = 0$ lie on the straight line $\mathbb{R}\{\mathbf{v}_1\}$, the span of $\{\mathbf{v}_1\}$, which corresponds to the x_1-axis in Figure 2.2(a), and those with $C_1 = 0$, on $\mathbb{R}\{\mathbf{v}_2\}$,

which corresponds to the x_2-axis. Trajectories with both C_1 and C_2 nonzero correspond to the hyperbola-like curves in Figure 2.2(a); these trajectories are asymptotic to the eigenspaces $\mathbb{R}\{\mathbf{v}_1\}$ or $\mathbb{R}\{\mathbf{v}_2\}$ as $t \to \pm\infty$. For many purposes, this level of detail is adequate.

In Exercise 18 we outline analogous shortcuts for other cases.

(b) If zero is an eigenvalue of A, then there are infinitely many solutions of $A\mathbf{x} = \mathbf{0}$, each of which represents a constant solution, or equilibrium, of $\mathbf{x}' = A\mathbf{x}$.

(c) Although visualization is more difficult, phase-portrait ideas extend to higher dimensions. Consider, for example, a 3×3 matrix A that has real eigenvalues $\lambda_1 < 0 < \lambda_2 \leq \lambda_3$ with (possibly generalized) eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$, respectively. Trajectories that lie in the λ_1-eigenspace $\mathbb{R}\{\mathbf{v}_1\}$ tend to the origin as $t \to \infty$. Trajectories that lie in the plane $\mathbb{R}\{\mathbf{v}_2, \mathbf{v}_3\}$, what we might call the positive eigenspace, converge to the origin as $t \to -\infty$. All other trajectories are curves that are asymptotic to these eigenspaces as $t \to \pm\infty$.

2.6 Solution of Inhomogeneous Problems

If $\mathbf{x}$ satisfies an inhomogeneous linear equation with constant coefficients, say

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(t), \tag{2.50}$$

then the unique solution to the IVP with $\mathbf{x}(0) = \mathbf{b}$ is given by[7]

$$\mathbf{x}(t) = e^{At}\mathbf{b} + \int_0^t e^{A(t-s)}\mathbf{g}(s)\, ds. \tag{2.51}$$

The derivation of this result, which plays a central role in Chapter 6, is posed as Exercise 4(b).

Incidentally, in Exercise 4(c) we ask you to show that if $\mathbf{g}(t)$ is a finite linear combination of terms of the form $e^{\mu t}\mathbf{v}$, where $\mu \in \mathbb{C}$ and $\mathbf{v}$ is a constant vector, then (2.50) may be solved explicitly (and (2.51) ignored). Since μ may be complex, this construction includes cases in which the inhomogeneous term is a trigonometric function.

Does (2.51) generalize to systems with variable coefficients,

$$\mathbf{x}' = A(t)\mathbf{x} + \mathbf{g}(t)\,? \tag{2.52}$$

[7]Note that the integrand in (2.51) is vector-valued. Such an integral can be interpreted componentwise, giving a collection of ordinary, scalar, integrals.

If $A(t_1)$ and $A(t_2)$ commute for all t_1, t_2, there is a generalization analogous to (1.18), *but not otherwise*; cf. Exercise 24.

2.7 Exercises

After the core exercises, there are subsections of linear-algebra problems and of problems involving drawing phase-plane portraits, followed by the usual PHD exercises.

2.7.1 Core Exercises

The core exercises have the following purposes:

To deal with unfinished business (general analysis, norms, exponentials, other)	1–4
To calculate exponentials of specific matrices	5
To develop general facility with systems of ODEs	6–8
To apply the tests of Propositions 2.4.5, 2.4.6, and 2.4.7	9

1. (a) Prove that the maximum in Equation (2.13) exists and is finite.

   Hint: Use compactness, which is discussed in Section B.1 of Appendix B.

   (b) Derive Leibniz's rule for differentiation of the product of two matrix-valued functions of a scalar variable:

$$\frac{d}{dt}[\phi(t)\psi(t)] \;=\; \phi'(t)\psi(t) + \phi(t)\psi'(t).$$

   Remark: Similarly, Leibniz's rule extends to other linear-algebra products, such as a matrix times a vector, as occurs for example in (2.51).

   (c) Verify (2.42).

   Hint: Here is a warmup problem that isolates the main issue: Prove that for every positive power p,

$$\max_{0\le t<\infty} t^p e^{-t}$$

   exists and is finite. The exact value of the maximum can in fact be computed using calculus, but this is unnecessary, and it is useful training to do this exercise merely with estimation, as follows. Argue that

$$\lim_{t\to\infty} t^p e^{-t} = 0.$$

   Deduce that there is a constant T such that $t^p e^{-t} < 1$ when $t > T$. Therefore,

$$\max_{0\le t<\infty} t^p e^{-t} \le \max\left\{ \max_{0\le t\le T} t^p e^{-t},\ 1 \right\},$$

   and the maximum over $[0, T]$ is finite by compactness.

2. (a) Prove Lemmas 2.2.2, 2.2.3, 2.2.4, 2.2.5, and 2.2.6.

Hint for Lemma 2.2.4: In the matrix part of the lemma, the lower bound for $\|A\|$ may be obtained by applying A to a basis vector $\mathbf{e}_k$. The upper bound may be obtained by writing A as a sum of $d_1 \times d_2$ matrices, each of which has only one nonzero entry.

Hint for Lemma 2.2.6: Derive the existence of L from Proposition B.2.4 in Appendix B, the analogous result for scalars. Regarding (2.15), for $N > M$, add and subtract terms A_n for n between M and N to write

$$\sum_{n=0}^{M} A_n - L = \left[\sum_{n=0}^{N} A_n - L\right] - \left[\sum_{n=M+1}^{N} A_n\right].$$

For every $\varepsilon > 0$, if N is large enough, the first sum on the right is bounded by ε. Since the second sum is finite, it may be estimated using the triangle inequality. Now let N tend to infinity.

(b) Establish Equation (2.31) for a matrix with real entries.

Hint: Show that $\|A\|_{\mathbb{R}} \leq \|A\|_{\mathbb{C}}$ because $\|A\|_{\mathbb{C}}$ is calculated as the maximum over a larger set. Use

$$|A(\mathbf{x} + i\mathbf{y})|_{\mathbb{C}}^2 = |A\mathbf{x}|_{\mathbb{R}}^2 + |A\mathbf{y}|_{\mathbb{R}}^2$$

to prove the reverse inequality.

(c) Rederive all the results of Sections 2.2 and 2.3 for matrices with complex entries.

Remark: This one is for the truly dedicated reader.

(d) Calculate that if A has dimension $d \times 1$ (i.e., A is a column vector) or $1 \times d$ (i.e., A is a row vector), then $\|A\|$ is just the norm of the vector.

(e) Prove that if A has a (real or complex) eigenvalue λ, then $\|A\| \geq |\lambda|$.

(f) Show that if S is an invertible matrix, then $\|S\| \, \|S^{-1}\| \geq 1$.

Remark: The quantity $\|S\| \, \|S^{-1}\|$ is often called the *condition number* of the matrix S; see Section 7.2 of Strang [79] regarding why this quantity is important.

3. (a) Prove the first two assertions in Proposition 2.2.12.

(b) Prove Proposition 2.3.1.

Hint: You may prove this result either by comparing terms in the series for the two sides of the equation or by differentiating $e^{-Bs} S e^{As}$.

(c) Justify the rearrangement of terms in Equation (2.26).

Hint: For a finite sum, trivially

$$\begin{aligned}\sum_{n=0}^{N} \tfrac{1}{n!}(bJt)^n &= \left(1 - \tfrac{1}{2!}(bt)^2 + \ldots + \tfrac{1}{(N-1)!}(bt)^{N-1}\right) I \\ &\quad + \left(bt - \tfrac{1}{3!}(bt)^3 + \ldots + \tfrac{1}{N!}(bt)^N\right) J,\end{aligned}$$

where for convenience we assume $N = 1 \pmod 4$, so that the last coefficient in each of the truncated series is positive; now let $N \to \infty$.

As a general rule, rearrangements of an absolutely convergent series are always permissible (see Theorem 3.55 of Rudin [68]). Errors that may result from rearranging terms in a series that is not absolutely convergent are illustrated in the Pearls of Appendix B.

(d) Prove Corollaries 2.4.3 and 2.4.4.

Hint for Corollary 2.4.4: Here is a way to rescale the entries of a $d \times d$ matrix A that is worth remembering: If c is a nonzero real number and if $S = \mathrm{Diag}(1, c, c^2 \ldots, c^{d-1})$, then

$$S^{-1}AS = \begin{bmatrix} a_{11} & ca_{12} & c^2a_{13} & \ldots & c^{d-1}a_{1d} \\ c^{-1}a_{21} & a_{22} & ca_{23} & \ldots & c^{d-2}a_{2d} \\ c^{-2}a_{31} & c^{-1}a_{32} & a_{33} & \ldots & c^{d-3}a_{3d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c^{-(d-1)}a_{d1} & c^{-(d-2)}a_{d2} & c^{-(d-3)}a_{d3} & \ldots & a_{dd} \end{bmatrix}. \quad (2.53)$$

Note that the power of c in the i,j-entry of the product is $j - i$. Thus if A is upper triangular (e.g., in Jordan form), then we can choose c small to make the off-diagonal entries of $S^{-1}AS$ as small as we like. If A has real eigenvalues, you may use such a scaling to prove the corollary. If A has complex eigenvalues, the same basic argument works but with some minor technical modifications.

4. (a) Prove Propositions 2.4.5 and 2.4.7.

(b) Show that (2.51) is the unique solution of the IVP for (2.50).

Hint: Argue that (2.50) is equivalent to

$$\frac{d}{dt}\left(e^{-At}\mathbf{x}\right) = e^{-At}\mathbf{g}(t)$$

and integrate.

(c) Show that if $\mathbf{g}(t) = e^{\mu t}\mathbf{v}$, where $\mu \in \mathbb{C}$ *is not an eigenvalue of* A and $\mathbf{v}$ is a constant vector, then (2.50) has a particular solution of the form $e^{\mu t}\mathbf{w}$.

Discussion: Because of linearity, this idea may be used to find a particular solution of (2.50) if $\mathbf{g}(t)$ is a finite linear combination of such terms. Of course, the general solution of (2.50) is a particular solution plus a solution of the homogeneous equation $\mathbf{x}' = A\mathbf{x}$.

If μ is an eigenvalue of A, particular solutions may still be constructed, but one or more components of the solution may have the form $te^{\mu t}$ or perhaps

with higher powers of t. A general theory is a little messy, but you might find it instructive to consider first a diagonalizable 2×2 matrix and then a 2×2 Jordan block.

5. Compute e^{tA} for the following matrices, and if initial conditions are provided, solve the initial value problem for the system $\mathbf{x}' = A\mathbf{x}$:

(a)
$$\begin{bmatrix} 3 & 1 \\ -1 & 1 \end{bmatrix}, \qquad \mathbf{x}(0) = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

(b)
$$\begin{bmatrix} a & 1 \\ 0 & a+\varepsilon \end{bmatrix}$$

Remark: Be sure to compare your answer for e^{tA} with (2.23), the exponential of the Jordan block with eigenvalue a.

(c)
$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & -1 \end{bmatrix}, \qquad \mathbf{x}(0) = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

(d)
$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

(e)
$$\begin{bmatrix} 0 & 1 \\ -\kappa^2 & 0 \end{bmatrix}$$

where κ is a real constant.

Hint: Here is a shortcut: invoking (2.53) with $c = \kappa$, we see that this matrix is similar to a matrix of the form (2.24), whose exponential we have already calculated.

(f)
$$\begin{bmatrix} a & -b & 1 & 0 \\ b & a & 0 & 1 \\ 0 & 0 & a & -b \\ 0 & 0 & b & a \end{bmatrix}$$

(g) Make up your own examples of matrices to exponentiate. (If you choose an example for which the calculations are too messy, you will nevertheless have learned something.)

6. Suppose A is a real square matrix with at least one eigenvalue λ, real or complex, such that $\Re\lambda < 0$. Show that the linear system $\mathbf{x}' = A\mathbf{x}$ has at least one nonzero real solution $\mathbf{x}(t)$ such that

$$\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{0}.$$

7. Rederive (2.23) by explicitly solving the ODE $\mathbf{x}' = A\mathbf{x}$.

   Hint: The x_2-equation does not involve x_1; solve this equation first and then attack the x_1-equation.

8. Write the general linear homogeneous 2×2 system

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

   in polar coordinates.[8]

9. (a) If

$$A = \begin{bmatrix} 1 & a \\ -1 & b \end{bmatrix},$$

   for what values of a, b is the origin a sink for the two-dimensional system $\mathbf{x}' = A\mathbf{x}$?

   (b) For what values of a, b does the matrix in Part (a) have complex eigenvalues?

   (c) If

$$A = \begin{bmatrix} -2 & a & 0 \\ 1 & 2 & b \\ 0 & 3 & -1 \end{bmatrix},$$

   for what values of a, b is the origin a sink for the three-dimensional system $\mathbf{x}' = A\mathbf{x}$?

2.7.2 Practice with Linear Algebra

10. *Introduction:* This problem identifies the similarity transformation that reduces a 2×2 real matrix A with eigenvalues $a \pm ib$, where $b \neq 0$, to its real canonical form C, i.e., the matrix S such that

$$S^{-1}AS = C = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}. \tag{2.54}$$

[8] The introduction of polar coordinates represents a *nonlinear* change of coordinates in an ODE. Linear changes of coordinates were exploited in Section 2.3.

Let $\mathbf{u} = \mathbf{v} + i\mathbf{w}$ be an eigenvector of A with eigenvalue $a - ib$; thus

$$A(\mathbf{v} + i\mathbf{w}) = (a - ib)(\mathbf{v} + i\mathbf{w}). \tag{2.55}$$

Let S be the 2×2 matrix $\mathrm{Col}(\mathbf{v}, \mathbf{w})$. Deduce (2.54) from (2.55).

11. Find the 2×2 matrix A that has the indicated eigenvalues and eigenvectors:

e-value	e-vector
$-1/\varepsilon$	$(1, 1)$
-1	$(1, 1+\delta)$

Hint: This exercise is easy if you make use of Proposition 2.3.2.

Discussion: The ulterior message in the exercise is to observe that $\|A\|$ may become large if ε and/or δ tends to zero. This behavior is not surprising for $\varepsilon \to 0$, since the eigenvalue $-1/\varepsilon$ gives a lower bound for $\|A\|$. You may find it surprising for $\delta \to 0$, which simply means that the two eigenvectors of the matrix become nearly parallel even as its eigenvalues remain bounded.

12. (a) Find matrices A, B such that $e^A e^B \neq e^{A+B}$.

Remark: It is informative to experiment and find your own example. There is no need to go beyond 2×2 matrices.

(b) Find matrices A, B such that e^{tA} and e^{tB} converge to zero but $\|e^{t(A+B)}\|$ tends to infinity.

Hint: You can find an example in which A is a Jordan block and $B = A^T$. Of course your matrices provide another example of what Part (a) requests.

13. (a) Show that if A is a square matrix with $\|A\| < 1$, then the Neumann series

$$I + A + A^2 \ldots \tag{2.56}$$

converges to $(I - A)^{-1}$.

(b) Let A be a $d \times d$ matrix with real entries. Show that the powers A^n tend to zero as $n \to \infty$ iff every eigenvalue of A satisfies

$$|\lambda_k(A)| < 1, \quad k = 1, \ldots, d.$$

Remark: With a bit more care you can show that this eigenvalue hypothesis implies that the series (2.56) converges.

(c) *True or false:* The largest eigenvalue (in absolute value, ties possible) of a square matrix A satisfies

$$|\lambda_{\max}| = \lim_{n\to\infty} \|A^n\|^{1/n}.$$

14. (a) Verify that the inequalities in the third column of Table 2.1 characterize the indicated classifications of the eigenvalues of A.

    (b) Write (1.24) as a 2×2 first-order system and classify how the type of the equilibrium at the origin depends on the parameters, according to the categories of Table 2.1.

15. Show that if A is a 3×3 matrix with eigenvalues $\alpha, \pm i\nu$, where α and ν are real, then the two sides of Condition (ii) in Proposition 2.4.6 are equal.

    Discussion: This exercise is intended to demystify Condition (ii) in Proposition 2.4.6. To make the issues more concrete, let's recall the application of Proposition 2.4.6 to the matrix (2.44). If $b = 0$, then $A(b)$ has block lower-triangular form, and its eigenvalues may be calculated explicitly to verify that they lie in the left half-plane. *Do they stay there as b varies?* Well, the eigenvalues of $A(b)$ vary continuously[9] with b, so the eigenvalues will indeed stay in the left half-plane *unless* an eigenvalue crosses the imaginary axis. *How can such a crossing occur?* Generically, in one of two ways, as sketched in Figure 2.3: (i) a single real eigenvalue may cross zero or (ii) a pair of complex conjugate eigenvalues can move across the imaginary axis. In the first case, the transition occurs when a real eigenvalue vanishes, which implies that the determinant also vanishes; thus, Condition (iii) tests for such a transition. This exercise shows that Condition (ii) tests for the other possible transition, i.e., for a pair of complex conjugate eigenvalues to lie on the imaginary axis.

    The seemingly minor issue in this exercise in fact segues to a major phenomenon appearing later in the book, i.e., Hopf bifurcation in Section 8.7.

2.7.3 Practice Sketching Phase Portraits

16. For each of the following matrices A, sketch the phase portrait for $\mathbf{x}' = A\mathbf{x}$. Your sketch should include information about the asymptotic behavior of trajectories. For two-dimensional systems, classify the origin according to the categories of Table 2.1.

$$(a)\ A = \begin{bmatrix} -2 & 4 \\ -2 & 2 \end{bmatrix}, \qquad (b)\ A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad (c)\ A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 2 \\ 0 & -2 & -1 \end{bmatrix}.$$

17. Sketch the phase portraits for (2.45) in cases in which it has a stable node different from Figure 2.2(b), i.e., when (i) $A = \lambda I$, where $\lambda < 0$ or (ii) A is a Jordan block with eigenvalue $\lambda < 0$.

[9] In the case of simple eigenvalues, this result is easily proved with the implicit function theorem. Multiple eigenvalues raise more subtle issues that we ignore here, but continuous dependence, appropriately interpreted, still holds. (Cf. Section C.3 of Appendix C.)

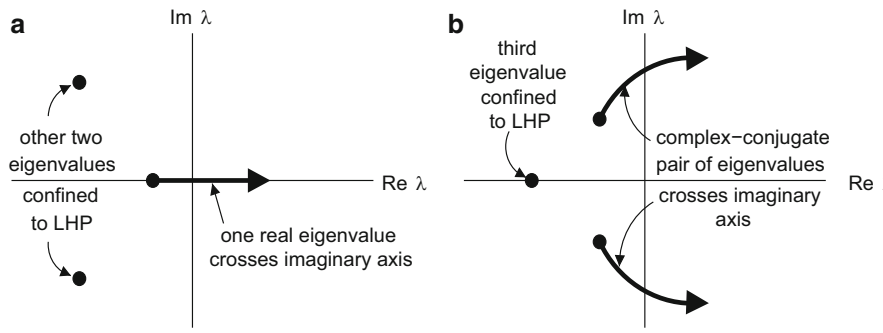

Figure 2.3: *Varying the entries in a* 3×3 *matrix can cause real parts of eigenvalue(s) to change sign in one of two generic ways: (a) a single, real eigenvalue moves across the imaginary axis, or (b) a pair of complex-conjugate eigenvalues moves across the imaginary axis.*

Remarks: Incidentally, in Exercise 26 we ask you to derive an equation for the orbits in Case (ii). Plots for unstable nodes may be derived by observing that $\mathbf{x}' = A\mathbf{x}$ and its time-reversed equation $\mathbf{x}' = -A\mathbf{x}$ have the same *orbits*, but with the flow direction reversed.

18. *Introduction:* Item (a) in Section 2.5.3 describes a shortcut for visualizing the phase portrait of (2.45) when this equation has a saddle point. In this exercise we describe analogous shortcuts when the equation has a stable node.

(a) Suppose A has eigenvalues $\lambda_2 < \lambda_1 < 0$ with eigenvectors $\mathbf{v}_1, \mathbf{v}_2$. Using the general solution (2.49), show that the eigenspaces $\mathbb{R}\{\mathbf{v}_1\}$ and $\mathbb{R}\{\mathbf{v}_2\}$ are straight trajectories and that all other trajectories are curved and approach the origin as $t \to \infty$, asymptotic to $\mathbb{R}\{\mathbf{v}_1\}$.

Remark: If $\lambda_1 = \lambda_2 < 0$ and A is diagonalizable, then $A = \lambda_1 I$, whose phase portrait you found in the previous exercise.

(b) Suppose that $\lambda_1 = \lambda_2 < 0$ and A is not diagonalizable. Let $\mathbf{v}_1$ be an eigenvector. Show that the eigenspace $\mathbb{R}\{\mathbf{v}_1\}$ is a straight trajectory and that all other trajectories are curved and approach the origin as $t \to \infty$, asymptotic to $\mathbb{R}\{\mathbf{v}_1\}$.

Hint: Choose a generalized eigenvector such that $(A - \lambda_1 I)\mathbf{v}_2 = \mathbf{v}_1$ and analyze the general solution of the equation,

$$e^{\lambda_1 t}\left[C_1\mathbf{v}_1 + C_2(\mathbf{v}_2 + t\mathbf{v}_1)\right].$$

(Note that the approach to the eigenspace is algebraic, rather than exponential as in Part (a).)

Remark: Near a center, trajectories are ellipses, and motion around them may be clockwise or counterclockwise. Similarly, the phase portrait for a focus is a distorted spiral.

19. Suppose A has eigenvalues $\lambda_2 < 0$, $\lambda_1 = 0$ with eigenvectors $\mathbf{v}_1, \mathbf{v}_2$, i.e., a singular limiting case of Part (a) of Exercise 18. Choose hypothetical directions for $\mathbf{v}_1, \mathbf{v}_2$ and sketch the phase portrait of $\mathbf{x}' = A\mathbf{x}$.

2.7.4 PHD Exercises

20. (a) Derive the following exact formula for the norm of a matrix:

$$\|A\| = \sqrt{\lambda_{max}(A^T A)}.$$

Hint: Deduce from the definition of $\|A\|$ that

$$\|A\|^2 = \max_{|\mathbf{x}|\leq 1} \langle A^T A\mathbf{x}, \mathbf{x}\rangle$$

and invoke the spectral theorem for symmetric matrices to estimate $\langle A^T A\mathbf{x}, \mathbf{x}\rangle$.

(b) Use this result to find the norm of

$$\begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix}$$

and compare this answer with the estimates of Lemma 2.2.4.

21. *Introduction:* If A is a $d \times d$ matrix with real entries, define the Euclidean norm of A,

$$\|A\|_E = \left\{ \sum_{j,k=1}^{d} a_{jk}^2 \right\}^{1/2}.$$

Determine which of the following is true and prove it:

(i) For all A different from zero, $\|A\| < \|A\|_E$.

(ii) For all A, $\|A\| \leq \|A\|_E$, with equality occurring for at least one nonzero matrix A.

(iii) There is a matrix A such that $\|A\| > \|A\|_E$.

22. *Introduction:* This exercise illustrates how the power series expansion of the function $\ln(1+x)$,

$$\ln(1+x) = x - x^2/2 + x^3/3 - x^4/4 + \ldots,$$

can be used to define a logarithm of a nonsingular matrix.

(a) Consider the Jordan block

$$A = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}$$

where $\lambda \neq 0$, which we rewrite as $A = \lambda(I + N)$, where I is the identity and N is nilpotent. Use the above series to define a provisional logarithm

$$\log A = (\log \lambda)I + N - N^2/2 + \dots ,$$

where of course the series terminates.

(b) Show that $e^{\log A} = A$.

Remark: A logarithm of any nonsingular matrix can be found by performing calculations of this type on each block in the Jordan form.

23. (a) Find the general solution for two coupled harmonic oscillators, say

$$\begin{aligned} x'' + x &= \varepsilon y \\ y'' + y &= \varepsilon x \end{aligned}$$

where for simplicity we take identical oscillators. Write your solution in the form

$$\begin{bmatrix} x \\ y \end{bmatrix} = C_1 \cos(\omega_1 t - \alpha_1) \begin{bmatrix} 1 \\ 1 \end{bmatrix} + C_2 \cos(\omega_2 t - \alpha_2) \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Hint: You could reduce this problem to a four-dimensional first-order system, but actually the linear algebra is simpler if you don't. Write your system as

$$\begin{bmatrix} x'' \\ y'' \end{bmatrix} + \begin{bmatrix} 1 & -\varepsilon \\ -\varepsilon & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{0}$$

and look for solutions of the form $(x, y) = e^{\lambda t}\mathbf{v}$.

(b) Solve the IVP with initial conditions

$$x(0) = 1, \qquad x'(0) = y(0) = y'(0) = 0.$$

Assuming that $\varepsilon \ll 1$, graph the energy (times 2) in the individual oscillators $(x')^2 + x^2$ and $(y')^2 + y^2$ as functions of time.

Remark: You can do this analytically, but it's fine if you prefer to use the computer. You will find that the energy shifts back and forth between the two oscillators.

24. (a) Consider the variable-coefficient system (2.52), and suppose that $A(t_1)$ and $A(t_2)$ commute for all t_1, t_2. Let

$$\overline{A}(t) = \int_0^t A(s)\, ds.$$

Show that

$$\mathbf{x}(t) = e^{\overline{A}(t)}\mathbf{b} + \int_0^t e^{\overline{A}(t)-\overline{A}(s)}\mathbf{g}(s)\, ds$$

solves the IVP for (2.52) with initial condition $\mathbf{x}(0) = \mathbf{b}$.

(b) Solve (2.52) in the case

$$A(t) = \begin{bmatrix} 0 & t \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{g}(t) \equiv \mathbf{0}$$

and show that this solution differs from what the above formula produces.

Remark: Even if in (2.52), the eigenvalues $\lambda_k(t)$ of the coefficient matrix $A(t)$ satisfy $\Re\lambda_k(t) < 0$ for all time (and say $\mathbf{g}(t) \equiv \mathbf{0}$), solutions of this system may suffer exponential growth. (For an example, you may peek ahead: the first-order system derived from the scalar ODE in Exercise 3.13, say with a small amount of friction added, exhibits this behavior.)

25. Suppose that the eigenvalues of A satisfy the hypothesis (2.37). Show that if $\lim_{t\to\infty} \mathbf{g}(t) = \mathbf{0}$, then the solution (2.51) of the inhomogeneous ODE (2.50) also tends to zero as $t \to \infty$.

26. Find an equation for the orbits of $\mathbf{x}' = A\mathbf{x}$ in the case that A is a 2×2 Jordan block with eigenvalue $\lambda < 0$.

Hint: One orbit is the x_1-axis. For other orbits you can express x_1 as a function of x_2 and one arbitrary constant.

2.8 Pearls of Wisdom

2.8.1 Alternative Norms

There are many different norms in the literature used to measure the size of vectors and matrices. For example, for vectors, two common choices are

$$|\mathbf{x}|_1 = \sum_{j=1}^{d} |x_j| \quad \text{and} \quad |\mathbf{x}|_\infty = \max_{1\le j\le d} |x_j|,$$

which give rise to matrix norms

$$\|A\|_1 = \max_{|\mathbf{x}|_1 \le 1} |A\mathbf{x}|_1 \quad \text{and} \quad \|A\|_\infty = \max_{|\mathbf{x}|_\infty \le 1} |A\mathbf{x}|_\infty.$$

2.8.2 Nondifferentiable Limits and the Cantor Set

In proving Proposition 2.2.10, we invoked Corollary B.2.6 of Appendix B to justify term-by-term differentiation of a series $\sum_n f_n(t)$ of functions. Here's another counterexample, in addition to those in the appendix, to show that for this result to hold, the series $\sum_n f_n'(t)$ of derivatives must converge uniformly. It's more convenient to formulate this example in the context of the analogous, but equivalent, issue for sequences. Let

$$\phi(t) = \begin{cases} 0\,, & t \le 0 \\ t^2(3-2t), & 0 < t < 1 \\ 1\,, & t \ge 1. \end{cases} \tag{2.57}$$

Graph this function! You will find that it is continuously differentiable.[10] For $n = 1, 2, \ldots$, let $f_n(t) = \phi(nt)$. Then

$$\lim_{n\to\infty} f_n(t) = \begin{cases} 1, & t > 0, \\ 0, & t \le 0, \end{cases}$$

is not even continuous, let alone differentiable. However, each function f_n is continuously differentiable, and for every t, the sequence $\{f_n'(t)\}$ converges (to zero).

Let's have some fun with this example by combining it with the Cantor set. The Cantor set is constructed by repeatedly removing middle thirds from intervals, as indicated in Figure 2.4(a). Although the figure may be clearer than formulas, let us not neglect the latter. To start, we define

$$\begin{aligned} C_1 &= [0,1] \sim (1/3, 2/3) \\ C_2 &= C_1 \sim \{(1/9, 2/9) \cup (7/9, 8/9)\}\,; \end{aligned}$$

we continue inductively for $n \ge 2$,

$$C_{n+1} = C_n \sim \bigcup_{k=0}^{3^n-1} \left(\frac{3k+1}{3^{n+1}}, \frac{3k+2}{3^{n+1}} \right); \tag{2.58}$$

[10] For your information, with a little extra work, it is possible to construct a $\mathcal{C}^\infty$ function that equals zero for $t \le 0$ and equals unity for $t \ge 1$. See [73], pp. 48–50.

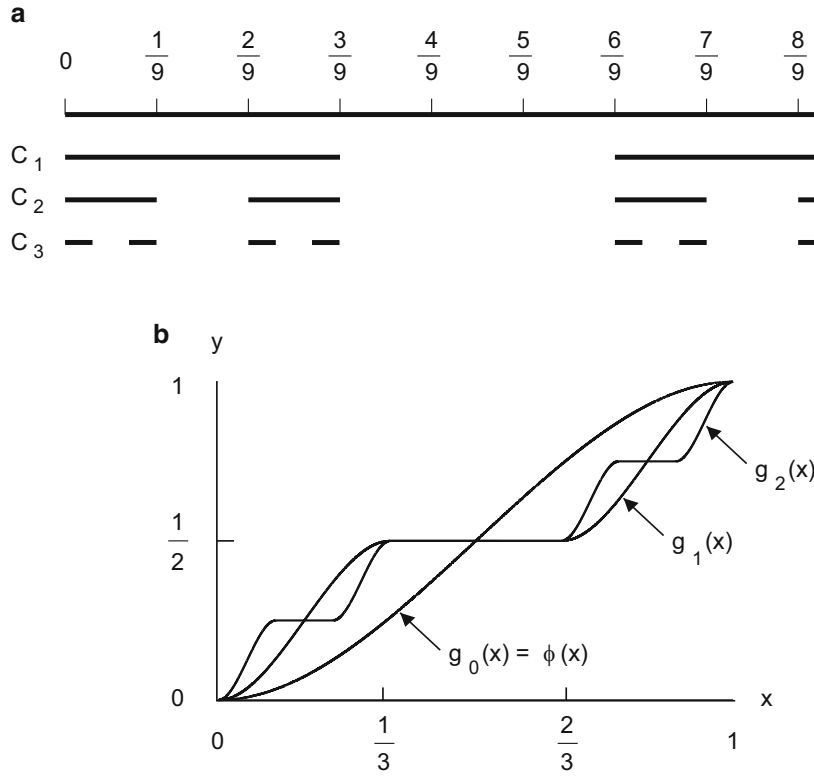

Figure 2.4: *(a) Schematic illustration of the construction of the Cantor set. (b) The functions $g_0(x)$, $g_1(x)$, and $g_2(x)$ in the construction of the Cantor function.*

and finally we let

$$C = \bigcap_{n=1}^{\infty} C_n.$$

You may show by induction that C_n consists of 2^n disjoint intervals of length $1/3^n$. Note that many of the intervals enumerated in the union in (2.58) don't remove anything from C_n; for example, if $n = 1$, the interval $(4/9, 5/9)$ is included in the union in (2.58), but this interval and more was already removed in defining C_1.

The Cantor *function* is the limiting function of the sequence whose construction is indicated in Figure 2.4(b). In formulas, define functions $g_n : [0, 1] \to \mathbb{R}$, where $g_0(x) = \phi(x)$ and

$$g_{n+1}(x) = \begin{cases} g_n(3x)/2, & \text{if } 0 \le x \le 1/3, \\ 1/2, & \text{if } 1/3 < x < 2/3, \\ 1/2 + g_n(3(x-2/3))/2, & \text{if } 2/3 \le x \le 1. \end{cases}$$

Although it is not readily apparent from this recursive definition, in fact $g_{n+1}(x) = g_n(x)$ for all $x \in [0,1] \sim C_n$. Moreover, $|g_{n+1}(x) - g_n(x)| \le 2^{-n}$ for all x. Thus, the sequence converges uniformly, so $g(x) = \lim_n g_n(x)$ is continuous. But this function has the paradoxical properties that somehow it manages to climb from zero to one as x ranges over $[0,1]$ but $g(x)$ is constant on intervals of total length

$$\frac{1}{3} + \frac{2}{3^2} + \frac{2^2}{3^3} + \ldots + \frac{2^{n-1}}{3^n} + \cdots = 1.$$

In particular, even though each function $g_n(x)$ is $\mathcal{C}^1$, the limit is far from differentiable.

2.8.3 More on Generic Behavior

In Section 1.8.2 we discussed the concept "generic" in choosing initial conditions for an IVP. In this section we apply the concept in another context, i.e., generic behavior for a homogeneous linear system with constant coefficients,

$$\mathbf{x}' = A\mathbf{x}. \tag{2.59}$$

This time we actually give a definition for the term.

The set of equations (2.59) is parametrized by $d \times d$ matrices, or the Euclidean space $\mathbb{R}^D$, where $D = d^2$. We shall say that a property $\mathcal{P}$ of matrices (or of such equations) is *generic* if it holds for all matrices in some open, dense set $\Omega \subset \mathbb{R}^D$. For example:

Proposition 2.8.1. *The set of $d \times d$ matrices whose eigenvalues are all simple is generic.*

For 2×2 matrices, the set of matrices with only simple eigenvalues may be characterized as the *complement* of the zero set of the discriminant polynomial on $\mathbb{R}^4$,

$$\{A : (a_{11} - a_{22})^2 + 4a_{12}a_{21} = 0\},$$

which implies that it is open and dense. Section 7.3 of [40] proves the result for general d.

A square matrix A is often called *hyperbolic*[11] if none of its eigenvalues lies on the imaginary axis; in symbols, $\Re\lambda_k(A) \neq 0$. For instance, in Table 2.1, all the equilibria described there are hyperbolic except a center.

Proposition 2.8.2. *The set of hyperbolic $d \times d$ matrices is generic.*

For 2×2 matrices, the set of hyperbolic matrices is the complement of

$$\{A : \det A = 0\} \cup \{A : \det A \geq 0,\ \mathrm{tr}A = 0\},$$

which is open and dense. Section 7.3 of [40] proves the result for general d.

By the following result, proved in Section 7.3 of [40], we may require both simple eigenvalues and hyperbolicity and still have a generic property.

Lemma 2.8.3. *If $\Omega_1, \ldots, \Omega_n$ are open, dense subsets of $\mathbb{R}^D$, then the intersection*

$$\Omega_1 \cap \cdots \cap \Omega_n$$

is also open and dense.

The following intuitive ideas about genericity are more important for us than a rigorous understanding of the supporting theory. (This is why we only give references for proofs, even though they are not difficult.) If Ω is dense, then for every matrix A_0, there are matrices arbitrarily close to A_0 that belong to Ω. On the other hand, if Ω is open, then for any $A_0 \in \Omega$, all matrices sufficiently close to A_0 also belong to Ω. In more intuitive language, *matrices that possess a generic property are both plentiful and robust.* In particular, if an equation is known only approximately (e.g., the coefficients are determined from experimental data), there is no real loss of generality in assuming that A has some convenient generic property.

Regarding the discussion of genericity in Section 1.8.2, note that what we identified as generic initial conditions there—the complement of the blue trajectory in Figure 1.11—is open and dense in the first quadrant.

Genericity ideas are also useful in understanding *nonlinear* ODEs. However, we shall keep the discussion informal in order to sidestep technical issues arising from working in infinite-dimensional spaces.

---

[11] We regard this terminology as unfortunate, since the word hyperbolic already has so many uses in mathematics, but it is well established and we adopt it. The term "hyperbolic" derives from the simplest system with such a coefficient matrix,

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix},$$

whose solutions move along hyperbolas $\{x^2 - y^2 = C\}$ where C is a constant. In general, however, orbits for hyperbolic systems have at best only a qualitative resemblance to hyperbolas, and possibly none at all.

Chapter 3

Nonlinear Systems: Local Theory

In this chapter we state and prove the basic existence and uniqueness theorems (in Sections 3.2 and 3.3, respectively) for the initial value problem (IVP) for systems of nonlinear ODEs. For the moment we consider only autonomous systems, say

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}) \;=\; \begin{bmatrix} F_1(x_1, x_2, \ldots, x_d) \\ F_2(x_1, x_2, \ldots, x_d) \\ \vdots \\ F_d(x_1, x_2, \ldots, x_d) \end{bmatrix} \tag{3.1}$$

where $\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^d$; or more generally, we may assume that $\mathbf{F}$ is defined only on an open subset $\mathcal{U} \subset \mathbb{R}^d$. In Section 3.4, we discuss extensions of the theory to nonautonomous systems.

3.1 Two Counterexamples

While the IVP for a linear system $\mathbf{x}' = A\mathbf{x}$ has a unique solution that exists for all time, nonlinear equations are not so kind to us. We begin the chapter with two examples to illustrate the difficulties.

Example 1: *Without special conditions on* $\mathbf{F}$*, the IVP for* (3.1) *may possess a solution only for a finite, possibly very short, time.* To see this, consider the IVP for a scalar unknown function $x(t)$,

$$x' = x^2, \qquad x(0) = 1. \tag{3.2}$$

The equation may be solved using separability:

$$\frac{dx}{x^2} = dt, \qquad \text{which integrates to} \qquad -\frac{1}{x} = t + C.$$

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8_3

Solving for x and imposing the initial condition (IC) to deduce that $C = -1$, we obtain

$$x(t) = \frac{1}{1-t}. \tag{3.3}$$

You may check that this formula satisfies both the equation and the initial condition, but note: *this solution exists only for* $t < 1$.

Strictly speaking, formula (3.3) makes sense provided $t \neq 1$, but in most modeling contexts, continuation of the solution beyond the blowup time, i.e., to $\{t > 1\}$, is rejected on physical grounds. Suppose, for example, that x represents a population; reemergence of x with large negative values after the singularity at $t = 1$ is nonsensical. We shall say that the solution ceases to exist at $t = 1$.

The blowup of (3.3) at $t = 1$ may be understood as follows: From the equation, we see that $x' > 0$, so the solution is always increasing. As $x(t)$ grows, the equation $x' = x^2$ forces the solution to increase ever more quickly, and the growth accelerates out of control in a finite time. It is instructive to compare (3.2) with the linear equation $x' = x$, whose solutions also grow without bound but remain finite for all time. The difference between these equations is that in (3.2), the RHS x^2 grows faster than linearly as $x \to \infty$; finite-time blowup is driven by such superlinear growth (cf. Exercise 2).

Incidentally, in Chapter 4 we will give sufficient conditions to guarantee that the solution to an IVP exists for all time.

Example 2: *Without special conditions on* $\mathbf{F}$*, solutions of the IVP for* (3.1) *need not be unique.* To see this, consider another scalar IVP,[1]

$$x' = \sqrt{|x|}, \qquad x(0) = 0. \tag{3.4}$$

Note that $x' \geq 0$, so that for positive t, we have $x(t) \geq 0$, and thus we may drop the absolute value in the equation. Solving $x' = \sqrt{x}$ by separability, we obtain the general solution

$$x(t) = \left[\frac{t+C}{2}\right]^2.$$

Choosing $C = 0$ to satisfy the IC, we get the solution $x(t) = t^2/4$. Please check that this function satisfies the equation for positive time.[2]

However, note that $x(t) \equiv 0$ also solves both the equation and the initial conditions. In other words, *the solution to this IVP is not unique.* Moreover, the situation

[1]This example is merely a simplification of Exercise 4(b) of Chapter 1 that makes the calculations all but trivial.

[2]This function is *not* a solution of (3.4) for $t < 0$. Geometrically, it follows from the fact that $x' \geq 0$ that $x(t) \leq 0$ when $t < 0$. Analytically, the issue is that the radical means the *positive* square root, so $\sqrt{t^2/4} = |t|/2 \neq (d/dt)(t^2/4)$. This example provides an unwelcome, but salutary, reminder that explicit solutions of ODEs need to be checked thoroughly.

is worse than is so far apparent: for every constant $t_0 \geq 0$, the function

$$x(t) = \begin{cases} 0 & \text{if } t \leq t_0 \\ (t - t_0)^2/4 & \text{if } t > t_0 \end{cases} \tag{3.5}$$

is a continuously differentiable solution of the equation that also satisfies the initial condition. *(Check this!)* In other words, there are *infinitely* many solutions of the IVP.

The problems with this example stem from the singularity of $\sqrt{|x|}$ at the origin. As we shall see in Section 3.3, uniqueness may be guaranteed if $\mathbf{F}$ is sufficiently regular, e.g., continuously differentiable.

3.2 Local Existence Theory

3.2.1 Statement of the Existence Theorem

In this section we formulate and prove the fundamental existence theorem for the IVP for (3.1). We will assume that the function $\mathbf{F}$ on the RHS of the ODE is continuous, and in fact we will impose a stronger condition that we now describe.

If S is a subset of $\mathbb{R}^{d_1}$, a function $\mathbf{F} : S \to \mathbb{R}^{d_2}$ is called *Lipschitz continuous*, or simply *Lipschitz*, if there is a constant L such that for all points $\mathbf{x}, \mathbf{y} \in S$,

$$|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})| \leq L|\mathbf{x} - \mathbf{y}|. \tag{3.6}$$

Condition (3.6) is much more restrictive than mere continuity. For example, on the real line, the function $F(x) = \sqrt{|x|}$, which appeared in the example (3.4), is continuous but not Lipschitz continuous.[3]

Let $\mathcal{U}$ be an open subset of $\mathbb{R}^{d_1}$, and let $\mathbf{F} : \mathcal{U} \to \mathbb{R}^{d_2}$. We shall call $\mathbf{F}$ *locally Lipschitz* if for every point $\mathbf{x}_0 \in \mathcal{U}$ there is a neighborhood $\mathcal{V}$ of $\mathbf{x}_0$ such that the restriction $\mathbf{F}|\mathcal{V}$ is Lipschitz.[4] For example, on the real line the function $F(x) = x^2$ is locally Lipschitz, even though it is not Lipschitz on the whole line. *(Check this!)* Sometimes, to emphasize the distinction from locally Lipschitz, we shall say that a function is *globally Lipschitz* on S to mean that it is Lipschitz on S.

We shall study the IVP (3.1) under the assumption that $\mathbf{F}$ is locally Lipschitz. As Proposition 3.2.2 below shows, a $\mathcal{C}^1$ function is locally Lipschitz; in fact, being locally Lipschitz is only slightly less restrictive than being $\mathcal{C}^1$. In most equations that we

[3]The Cantor function, which was discussed in the Pearls of Chapter 2, gives a more dramatic illustration of this point. For yet more drama, see Exercise 20.

[4]Incidentally, if $\mathbf{F}$ is locally Lipschitz on $\mathcal{U}$, then the restriction of $\mathbf{F}$ to a compact subset of $\mathcal{U}$ is Lipschitz; see Proposition 3.3.2. The proof is trickier than you might expect.

study in later chapters, $\mathbf{F}$ will actually be $\mathcal{C}^1$; we consider the more general condition here only because, as we shall see, the local existence and uniqueness theory is so well suited to a Lipschitz condition.

Below we give our fundamental existence theorem for the IVP, whose formulation we repeat for convenient reference: given $\mathbf{b} \in \mathbb{R}^d$, find a continuously differentiable function[5] $\mathbf{x}(t)$ such that

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0) = \mathbf{b}. \tag{3.7}$$

Theorem 3.2.1. *Let $\mathcal{U} \subset \mathbb{R}^d$ be open and contain the initial data $\mathbf{b}$, and let $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$ be locally Lipschitz. Then there exist an interval $(-\eta, \eta)$ and a $\mathcal{C}^1$ function $\mathbf{x} : (-\eta, \eta) \to \mathcal{U}$ such that (3.7) is satisfied.*

The construction of $\mathbf{x}(t)$ will show that the solution is unique on the possibly very short interval $(-\eta, \eta)$ in the theorem. In fact, uniqueness holds in far greater generality, as we shall show in Section 3.3. Anticipating those stronger results, we ignore the information regarding uniqueness that may be obtained through proving Theorem 3.2.1.

We must develop substantial preliminaries before we are ready to prove Theorem 3.2.1, which we do in Section 3.2.5. Although the results of the next subsection are not actually needed to prove the theorem, they will be needed later, and we include them here because they help elucidate Lipschitz continuity.

3.2.2 $\mathcal{C}^1$ Implies Lipschitz Continuity

Proposition 3.2.2. *If $\mathcal{U} \subset \mathbb{R}^{d_1}$ is open and $\mathbf{F} : \mathcal{U} \to \mathbb{R}^{d_2}$ is $\mathcal{C}^1$, then $\mathbf{F}$ is locally Lipschitz.*

Incidentally, the converse of this result is not true. For example, on the real line, $F(x) = |x|$ is locally Lipschitz (globally Lipschitz, in fact) despite not being differentiable at the origin.[6]

We offer two proofs of the proposition, the first only for scalar-valued functions of one variable and the second for the general case. *The second proof is greatly to be preferred.* We offer the first proof primarily because it illustrates a common bad habit among beginning analysis students—overreliance on the mean value theorem—and we want to have an identified target to shoot down.

Proof 1 (only for $d_1 = d_2 = 1$). Given $x_0 \in \mathcal{U}$, choose a compact interval $\mathcal{I}$ such that

$$x_0 \in \operatorname{Int} \mathcal{I} \subset \mathcal{I} \subset \mathcal{U},$$

[5]This might be an appropriate place to repeat our warning from Chapter 1: the symbols $\mathbf{x}$, $\mathbf{y}$, etc., may denote a point in $\mathbb{R}^d$ (as in (3.6)) or a vector-valued function of time (as in (3.7)), depending on context. Whenever you see one of these letters, ask yourself which usage is intended.

[6]In the Pearls, Section 3.6.1, we invoke a little analysis to construct a more dramatic example of this point.

where Int means interior, and let $L = \max_{\mathcal{I}} |F'|$. Given $x, y \in \mathcal{I}$, we have from the mean value theorem that there is a point ξ between x and y such that

$$F(x) - F(y) = F'(\xi)(x - y),$$

so by the choice of L as the maximum, we deduce

$$|F(x) - F(y)| \leq L|x - y|. \qquad \square$$

This proof may be generalized to all d_1, but there are problems with it if $d_2 > 1$. The difficulty is that one needs to apply the mean value theorem separately to each of the d_2 components of $\mathbf{F}$. This is not impossible, but it results in a clunky proof. The following is a far more elegant alternative, and we urge you to absorb the clever application of the fundamental theorem of calculus in Lemma 3.2.3. In this result, and indeed throughout the book, we use the notation $\mathbf{DF}$ for the Jacobian matrix of $\mathbf{F}(x)$ with entries $\partial F_k/\partial x_j$, where $j = 1, \ldots, d_1$ and $k = 1, \ldots, d_2$.

Proof 2 (for general d_1, d_2). Given $\mathbf{x}_0 \in \mathcal{U}$, we choose an (open) ball

$$B(\mathbf{x}_0, \varepsilon) = \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x} - \mathbf{x}_0| < \varepsilon\}$$

whose closure is contained in $\mathcal{U}$. We will show that $\mathbf{F}$ is Lipschitz on this ball with Lipschitz constant derived from the norm of the Jacobian,

$$L = \max_{\mathbf{z} \in \overline{B(\mathbf{x}_0, \varepsilon)}} \|\mathbf{DF}(\mathbf{z})\|. \tag{3.8}$$

Let two points $\mathbf{x}, \mathbf{y} \in \overline{B(\mathbf{x}_0, \varepsilon)}$ be given. We isolate the following simple lemma as a separate result so that we can refer to it later.

Lemma 3.2.3. *In the above notation,*

$$\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y}) = \left\{ \int_0^1 \mathbf{DF}(\mathbf{y} + s(\mathbf{x} - \mathbf{y}))\, ds \right\} \cdot (\mathbf{x} - \mathbf{y}). \tag{3.9}$$

Proof. Note that the argument of $\mathbf{DF}$ in the above integrand,

$$\boldsymbol{\ell}(s) = \mathbf{y} + s(\mathbf{x} - \mathbf{y}), \qquad 0 \leq s \leq 1, \tag{3.10}$$

defines the line segment from $\mathbf{y}$ to $\mathbf{x}$, which is entirely contained in $\overline{B(\mathbf{x}_0, \varepsilon)} \subset \mathcal{U}$. Thus, the composition $\mathbf{F} \circ \boldsymbol{\ell} : [0,1] \to \mathbb{R}^{d_2}$ is defined and $\mathcal{C}^1$. By the fundamental theorem of calculus,[7]

$$\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y}) = \int_0^1 \frac{d}{ds} \left[\mathbf{F} \circ \boldsymbol{\ell}\right](s)\, ds. \tag{3.11}$$

According to the chain rule (reviewed in Section B.3 of Appendix B), $(d/ds)[\mathbf{F} \circ \boldsymbol{\ell}] = \mathbf{DF} \cdot \boldsymbol{\ell}'$, and differentiation of (3.10) yields $\boldsymbol{\ell}'(s) = \mathbf{x} - \mathbf{y}$. Thus (3.9) follows. □

Proof 2 of Proposition 3.2.2, concluded. Taking norms in (3.9), we deduce from Property (iv) in Lemma 2.2.3 that

$$|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})| \leq \left\| \int_0^1 \mathbf{DF}(\mathbf{y} + s(\mathbf{x} - \mathbf{y}))\, ds \right\| \, |\mathbf{x} - \mathbf{y}|. \tag{3.12}$$

We estimate the first factor on the RHS by the triangle inequality for matrix-valued integrals,[8]

$$\left\| \int_0^1 \mathbf{DF}(\mathbf{y} + s(\mathbf{x} - \mathbf{y}))\, ds \right\| \leq \int_0^1 \|\mathbf{DF}(\mathbf{y} + s(\mathbf{x} - \mathbf{y}))\|\, ds, \tag{3.13}$$

and we bound the RHS of (3.13) using the maximum of the integrand. Substituting into (3.12), we obtain

$$|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})| \leq \left\{ \max_{0 \leq s \leq 1} \|\mathbf{DF}(\mathbf{y} + s(\mathbf{x} - \mathbf{y}))\| \right\} |\mathbf{x} - \mathbf{y}|. \tag{3.14}$$

Since the line segment between $\mathbf{x}$ and $\mathbf{y}$ is contained in $\overline{B(\mathbf{x}_0, \varepsilon)}$, we obtain the required bound $|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})| \leq L|\mathbf{x} - \mathbf{y}|$, where L is defined by (3.8). □

We ask you to verify that the exact same construction as in the second proof above supports the following extension of Proposition 3.2.2.

Corollary 3.2.4. *If $\mathcal{U} \subset \mathbb{R}^{d_1}$ is open, if $\mathbf{F} : \mathcal{U} \to \mathbb{R}^{d_2}$ is $\mathcal{C}^1$, and if $\mathcal{K} \subset \mathcal{U}$ is compact and convex, then $\mathbf{F}|\mathcal{K}$ is Lipschitz with Lipschitz constant*

$$L = \max_{\mathcal{K}} \|\mathbf{DF}\|. \tag{3.15}$$

---

[7] Note that the functions in (3.11) are vector-valued. Each component of this equation is simply the standard one-variable fundamental theorem of calculus.

[8] For a scalar function it is obvious from the interpretation of the integral as an area that

$$\left| \int_0^1 f(s)\, ds \right| \leq \int_0^1 |f(s)|\, ds.$$

To derive the analogous result for a vector- or matrix-valued function, apply the triangle inequality to approximating Riemann sums and then take limits.

Remark. This result may be strengthened: even if $\mathcal{K}$ is not convex, $\mathbf{F}|\mathcal{K}$ is still Lipschitz. We will prove this in Proposition 3.3.2. However, the proof involves an intricate application of compactness, and a simple estimate like (3.15) for a Lipschitz constant is not available. We postpone the proof of the stronger result until it is actually needed.

3.2.3 Reformulation of the IVP as an Integral Equation

The proof of Theorem 3.2.1 is based on analyzing an equivalent integral equation, i.e., (3.16) below. The integral equation is more tractable than (3.7), because integration is a much less singular operation than differentiation. Also note that the two separate equations in (3.7)—the ODE and the initial conditions—are combined in a single integral equation.

In the proposition, it suffices if $\mathbf{F}$ is merely continuous, so we temporarily weaken our hypotheses.

Proposition 3.2.5. *Let $\mathcal{U} \subset \mathbb{R}^d$ be open, and let $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$ be continuous. If $\mathbf{x} \in \mathcal{C}^1((\alpha, \beta), \mathcal{U})$ satisfies the IVP (3.7), then $\mathbf{x}$ satisfies the integral relation*

$$\mathbf{x}(t) = \mathbf{b} + \int_0^t \mathbf{F}(\mathbf{x}(s))\, ds, \qquad \alpha < t < \beta. \tag{3.16}$$

Conversely, if $\mathbf{x}$ is continuous on (α, β) and satisfies (3.16), then $\mathbf{x}$ is $\mathcal{C}^1$ and satisfies (3.7).

The proof of this result is a straightforward application of the fundamental theorem of calculus, and we leave it as an exercise.

Despite its appearance, equation (3.16) is *not* a formula that tells us what the solution is, because we need to know $\mathbf{x}(s)$ in order to evaluate the integral.

3.2.4 The Contraction-Mapping Principle

In Chapter 2 we encountered norms on a vector space X, i.e., a function $\|\cdot\| : \mathsf{X} \to [0, \infty)$ such that for all vectors $\mathbf{x}, \mathbf{y} \in \mathsf{X}$ and scalar $c \in \mathbb{R}$,

$$\begin{array}{ll} (a) & \|\mathbf{x}\| \geq 0, \quad \text{and } \|\mathbf{x}\| = 0 \text{ iff } \mathbf{x} = \mathbf{0}, \\ (b) & \|c\mathbf{x}\| = |c|\, \|\mathbf{x}\|, \\ (c) & \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|. \end{array} \tag{3.17}$$

Previously, the norms were defined on finite-dimensional spaces of vectors or matrices. For the proof of Theorem 3.2.1, we need a norm on the infinite-dimensional space $\mathcal{C}([-\eta, \eta], \mathbb{R}^d)$, the set of continuous functions from the closed interval $[-\eta, \eta]$

into $\mathbb{R}^d$; specifically, for $\mathbf{x} \in \mathcal{C}([-\eta, \eta], \mathbb{R}^d)$ we define

$$\|\mathbf{x}\| = \max_{-\eta \le t \le \eta} |\mathbf{x}(t)|. \tag{3.18}$$

Since $[-\eta, \eta]$ is compact, the maximum exists. In Exercise 1(b), you are asked to show that (3.18) satisfies the axioms (3.17). Convergence of a sequence in $\mathcal{C}([-\eta, \eta], \mathbb{R}^d)$ with respect to this norm is simply uniform convergence of a sequence of functions.

Let $\mathsf{X}, \|\cdot\|$ be a normed vector space. A sequence $\{\mathbf{x}_n\} \subset \mathsf{X}$ is called *Cauchy* if for every $\varepsilon > 0$, there is an integer N such that

$$m, n > N \quad \Longrightarrow \quad \|\mathbf{x}_m - \mathbf{x}_n\| < \varepsilon.$$

The space in question is called *complete* if every Cauchy sequence converges, i.e., if there exists an element $\mathbf{x}_\infty$ in X such that

$$\lim_{n \to \infty} \|\mathbf{x}_n - \mathbf{x}_\infty\| = 0.$$

These definitions generalize familiar concepts for real numbers. Incidentally, a complete normed vector space is called a *Banach* space. Thus, in this terminology, the following proposition asserts that $\mathcal{C}([-\eta, \eta], \mathbb{R}^d)$ is a Banach space.

Proposition 3.2.6. *The space $\mathcal{C}([-\eta, \eta], \mathbb{R}^d)$ is complete.*

This theorem is merely a restatement of the result (Theorem B.2.2 in Appendix B) that the uniform limit of a sequence of continuous functions is itself continuous.

After two more definitions, we will be ready to state and prove the contraction-mapping principle. Let S be a subset of a normed linear space[9] X, and let $\mathfrak{T} : \mathsf{S} \to \mathsf{S}$ be some mapping of that set into itself. We use a Gothic letter for the mapping as a warning that it may be a more complicated mathematical object than others we have encountered so far: if, for example, $\mathsf{X} = \mathcal{C}([-\eta, \eta], \mathbb{R}^d)$, then $\mathfrak{T}$ needs a vector-valued function $\mathbf{x}(t)$, $-\eta \le t \le \eta$, as its argument, and the result of applying $\mathfrak{T}$ to $\mathbf{x}$, which we write as $\mathfrak{T}[\mathbf{x}]$ with square brackets, is also a function on $[-\eta, \eta]$. We shall call $\mathfrak{T}$ a *contraction* if there is a constant $C < 1$ such that for all $\mathbf{x}, \mathbf{y} \in \mathsf{S}$,

$$\|\mathfrak{T}[\mathbf{x}] - \mathfrak{T}[\mathbf{y}]\| \le C\|\mathbf{x} - \mathbf{y}\|; \tag{3.19}$$

in words, for $\mathfrak{T}$ to be a contraction, it must be Lipschitz continuous with a Lipschitz constant less than unity. (Note that we are generalizing the notion of Lipschitz continuity from $\mathbb{R}^d$ to an infinite-dimensional space.) Finally, we shall call a point $\mathbf{x} \in \mathsf{S}$ a *fixed point* of $\mathfrak{T}$ if $\mathfrak{T}[\mathbf{x}] = \mathbf{x}$.

[9] Incidentally, the contraction mapping theorem extends easily to the more general context of a complete metric space. This result has no connection to the fact that X is a linear space. The only change needed is to replace $\|\mathbf{x} - \mathbf{y}\|$ by the distance function $d(\mathbf{x}, \mathbf{y})$. However, to minimize formalism, we don't introduce this more general, but unnecessary, context.

Theorem 3.2.7. *If S is a closed subset of a Banach space X and if $\mathfrak{T} : \mathsf{S} \to \mathsf{S}$ is a contraction, then $\mathfrak{T}$ has a unique fixed point in S.*

Proof. Choose a vector $\mathbf{x}_0 \in \mathsf{S}$ arbitrarily. Define a sequence inductively as follows: having chosen $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n$, let $\mathbf{x}_{n+1} = \mathfrak{T}[\mathbf{x}_n]$. Observe that

$$\|\mathbf{x}_{n+1} - \mathbf{x}_n\| = \|\mathfrak{T}[\mathbf{x}_n] - \mathfrak{T}[\mathbf{x}_{n-1}]\| \leq C \, \|\mathbf{x}_n - \mathbf{x}_{n-1}\|.$$

Iterating this inequality, we deduce that

$$\|\mathbf{x}_{n+1} - \mathbf{x}_n\| \leq C^n \|\mathbf{x}_1 - \mathbf{x}_0\|. \tag{3.20}$$

Since $C < 1$, (3.20) implies that $\{\mathbf{x}_n\}$ is Cauchy. *(Check this!)* Because X is complete, we conclude that $\{\mathbf{x}_n\}$ has a limit $\mathbf{x}_\infty$ in X; moreover, since S is closed, in fact $\mathbf{x}_\infty \in \mathsf{S}$.

We claim that $\mathbf{x}_\infty$ is a fixed point of $\mathfrak{T}$. To see this, observe that

$$\mathfrak{T}[\mathbf{x}_\infty] = \mathfrak{T}[\lim_{n\to\infty} \mathbf{x}_n] = \lim_{n\to\infty} \mathfrak{T}[\mathbf{x}_n] = \lim_{n\to\infty} \mathbf{x}_{n+1} = \mathbf{x}_\infty,$$

where we have used the continuity of $\mathfrak{T}$ to pull the limit outside the argument of $\mathfrak{T}$.

To show that the fixed point is unique, suppose $\mathbf{x}, \mathbf{y}$ are both fixed points of $\mathfrak{T}$. Then

$$\|\mathbf{x} - \mathbf{y}\| = \|\mathfrak{T}[\mathbf{x}] - \mathfrak{T}[\mathbf{y}]\| \leq C \, \|\mathbf{x} - \mathbf{y}\|,$$

or

$$(1 - C) \, \|\mathbf{x} - \mathbf{y}\| \leq 0.$$

Since $1 - C > 0$, we deduce $\|\mathbf{x} - \mathbf{y}\| \leq 0$. Then by Property (3.17a), we obtain $\mathbf{x} = \mathbf{y}$. □

3.2.5 Proof of the Existence Theorem

To prove Theorem 3.2.1 we will construct a solution of (3.7) by finding a fixed point of a mapping based on the integral equation (3.16), as follows: Choose an (open) ball $B(\mathbf{b}, \delta)$ around the initial condition $\mathbf{b}$ in $\mathbb{R}^d$ such that

$$\begin{aligned} &\text{(a) } \overline{B(\mathbf{b}, \delta)} \subset \mathcal{U} \quad \text{and} \\ &\text{(b) the restriction } \mathbf{F}|\,\overline{B(\mathbf{b}, \delta)} \text{ is Lipschitz continuous.} \end{aligned} \tag{3.21}$$

Using the same radius δ, let $\mathsf{S} \subset \mathcal{C}([-\eta,\eta],\mathbb{R}^d)$ be defined by[10]

$$\mathsf{S} = \{\mathbf{x} \in \mathcal{C}([-\eta,\eta],\mathbb{R}^d) : \forall t \in [-\eta,\eta] \;\; |\mathbf{x}(t) - \mathbf{b}| \le \delta\}. \tag{3.22}$$

For every $\mathbf{x} \in \mathsf{S}$, the composition $\mathbf{F}(\mathbf{x}(s))$ in the integrand of (3.16) makes sense and is a continuous function of s. Hence we may define a mapping from S into $\mathcal{C}([-\eta,\eta],\mathbb{R}^d)$, in symbols $\mathfrak{T} : \mathsf{S} \to \mathcal{C}([-\eta,\eta],\mathbb{R}^d)$, by the RHS of (3.16), i.e.,

$$\mathfrak{T}[\mathbf{x}](t) = \mathbf{b} + \int_0^t \mathbf{F}(\mathbf{x}(s))\,ds, \qquad -\eta \le t \le \eta. \tag{3.23}$$

Since we are dealing with infinite-dimensional spaces, let's review the notation, even though this may bore you. The argument of $\mathfrak{T}$ is a *function*, written $\mathbf{x}$ without any argument, and the result $\mathfrak{T}[\mathbf{x}]$ is also a function. To know what function $\mathfrak{T}[\mathbf{x}]$ is, we have to be told its value for every point $t \in [-\eta,\eta]$, and that is what (3.23) gives us.

The following two claims will allow us to apply Theorem 3.2.7 to extract a fixed point of $\mathfrak{T}$ in $\mathcal{C}([-\eta,\eta],\mathbb{R}^d)$. By Proposition 3.2.5, such a fixed point is the desired solution[11] of (3.7) on $(-\eta,\eta)$. Thus, the proof of Theorem 3.2.1 will be complete when we prove the claims.

Claim 1: *If η is sufficiently small, then for every $\mathbf{x} \in \mathsf{S}$, the image $\mathfrak{T}[\mathbf{x}]$ belongs to S.*

In other words, although as originally defined the range of $\mathfrak{T}$ was $\mathcal{C}([-\eta,\eta],\mathbb{R}^d)$, by reducing η if needed, we may regard $\mathfrak{T}$ as a mapping from S into S.

Proof. We need to show that for every $\mathbf{x} \in \mathsf{S}$,

$$\|\mathfrak{T}[\mathbf{x}] - \mathbf{b}\| \le \delta.$$

From (3.16) we compute that

$$(\mathfrak{T}[\mathbf{x}] - \mathbf{b})(t) = \int_0^t \mathbf{F}(\mathbf{x}(s))\,ds,$$

so

$$|\mathfrak{T}[\mathbf{x}] - \mathbf{b}|(t) \le \int_{I(t)} |\mathbf{F}(\mathbf{x}(s))|\,ds, \tag{3.24}$$

[10] In words, $B(\mathbf{b},\delta)$ is the ball in the Euclidean space $\mathbb{R}^d$ of radius δ around the vector $\mathbf{b}$, while S is the (closed) ball in the infinite-dimensional space $\mathcal{C}([-\eta,\eta],\mathbb{R}^d)$ of radius δ around the constant function $\mathbf{b}$. In symbols,

$$\mathsf{S} = \{\mathbf{x} \in \mathcal{C}([-\eta,\eta],\mathbb{R}^d) : \|\mathbf{x} - \mathbf{b}\| \le \delta\}.$$

[11] The fixed point $\mathbf{x}$ is a continuous function on the *closed* interval $[-\eta,\eta]$, which is more than (3.7) requires.

where $I(t)$ is the interval[12]

$$I(t) = [\min\{0, t\}, \max\{0, t\}].$$

Let

$$K = \max_{\mathbf{z} \in \overline{B(\mathbf{b},\delta)}} |\mathbf{F}(\mathbf{z})|.$$

Since $\mathbf{x} \in \mathsf{S}$, the integrand in (3.24) satisfies $|\mathbf{F}(\mathbf{x}(s))| \leq K$. Thus, observing that $\int_{I(t)} ds \leq \eta$, we conclude that $|\mathfrak{T}[\mathbf{x}] - \mathbf{b}|(t) \leq \eta K$, so the claim will be satisfied, provided η is chosen such that

$$\eta K \leq \delta. \tag{3.25}$$

□

Claim 2: *If η is sufficiently small, then $\mathfrak{T}$ is a contraction on S.*

Proof. By (3.21b), there is a Lipschitz constant L for $\mathbf{F}$ over $\overline{B(\mathbf{b}, \delta)}$. Let $\mathbf{x}, \mathbf{y} \in \mathsf{S}$ be given. From (3.23), we have

$$|\mathfrak{T}[\mathbf{x}] - \mathfrak{T}[\mathbf{y}]|(t) \leq \int_{I(t)} |\mathbf{F}(\mathbf{x}(s)) - \mathbf{F}(\mathbf{y}(s))| \, ds.$$

By the Lipschitz property,

$$|\mathfrak{T}[\mathbf{x}] - \mathfrak{T}[\mathbf{y}]|(t) \leq L \int_{I(t)} |\mathbf{x}(s) - \mathbf{y}(s)| \, ds.$$

Of course

$$|\mathbf{x}(s) - \mathbf{y}(s)| \leq \|\mathbf{x} - \mathbf{y}\|,$$

and estimating $\int_{I(t)} ds \leq \eta$, we deduce that

$$\|\mathfrak{T}[\mathbf{x}] - \mathfrak{T}[\mathbf{y}]\| \leq \eta L \|\mathbf{x} - \mathbf{y}\|.$$

The claim follows if

$$\eta L < 1. \tag{3.26}$$

□

Thus the proof of Theorem 3.2.1 is now complete. Since the above proof gives a fixed point for every η that satisfies (3.25) and (3.26), we may formulate a more quantitative version of the theorem.

Corollary 3.2.8. *Suppose the function $\mathbf{F}$ on the RHS of (3.7) is defined and Lipschitz continuous on a neighborhood of the closed ball $\overline{B(\mathbf{b}, \delta)}$. Let $K = \max\{|\mathbf{F}(\mathbf{x})| :$*

[12]With this notation we may write a single formula that is valid for both $t > 0$ *and* $t < 0$.

$\mathbf{x} \in \overline{B(\mathbf{b}, \delta)}\}$ *and let* L *be a Lipschitz constant for* $\mathbf{F}$ *on* $\overline{B(\mathbf{b}, \delta)}$. *Then the IVP* (3.7) *is solvable for* $t \in (-\eta, \eta)$, *provided*

$$\eta < \min\{\delta/K, 1/L\}. \tag{3.27}$$

In this result, the size of η is limited by both the magnitude of $\mathbf{F}$ and its Lipschitz constant. In point of fact, the restriction $\eta < 1/L$ *has no intrinsic significance*; it is required only for constructing the solution via a fixed-point argument. Indeed, the IVP has a solution for $-\eta < t < \eta$, provided merely that $\eta < \delta/K$. This range of t has a natural interpretation: if a particle starts at the center of a ball of radius δ and moves with speed at most K, then a time of at least δ/K is required for the particle to reach the boundary. You may use this interpretation to prove the result now, or you can wait until Exercise 4.16, where we ask you to derive it using techniques developed in that chapter.

3.2.6 An Illustrative Example and Picard Iteration

Unscrambling the proof of the fixed-point theorem, we see that the construction of the solution of the IVP ultimately comes down to the limit of an iterated sequence: $\mathbf{x}_0$ is chosen arbitrarily (e.g., $\mathbf{x}_0(t) \equiv \mathbf{b}$), and subsequent $\mathbf{x}$'s are chosen iteratively:

$$\mathbf{x}_{n+1} = \mathfrak{T}[\mathbf{x}_n]. \tag{3.28}$$

Let's compute the iterates for the simplest of IVPs, the scalar problem

$$x' = x, \qquad x(0) = 1.$$

Equation (3.28) becomes

$$x_{n+1} = 1 + \int_0^t x_n(s)\, ds.$$

If we choose $x_0(t) \equiv 1$, then we obtain

$$x_n(t) = 1 + t + \frac{1}{2!}t^2 + \cdots + \frac{1}{n!}t^n.$$

In other words, the nth iterate is just the polynomial approximation of degree n to the exponential e^t. Thus, the iteration works very well indeed for this simple example.

Incidentally, some authors prove Theorem 3.2.1 directly by iteration of the integral equation (3.16), which avoids the fixed-point theorem. This is called *Picard iteration*. The two approaches are compared in the Pearls. If you'd like to practice Picard iteration on your own, try Exercise 6.

3.2.7 Concluding Remarks

Strictly speaking, the definition of a solution of an IVP includes a domain, $\alpha < t < \beta$. For the time being, functions defined on different intervals must be regarded as different solutions. Ultimately, these pedantic distinctions can be removed—we will show this in several stages culminating in Proposition 4.1.1 in Chapter 4—but based on what we know so far they are required for logical consistency.

The following two results are mildly interesting in their own right, but more important, they are useful as tools in certain proofs below. In Exercise 1(c) we give you hints for proving both results.

Lemma 3.2.9. *Suppose $\mathbf{x}_1, \mathbf{x}_2$ are solutions of an ODE such that*

(i) $\mathbf{x}_1$ *is continuous on* $(\alpha, \beta]$ *and satisfies* $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ *on* (α, β),
(ii) $\mathbf{x}_2$ *is continuous on* $[\beta, \gamma)$ *and satisfies* $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ *on* (β, γ).

If $\mathbf{x}_1(\beta) = \mathbf{x}_2(\beta)$ and if this common value belongs to the domain[13] *of $\mathbf{F}$ (on which this function is continuous), then the definition*

$$\mathbf{x}(t) = \begin{cases} \mathbf{x}_1(t) & \text{if } \alpha < t \le \beta, \\ \mathbf{x}_2(t) & \text{if } \beta \le t < \gamma, \end{cases} \tag{3.29}$$

yields a solution of the ODE on the combined interval (α, γ).

Corollary 3.2.10. *If $\mathbf{x}$ is a $\mathcal{C}^1$ solution of an ODE in an interval (α, β) that is continuous on $(\alpha, \beta]$ and if $\mathbf{x}(\beta)$ belongs to the domain of $\mathbf{F}$, then $\mathbf{x}$ may be extended to a solution of the equation in a slightly larger interval $(\alpha, \beta + \varepsilon)$, and similarly for the left endpoint.*

3.3 Uniqueness Theory

3.3.1 Gronwall's Lemma

The main workhorse of the uniqueness proof, Gronwall's lemma, is a simple inequality, but it provides extremely useful estimates for solutions of an ODE in many contexts besides the present one.

[13] Here is an example to show that the annoying hypothesis "$\mathbf{x}_j(\beta)$ belongs to the domain of $\mathbf{F}$" is really necessary. It's a bit technical, so don't pursue this unless you, like us, are amused by such things. On the domain $\mathcal{U} = \{(x, y) : x^2 > y^3\}$ consider the ODE $(x', y') = (f(x, y), 0)$, where $f(x, y) = (1/3)(x^2 - y^3)^{-1}$. Then $(x_1(t), y_1(t)) = (t^{1/3}, 0)$ for $t \le 0$ and $(x_2(t), y_2(t)) = (t^{1/3}, 0)$ for $t \ge 0$ satisfy all the hypotheses of the theorem (with $\beta = 0$) except the containment hypothesis, but the function defined by (3.29) is not differentiable at $t = 0$.

Lemma 3.3.1. *Let $g : [0, T] \to \mathbb{R}$ be continuous, and suppose there are nonnegative constants C, K such that*

$$g(t) \le C + K \int_0^t g(s)\, ds, \qquad 0 \le t \le T. \tag{3.30}$$

Then

$$g(t) \le Ce^{Kt}, \qquad 0 \le t \le T. \tag{3.31}$$

Here is a corollary of Gronwall's lemma whose hypotheses are more intuitive: if g is differentiable and satisfies

$$g' \le Kg, \qquad g(0) \le C, \tag{3.32}$$

then g is bounded by an exponential as in (3.31). This could easily be proved directly; alternatively, it follows from the lemma, because integration of condition (3.32) yields (3.30). In contrast to (3.32), Gronwall's inequality does *not* require that g be differentiable, and this makes application of the lemma much more flexible.

Proof of Lemma 3.3.1. We define a function by the RHS of (3.30),

$$G(t) = C + K \int_0^t g(s)\, ds.$$

The function G is $\mathcal{C}^1$, and it satisfies

$$\text{(a) } g(t) \le G(t) \qquad \text{and} \qquad \text{(b) } G'(t) = Kg(t). \tag{3.33}$$

Applying Leibniz's rule to differentiate the product $e^{-Kt}G(t)$ and invoking (3.33), we calculate

$$\frac{d}{dt}\left[e^{-Kt}G(t)\right] = e^{-Kt} \cdot Kg(t) - Ke^{-Kt} \cdot G(t) = Ke^{-Kt}\left[g(t) - G(t)\right] \le 0.$$

Thus, $e^{-Kt}G(t)$ is nonincreasing, so $e^{-Kt}G(t) \le G(0) = C$. Invoking (3.33a) again, we deduce that

$$g(t) \le G(t) \le Ce^{Kt}.$$

□

Remark: One minor generalization of Gronwall's lemma relaxes the hypotheses and assumes that g is merely *piecewise* continuous;[14] check that the same conclusion still follows. Exercise 8 gives two other generalizations, but there are many more.

[14]For the record: a function g is called *piecewise continuous* on an interval $\mathcal{I}$ if there is a finite set of points $\{a_k : k = 1, 2, \ldots, p\}$ in $\mathcal{I}$ such that (i) g is continuous on $\mathcal{I} \sim \cup_k \{a_k\}$ and (ii) at each point a_k, the one-sided limits of g exist (and are finite).

3.3.2 More on Lipschitz Functions

Proposition 3.3.2. *Suppose $\mathcal{U} \subset \mathbb{R}^{d_1}$ is open and $\mathbf{F} : \mathcal{U} \to \mathbb{R}^{d_2}$ is locally Lipschitz. Then for every compact set $\mathcal{K} \subset \mathcal{U}$, the restriction $\mathbf{F}|\mathcal{K}$ is (globally) Lipschitz.*

In particular, if $\mathbf{F}$ is $\mathcal{C}^1$ and $\mathcal{K}$ is compact, then $\mathbf{F}|\mathcal{K}$ is (globally) Lipschitz. If $\mathcal{K}$ is *convex*, Proposition 3.2.2 gives the same conclusion, along with the simple estimate (3.15) for the Lipschitz constant of $\mathbf{F}|\mathcal{K}$. When $\mathcal{K}$ is not convex, the following intricate nonconstructive compactness argument is needed for the proof. (See Exercise 12 for an illustration of the complicating issues.)

Proof. For each $\mathbf{x} \in \mathcal{K}$, choose a ball $B(\mathbf{x}, \delta_\mathbf{x})$ such that (i) its closure $\overline{B(\mathbf{x}, \delta_\mathbf{x})}$ is contained in $\mathcal{U}$ and (ii) $\mathbf{F}$ is Lipschitz continuous on $\overline{B(\mathbf{x}, \delta_\mathbf{x})}$. The collection

$$\{B(\mathbf{x}, \delta_\mathbf{x}/2) : \mathbf{x} \in \mathcal{K}\}$$

is an open cover of $\mathcal{K}$. (Note that the radii here have been halved.) Choose a finite subcover of $\mathcal{K}$, say $B(\mathbf{x}_j, \delta_j/2)$, $j = 1, 2, \ldots, J$; let Λ_j be a Lipschitz constant for $\mathbf{F}$ on $B(\mathbf{x}_j, \delta_j)$ (radius *not* halved); and let

$$L_1 = \max_{j=1,\ldots,J} \Lambda_j, \qquad L_2 = 4 \max_{\mathbf{x}\in\mathcal{K}} |\mathbf{F}(\mathbf{x})| / \min_j \delta_j.$$

Let us show that $\mathbf{F}$ is Lipschitz over $\mathcal{K}$ with Lipschitz constant $L = \max\{L_1, L_2\}$. To prove this, suppose $\mathbf{x}, \mathbf{y} \in \mathcal{K}$. The first point, $\mathbf{x}$, belongs to one of the balls in the finite subcover, say $\mathbf{x} \in B(\mathbf{x}_k, \delta_k/2)$. We consider two cases: (i) If $\mathbf{y}$ belongs to the (full-radius) ball $B(\mathbf{x}_k, \delta_k)$ with the same index, then by construction,

$$|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})| \le \Lambda_k |\mathbf{x} - \mathbf{y}| \le L_1 |\mathbf{x} - \mathbf{y}|.$$

(ii) If $\mathbf{y}$ lies outside $B(\mathbf{x}_k, \delta_k)$, then it may be seen from Figure 3.1 that

$$|\mathbf{x} - \mathbf{y}| \ge \delta_k/2. \tag{3.34}$$

(*It is useful practice to derive this result analytically.*) Of course

$$|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})| \le 2 \max_{\mathcal{K}} |\mathbf{F}|; \tag{3.35}$$

we multiply the RHS of (3.35) by $2|\mathbf{x} - \mathbf{y}|/\delta_k$, a quantity that by (3.34) is greater than unity, to obtain

$$|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})| \le 4 \max_{\mathcal{K}} |\mathbf{F}| \, \frac{|\mathbf{x} - \mathbf{y}|}{\delta_k} \le L_2 |\mathbf{x} - \mathbf{y}|.$$

Thus both cases are covered, and the proof is complete. □

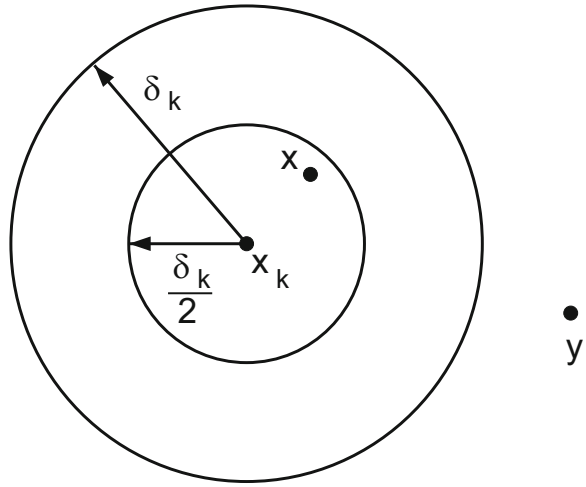

Figure 3.1: *Graphical verification of the inequality* (3.34).

For future use let us record a somewhat dull corollary of the compactness construction in the above argument that will be needed below.

Corollary 3.3.3. *If $\mathcal{K} \subset \mathcal{U} \subset \mathbb{R}^d$, where $\mathcal{K}$ is compact and $\mathcal{U}$ is open, then there exist a larger compact set $\mathcal{K}' \subset \mathcal{U}$ and a $\delta > 0$ such that for every $\mathbf{x} \in \mathcal{K}$, the closed ball $\overline{B(\mathbf{x}, \delta)}$ is contained in $\mathcal{K}'$.*

Proof. In the notation of the previous proof, let $\mathcal{K}' = \cup_{j=1}^{J} \overline{B(\mathbf{x}_j, \delta_j)}$ and let $\delta = \min_j(\delta_j/2)$. □

3.3.3 The Uniqueness Theorem

Theorem 3.3.4. *Suppose that $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$ is locally Lipschitz. Let $\mathbf{x}_1, \mathbf{x}_2$ be two solutions of the initial value problem*

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0) = \mathbf{b}, \tag{3.36}$$

say defined for $\alpha_j < t < \beta_j, \quad j = 1, 2$. Then for all t in the range $\max\{\alpha_1, \alpha_2\} < t < \min\{\beta_1, \beta_2\}$ where both solutions are defined, $\mathbf{x}_1(t) = \mathbf{x}_2(t)$.

Remarks: (i) Although "local theory" is part of the title of this chapter and the existence theorem was very local indeed, this uniqueness theorem is in fact global: uniqueness holds on every interval over which the IVP happens to have a solution, no matter how long. (ii) In light of Lemma 3.2.9, we may paste together the two solutions considered in the theorem to obtain a solution on the combined interval $\min\{\alpha_1, \alpha_2\} < t < \max\{\beta_1, \beta_2\}$.

Proof. We consider only positive time; with trivial modifications, the arguments can be adapted to handle $t < 0$. We want to apply Gronwall's lemma to the function[15] $g(t) = |\mathbf{x}_1(t) - \mathbf{x}_2(t)|$. Both solutions satisfy integral equations as in (3.16). Subtracting these, we have

$$\mathbf{x}_1(t) - \mathbf{x}_2(t) = \int_0^t [\mathbf{F}(\mathbf{x}_1(s)) - \mathbf{F}(\mathbf{x}_2(s))]\, ds, \qquad 0 \le t < \min\{\beta_1, \beta_2\}. \tag{3.37}$$

Thus

$$g(t) = |\mathbf{x}_1(t) - \mathbf{x}_2(t)| \le \int_0^t |\mathbf{F}(\mathbf{x}_1(s)) - \mathbf{F}(\mathbf{x}_2(s))|\, ds. \tag{3.38}$$

Temporarily we restrict t to an interval $[0, T]$ where $T < \min\{\beta_1, \beta_2\}$. Since $[0, T]$ is compact, so is the union of images $\mathcal{K} = \mathbf{x}_1([0, T]) \cup \mathbf{x}_2([0, T])$. Therefore, by Proposition 3.3.2, there is a Lipschitz constant L for $\mathbf{F}$ on $\mathcal{K}$. Hence for $0 \le s \le T$, the integrand on the RHS of (3.38) may be estimated as follows:

$$|\mathbf{F}(\mathbf{x}_1(s)) - \mathbf{F}(\mathbf{x}_2(s))| \le L|\mathbf{x}_1(s) - \mathbf{x}_2(s)| = Lg(s).$$

Substituting into (3.38), we see that

$$g(t) \le L \int_0^t g(s)\, ds. \tag{3.39}$$

Thus from Gronwall's inequality with $C = 0$ we deduce that $g(t) \le 0$ for $0 \le t \le T$. But g is nonnegative, so $g \equiv 0$, and thus $\mathbf{x}_1(t) = \mathbf{x}_2(t)$ for this range of t. Finally, we may take T arbitrarily close to $\min\{\beta_1, \beta_2\}$, so we have equality for all t where both solutions are defined. □

Remark: We pause to shoot down a possible misconception. Students sometimes imagine that two solutions of an autonomous equation $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ may cross one another, but this is false. To put a positive slant on this negative remark: If at times t_1 and t_2 two solutions $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ satisfy $\mathbf{x}_1(t_1) = \mathbf{x}_2(t_2)$, then $\mathbf{x}_2(t) = \mathbf{x}_1(t - (t_2 - t_1))$ for all t for which both sides of this equality are defined. *(Prove this!)*

---

[15] Note that the absolute-value function is not differentiable, so it is possible that g is not differentiable. Thus, the weak hypothesis in Gronwall's lemma simplifies the proof of the theorem.

This is just one instance of how proofs in ODEs, which is an old subject, have been polished over the years. Indeed, beware of reading through this and other proofs too quickly and missing the cleverness. For example, in this proof, before applying Gronwall's lemma, we prepare for it by (i) restricting t to a large closed subinterval of $[0, \beta)$ to obtain compactness and (ii) invoking Proposition 3.3.2 to derive a Lipschitz constant that works on all of $[0, T]$. *After these preparations have been made*, the proof may be appropriately described as a straightforward application of Gronwall's lemma.

It will be convenient below to have uniqueness under a (superficially) weaker hypothesis. Thus, we define a *solution of* (3.36) *in forward time*, by which we mean a continuous function $\mathbf{x} : [0, \beta) \to \mathcal{U}$ with $\mathbf{x}(0) = \mathbf{b}$ that is continuously differentiable on the open interval $(0, \beta)$ and satisfies the ODE there. Such solutions are unique:

Corollary 3.3.5. *Suppose that* $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$ *in* (3.36) *is locally Lipschitz. Let* $\mathbf{x}_1, \mathbf{x}_2$ *be two solutions in forward time of the initial value problem* (3.36)*, say defined for* $0 \le t < \beta_j, \quad j = 1, 2$. *Then for all* t *in the range* $0 \le t < \min\{\beta_1, \beta_2\}$ *where both solutions are defined,* $\mathbf{x}_1(t) = \mathbf{x}_2(t)$.

In fact, it follows from Corollary 3.2.10 that each solution $\mathbf{x}_j$ has an extension to a solution on an open interval $(-\varepsilon, \beta_j)$ that contains the origin, and Corollary 3.3.5 may be proved by applying Theorem 3.3.4 to the extensions. Thus, *the generalization to an apparently wider class of solutions is actually vacuous.* We introduce it here only because in some proofs below it is technically convenient to be able to apply the uniqueness result to solutions of the IVP that are defined only for $t \ge 0$. In this way, we may avoid the distraction of constructing extensions in the middle of another proof.

3.4 Generalization to Nonautonomous Systems

3.4.1 Nonlinear Systems

Both the existence and uniqueness theorems generalize to nonautonomous IVPs, say

$$\mathbf{x}' = \mathbf{G}(\mathbf{x}, t), \qquad \mathbf{x}(0) = \mathbf{b}, \tag{3.40}$$

as we discuss in Exercise 17.

Note that to simplify the notation in (3.40), we have imposed the initial condition at $t = 0$. Unlike those for autonomous equations, solutions of nonautonomous equations do not have translational invariance. Thus strictly speaking, imposing an initial condition at a different time, say $\mathbf{x}(t_0) = \mathbf{b}$, gives a different problem. However, no real generality is lost by assuming $t_0 = 0$ in (3.40), since the general case can easily be reduced to (3.40) by appropriate translation.

3.4.2 Linear Systems

The theory for nonautonomous *linear* systems, say an IVP

$$\mathbf{x}' = A(t)\mathbf{x} + \mathbf{g}(t), \quad \mathbf{x}(0) = \mathbf{b}, \tag{3.41}$$

is refreshingly simple and deserves explicit formulation.

Theorem 3.4.1. *Suppose the coefficient matrix $A(t)$ and the inhomogeneous term $\mathbf{g}(t)$ in (3.41) are continuous in the (possibly infinite) open interval (T_1, T_2) that contains $t = 0$; then this IVP has a unique solution that exists for $T_1 < t < T_2$.*

In particular, solutions of a linear system do not blow up in finite time, no matter how rapidly $\|A(t)\|$ or $|\mathbf{g}(t)|$ may grow with t. (If this result makes you think linear equations are dull, see Exercise 13.)

Two options for proving this result are available. One method uses Picard iteration; Exercise 16 guides you through this. The other, for which we are not yet ready, combines local results with theory from Chapter 4; see Exercise 5(a) in that chapter.

3.5 Exercises

After the core exercises, there is a subsection of exercises on linear ODEs with (time-dependent) periodic coefficients. This phrase describes a whole subfield known as Floquet theory.

3.5.1 Core Exercises

The core exercises have the following purposes:

To deal with unfinished business	1, 12
To develop intuition about blowup through an example	2
To understand Lipschitz continuity through examples	3
To understand integral operators	4, 5
To explore the approach of Picard iteration	6
To extend theory developed in the text	7, 8
To increase your facility with ODEs	9, 10
To connect theory and computing	11

1. (a) Prove Proposition 3.2.5.

   (b) Show that the definition (3.18) satisfies the axioms (3.17).

   (c) Prove Lemma 3.2.9 and Corollary 3.2.10.

   Hint for Lemma 3.2.9: By hypothesis, $\mathbf{x}_1' = \mathbf{F}(\mathbf{x}_1(t))$ for $t \in (\alpha, \beta)$. But $\mathbf{F} \circ \mathbf{x}_1$, a composition of continuous functions, is continuous on the half-closed interval $(\alpha, \beta]$. Thus,

   $$\lim_{t \to \beta^-} \mathbf{x}_1'(t) = \mathbf{F}(\mathbf{x}(\beta)).$$

   Argue similarly for $\mathbf{x}_2(t)$ and combine your results to conclude that the function (3.29) is $\mathcal{C}^1$ on the combined interval.

   Hint for Corollary 3.2.10: Apply Theorem 3.2.1 to solve an IVP for $\mathbf{y}' = \mathbf{F}(\mathbf{y})$ with initial condition $\mathbf{y}(\beta) = \mathbf{x}(\beta)$ on $\beta - \eta < t < \beta + \eta$. Then use Lemma 3.2.9 to obtain a solution on $(\alpha, \beta + \eta)$.

2. (a) Use separability to solve the scalar IVP

$$x' = |x|^p, \qquad x(0) = b,$$

where $p > 0$ and $b > 0$.

Discussion: You will see that the solution blows up in finite time if $p > 1$, i.e., if the RHS of the ODE has superlinear growth. According to Theorem 4.2.1 of Chapter 4, solutions of an equation with linear growth cannot blow up in finite time. However, even if an equation has superlinear growth, this does not imply that solutions blow up in finite time. One counterexample is $x' = |x|^p \sin x$. (*Why don't solutions blow up?*) Part (b) of this exercise gives another counterexample, which is two-dimensional.

(b) By finding explicit solutions, show that every IVP for the system (in polar coordinates)

$$\begin{aligned} r' &= r \\ \theta' &= r^2 \end{aligned}$$

has a solution for all time (positive and negative).

Advice: For practice, be sure you can write this system in Cartesian coordinates. Incidentally, you will find that in Cartesian coordinates, the RHS of the equation grows cubically.

3. *Question:* Are the following functions Lipschitz continuous on the indicated sets? (Incidentally, all of these examples are scalar-valued functions; vector-valued functions would not pose any additional difficulties except for the need to examine more components.)

(a) $(x+2)^{1/3}$ on the interval $[-1, 1]$? On $[-\infty, -1]$?

(b) $(x^2+2)^{1/3}$ on the interval $[-1, 1]$? On $[-\infty, -1]$?

(c) $(2x^3+1)^{1/3}$ on the interval $[-1, 1]$? On $[-\infty, -1]$?

(d) $(2x^4+1)^{1/3}$ on the interval $[-1, 1]$? On $[-\infty, -1]$?

(e) $\sin(e^{-x})$ on the interval $[-1, 1]$? On $[-\infty, -1]$?

(f) $x \sin(1/x)$ on the interval $[-1, 1]$? On $[-\infty, -1]$?

(g) $|x + y^2 - 2|$ on the square $\{|x| \le 2, |y| \le 2\}$? On $\mathbb{R}^2$?

Hint: Observe that the function is the composition of $x + y^2 - 2$ with the absolute value function. Show that the composition of two Lipschitz functions is Lipschitz continuous. Thus, $|x+y^2-2|$ is Lipschitz on the indicated set if $x + y^2 - 2$ is.

(h) $\sqrt{x^2+1}/(x^2+y^2-1)$ on the bounded annulus $\{2 \le x^2 + y^2 \le 8\}$? On the unbounded annulus $\{2 \le x^2 + y^2 < \infty\}$?

Hint: The function is certainly not Lipschitz on all of $\mathbb{R}^2$, because the denominator goes to zero on the unit circle. Propositions 3.2.2 and 3.3.2 may be combined to show that this function is Lipschitz on the bounded annulus. More thought is required to analyze the unbounded annulus. Have fun!

4. Consider the integral operator (on scalar functions)

$$\mathfrak{T}[x](t) = e^t\, b + \int_0^t \cos(s+t)\, x^2(s)\, ds,$$

where b is a constant. Find η, δ such that $\mathfrak{T}$ is a contraction on the set

$$\mathsf{S} = \{x \in \mathcal{C}([-\eta, \eta], \mathbb{R}) : \|x - b\| \le \delta\}.$$

5. (a) Formulate the integral equation (3.16) for the IVP

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \sqrt{100 - x^2 - y^2} \\ (\cos x)\, e^{y/5} \end{bmatrix}, \qquad \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} = \begin{bmatrix} 7 \\ 7 \end{bmatrix}.$$

   (b) Find (numerical values for) η, δ such that on the set (3.22), this operator defines a contraction of S into itself.

6. (a) Find the nth Picard iterate resulting from the IVP

$$x' = -tx, \qquad x(0) = 2,$$

   assuming $x_0(t) \equiv 2$.

   (b) Solve this IVP analytically and compare your answer in Part (a) with the power series expansion of the solution.

7. Consider an IVP

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0) = \mathbf{b}$$

such that the first component of the RHS satisfies $F_1(0, \tilde{\mathbf{x}}) = 0$ for all $\tilde{\mathbf{x}} \in \mathbb{R}^{d-1}$ with $(0, \tilde{\mathbf{x}}) \in \mathcal{U}$, where $\tilde{\mathbf{x}}$ is shorthand for $(x_2, \ldots, x_d)$. Prove that if $b_1 = 0$, then $x_1(t) \equiv 0$ for as long as the solution exists.

Hint: Prove this result, which is a kind of uniqueness theorem for the first component of the solution of an IVP, with an argument similar to the proof of the uniqueness theorem.

8. *Introduction:* In this exercise you derive two generalizations of Gronwall's inequality, Lemma 3.3.1, which will be used in later chapters.

(a) Show that if $g : [0, T] \to \mathbb{R}$ is continuous and if there are nonnegative constants C, B, K such that

$$g(t) \le C + Bt + K \int_0^t g(s)\, ds, \qquad 0 \le t \le T,$$

then

$$g(t) \le Ce^{Kt} + B\frac{e^{Kt} - 1}{K}, \qquad 0 \le t \le T.$$

Hint: One approach is simply to mimic the proof of Lemma 3.3.1. A more elegant alternative, which does not require reexamining the proof of Lemma 3.3.1, involves applying Gronwall's inequality as given in Section 3.3.1 to the function $h(t) = g(t) + B/K$.

(b) Show that if $g : [0, T] \to \mathbb{R}$ is continuous and if there are nonnegative constants C, M, K with $M < K$ such that

$$g(t) \le C(e^{Mt} - 1) + K \int_0^t g(s)\, ds, \qquad 0 \le t \le T,$$

then

$$g(t) \le \frac{C}{K/M - 1}(e^{Kt} - e^{Mt}), \qquad 0 \le t \le T. \tag{3.42}$$

Hint: As in Part (a), this inequality can be reduced to Lemma 3.3.1 by the addition of a carefully chosen term to $g(t)$. Alternatively, you can mimic the original proof of Lemma 3.3.1.

9. (a) Let $\mathbf{x}_1(t), \ldots, \mathbf{x}_k(t)$ be solutions of a homogeneous d-dimensional linear system

$$\mathbf{x}' = A(t)\mathbf{x}, \tag{3.43}$$

where $A(t)$ is continuous, and assume that at time zero, $\mathbf{x}_1(0), \ldots, \mathbf{x}_k(0)$ are linearly independent vectors in $\mathbb{R}^d$. Show that at later times, $\mathbf{x}_1(t), \ldots, \mathbf{x}_k(t)$ are still linearly independent.

Hint: This is an easy exercise. Just consider appropriate linear combinations of $\mathbf{x}_1(t), \ldots, \mathbf{x}_k(t)$ and invoke the uniqueness theorem.

(b) Let $\mathbf{x}_1(t), \ldots, \mathbf{x}_d(t)$ be solutions of (3.43), let $W(t)$ be the square matrix $W(t) = \mathrm{Col}(\mathbf{x}_1(t), \ldots, \mathbf{x}_d(t))$, and let $\phi(t) = \det W(t)$. Give a heuristic proof[16] that

$$\phi'(t) = [\mathrm{tr}A(t)]\, \phi.$$

[16] In Chapter 4, when we introduce the order notation, we ask you to use this notation to make the argument rigorous. Incidentally, the letter "W" is a mnemonic for "Wronskian" (cf. Section 2.1 of [10]).

Hint: Of course

$$\phi'(t) = \lim_{\varepsilon \to 0} \frac{\det W(t+\varepsilon) - \det W(t)}{\varepsilon}. \tag{3.44}$$

Now

$$W(t+\varepsilon) \approx W(t) + \varepsilon W'(t) = W(t)\,[I + \varepsilon W^{-1}(t) W'(t)].$$

Regarding the RHS of the last equation, recall that the determinant of a product equals the product of the determinants. Then, with appropriate hand-waving, argue that $\det[I + \varepsilon W^{-1}W'] \approx 1 + \varepsilon \operatorname{tr}(W^{-1}W')$ for small ε. Justify the following equalities:

$$\operatorname{tr} W^{-1}W' = \operatorname{tr} W^{-1}(AW) = \operatorname{tr} (AW)W^{-1} = \operatorname{tr} A(WW^{-1}) = \operatorname{tr} A.$$

Substitute this information into (3.44) and manipulate the result.

10. *Introduction:* Let $\mathbf{x}(t)$ be a continuous function on the closed interval $[0, T]$ that satisfies the ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ for $0 < t < T$ and assume moreover that $\mathbf{x}(T) = \mathbf{x}(0)$ belongs to the domain of $\mathbf{F}$. Extend $\mathbf{x}$ to a periodic function $\tilde{\mathbf{x}}(t)$, defined for all time, such that

$$\tilde{\mathbf{x}}(t+T) = \tilde{\mathbf{x}}(t). \tag{3.45}$$

Show that $\tilde{\mathbf{x}}$ satisfies the equation $\tilde{\mathbf{x}}' = \mathbf{F}(\tilde{\mathbf{x}})$ for all t.

Remark: This easy exercise gives you a mini head start in thinking about periodic solutions of ODEs, which are the focus of Chapter 7.

11. Solve numerically the IVP

$$x' = \sqrt{|x|}, \qquad x(0) = -1.$$

Discussion: Adapting our solutions from Example 2 in Section 3.1, we can see that this IVP has infinitely many solutions: once $x(t)$ reaches zero, it may sit at this level for an arbitrarily long time, say until $t = T$, before switching to the parabola $x(t) = (t-T)^2/4$. However, you will find that the numerical solution completely misses these subtleties. *Moral:* It is risky to compute solutions without an adequate theoretical understanding of existence and uniqueness issues for the problem you are studying.

12. On the cut plane

$$\mathcal{U} = \{(r, \theta) : r > 0,\ -\pi < \theta < \pi\},$$

the function $f(r, \theta) = r \sin(\theta/2)$ is continuous and locally Lipschitz. *(Are you clear why this is true with the cut but false without it?)* If $\varepsilon > 0$, define the compact subset of $\mathcal{U}$

$$\mathcal{K} = \{(r, \theta) : 1/2 \le r \le 3/2,\ -\pi + \varepsilon \le \theta \le \pi - \varepsilon\},$$

and consider the two points $\mathbf{P}_\pm$ in $\mathcal{K}$ with $r = 1$, $\theta = \pm(\pi - \varepsilon)$. Argue that

$$\frac{|f(\mathbf{P}_+) - f(\mathbf{P}_-)|}{|\mathbf{P}_+ - \mathbf{P}_-|} \approx \frac{1}{\varepsilon}.$$

Discussion: In other words, although f is Lipschitz on $\mathcal{K}$, the Lipschitz constant of f goes through the roof if ε is small. This example explains why the proof of Theorem 3.3.2 was such a chore.

3.5.2 Linear ODEs with Periodic Coefficients

The study of linear ODEs with periodic coefficients is known as *Floquet theory*. We briefly develop the theory in Section 7.10.1. In the following exercises we illustrate the behavior of solutions of such equations. Definitely do Exercise 13; it is both interesting and instructive. The other two exercises are optional. Although they are a little technical, in them you can derive the behavior of solutions analytically, without recourse to the computer.

13. Write Mathieu's equation (1.5b)

$$x'' + (\kappa + 2\varepsilon \cos t)x = 0 \tag{3.46}$$

as a first-order system; assuming $\kappa = 1/4$, solve the IVP numerically for $0 \le t \le 1000$, say with $\varepsilon = 0.05$. (Exact initial conditions don't matter much, provided $x(0)$ and $x'(0)$ don't both vanish.)

Discussion: It is easy to imagine that the effect of the small oscillatory coefficient $2\varepsilon \cos t$ ought to average to zero, so that solutions of (3.46) ought to roughly imitate the periodic behavior that results from setting $\varepsilon = 0$ in the equation. This can happen, but not for the parameters proposed here. Equation (3.46) exhibits a form of *resonance*, which is discussed in the Pearls (cf. Figure 3.2 below).

If you are curious, you may experiment by adding friction (include an additional term, $\beta x'$) and "detuning" the equation (change κ slightly away from $1/4$). You will find that resonant growth persists under sufficiently small perturbations of either type. We will derive this behavior analytically in Exercise 7.16.

Incidentally, (3.46) also exhibits resonant growth near $\kappa = 0$ and $\kappa = 1$, although in the former case, ε must exceed a threshold, and in the latter case, the phenomenon is less robust.

It is interesting to compare this behavior of (3.46) with a pendulum vibrated at its base, Exercise 1.12. Two minor differences are that the pendulum equation was nonlinear and friction was included. Far more significantly, however, the basic behavior in the two equations is reversed—in (3.46), the time-periodic perturbation was small and caused exponential growth, while in (1.53), it was large and its effect was stabilizing.

14. *Introduction:* Meissner's equation, a more tractable analogue of Mathieu's equation, is

$$x'' + [\kappa + \varepsilon S(t)]x = 0, \tag{3.47}$$

where $S(t)$ is the 2π-periodic square wave[17] such that

$$S(t) = \begin{cases} 1 & \text{for } 0 < t < \pi \\ -1 & \text{for } \pi < t < 2\pi. \end{cases}$$

Show that if $\kappa = 0$, then solutions of (3.47) are bounded if

$$|\cosh(\pi\sqrt{\varepsilon})\,\cos(\pi\sqrt{\varepsilon})| < 1 \tag{3.48}$$

and grow exponentially if this quantity exceeds 1.

Hint: Rewrite (3.47), assuming $\kappa = 0$, as a first-order system. On intervals of length π, this system may be viewed as a constant-coefficient linear system:

$$\begin{array}{ll} \mathbf{x}' = A_+\mathbf{x}, & \text{for } 2n\pi < t < (2n+1)\pi \\ \mathbf{x}' = A_-\mathbf{x}, & \text{for } (2n+1)\pi < t < (2n+2)\pi \end{array} \tag{3.49}$$

where n is an arbitrary integer and

$$A_\pm = \begin{bmatrix} 0 & 1 \\ \mp\varepsilon & 0 \end{bmatrix}.$$

Argue that for every positive integer n,

$$\mathbf{x}(2n\pi) = \Pi^n\,\mathbf{x}(0), \qquad \text{where } \Pi = (e^{\pi A_-})\cdot(e^{\pi A_+}). \tag{3.50}$$

Solutions will be bounded if the eigenvalues of Π satisfy $|\lambda(\Pi)| = 1$ and are distinct, while (most) solutions will grow if one of the eigenvalues satisfies $|\lambda(\Pi)| > 1$.

To determine the eigenvalues of Π, calculate e^{tA_+} (cf. Exercise 2.5(e)) and e^{tA_-}. Then show that $\det\Pi = 1$ and $\operatorname{tr}\Pi = 2\cosh(\pi\sqrt{\varepsilon})\,\cos(\pi\sqrt{\varepsilon})$. Obtain the claimed result from examining the quadratic formula for the eigenvalues of Π.

Incidentally, near $\varepsilon = 0$, (3.48) is satisfied if $0 < \varepsilon < 0.356$. For large ε, (3.48) is violated except for small intervals around the zeros of $\cos(\pi\sqrt{\varepsilon})$.

15. Show that if $\kappa = 1/4$, solutions of (3.47) grow exponentially if $\varepsilon \neq 0$ is small.

Hint: On intervals of length π, (3.47) may written in the form (3.49) with

$$A_\pm = \begin{bmatrix} 0 & 1 \\ -\kappa_\pm^2 & 0 \end{bmatrix}$$

where $\kappa_\pm = \sqrt{1/4 \pm \varepsilon}$. As above, the behavior of solutions depends on the eigenvalues of $\Pi = (e^{\pi A_-})\cdot(e^{\pi A_+})$. Show that $\det\Pi = 1$ and

$$\operatorname{tr}\Pi = 2\cos(\kappa_+\pi)\,\cos(\kappa_-\pi) - \left(\frac{\kappa_+}{\kappa_-} + \frac{\kappa_-}{\kappa_+}\right)\sin(\kappa_+\pi)\,\sin(\kappa_-\pi).$$

[17] This system has discontinuous coefficients, but it is a case in which the discontinuities are not a problem (cf. Section 3.6.1).

Substituting[18] $\kappa_\pm = 1/2 \pm \varepsilon - \varepsilon^2 + \mathcal{O}(\varepsilon^3)$, perform the somewhat messy calculation to deduce that $\operatorname{tr} \Pi = -2 - 16\varepsilon^2 + \mathcal{O}(\varepsilon^4)$. Apply the quadratic formula to conclude that for small nonzero ε, the absolute value of one of the eigenvalues of Π must be greater than unity.

3.5.3 PHD Exercises

16. Prove Theorem 3.4.1.

Remark: As we mentioned in the text, Theorem 3.4.1 can be proved using ideas from Chapter 4, and that proof is perhaps more in the mainstream of this book than the one outlined below. Nevertheless, the following proof has the merit of introducing a number of interesting and useful ideas from analysis.

Hint: The linear IVP (3.41) is equivalent to the integral equation

$$\mathbf{x}(t) = \int_0^t A(s)\mathbf{x}(s)\, ds + \mathbf{h}(t), \qquad T_1 < t < T_2, \tag{3.51}$$

where

$$\mathbf{h}(t) = \mathbf{b} + \int_0^t \mathbf{g}(s)\, ds.$$

If T_1 or T_2 is large, (3.51) need not define a contraction on $\mathcal{C}([T_1, T_2], \mathbb{R}^d)$. Nevertheless, let us show that Picard iteration converges uniformly for $t \in \mathcal{I}$, where $\mathcal{I}$ is an arbitrary compact subinterval of (T_1, T_2). To this end, we rewrite (3.51) more symbolically as

$$\mathbf{x} = \mathcal{L}\,\mathbf{x} + \mathbf{h}, \tag{3.52}$$

where $\mathcal{L}$ is the linear operator $\mathcal{L} : \mathcal{C}(\mathcal{I}, \mathbb{R}^d) \to \mathcal{C}(\mathcal{I}, \mathbb{R}^d)$ defined by

$$[\mathcal{L}\,\mathbf{x}](t) = \int_0^t A(s)\mathbf{x}(s)\, ds.$$

To begin Picard iteration on (3.52), let $\mathbf{x}_0 = \mathbf{h}$, and to continue, let $\mathbf{x}_{n+1} = \mathcal{L}\,\mathbf{x}_n + \mathbf{h}$. Check that

$$\mathbf{x}_n = [I + \mathcal{L} + \mathcal{L}^2 + \ldots + \mathcal{L}^n]\,\mathbf{h}.$$

Now show that the terms on the RHS of this equation may be written as iterated integrals: for $t \geq 0$, we have

$$[\mathcal{L}^k\,\mathbf{h}](t) = \int_0^t \int_0^{s_1} \cdots \int_0^{s_{k-1}} A(s_1)A(s_2)\ldots A(s_k)\mathbf{h}(s_k)\, ds_k ds_{k-1} \ldots ds_1.$$

Let $K = \max_{t\in\mathcal{I}} \|A(t)\|$ and let $\|\mathbf{h}\| = \max_{t\in\mathcal{I}} |\mathbf{h}|$. Estimating maxima in the above equation, conclude that

$$|\,[\mathcal{L}^k\,\mathbf{h}](t)| \leq K^k \|\mathbf{h}\| \int_0^t \int_0^{s_1} \cdots \int_0^{s_{k-1}} ds_k ds_{k-1} \ldots ds_1 = \frac{K^k t^k}{k!}\|\mathbf{h}\|.$$

[18] The letter $\mathcal{O}$ is mnemonic for *order*; thus $\mathcal{O}(\varepsilon^3)$ is shorthand for omitted terms that are of order ε^3 or higher. A more serious usage of this notation is introduced in Section 4.6.4.

Combining this estimate with an analogous estimate for $t \le 0$ gives $\|\mathcal{L}^k \mathbf{h}\| \le (K^k |\mathcal{I}|^k/k!)\,\|\mathbf{h}\|$, where $|\mathcal{I}|$ denotes the length of $\mathcal{I}$. With this estimate you can show that Picard iteration for (3.51) converges as claimed, which establishes existence.

Uniqueness follows from an application of Gronwall's lemma, as in the proof of Theorem 3.3.4.

17. Formulate and prove generalizations of the existence result Theorem 3.2.1 and the uniqueness result Theorem 3.3.4 to the nonautonomous problem (3.40), assuming that $\mathbf{G}$ is locally Lipschitz in *both* $\mathbf{x}$ and t.

Hint: Recall the transformation of $\mathbf{x}' = \mathbf{G}(\mathbf{x}, t)$ to an autonomous system in $d+1$ variables, as described in the Pearls of Chapter 1.

Discussion: Both existence and uniqueness theorems for (3.40) can be derived under a weaker hypothesis; it suffices if $\mathbf{G}$ is uniformly locally Lipschitz in $\mathbf{x}$ and merely continuous in t. For the record, if $\mathbf{G}(\mathbf{x}, t)$ is defined on $\mathcal{U} \times \mathcal{I}$, where $\mathcal{U} \subset \mathbb{R}^d$ is open and $\mathcal{I} \subset \mathbb{R}$ is an open interval, we say that $\mathbf{G}$ is *uniformly locally Lipschitz* in $\mathbf{x}$ if for every $(\mathbf{x}_0, t_0) \in \mathcal{U} \times \mathcal{I}$, there exist a neighborhood $\mathcal{V} \times \mathcal{J}$ of $(\mathbf{x}_0, t_0)$ and a constant L such that

$$(\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{V})\,(\forall t \in \mathcal{J})\ \ |\mathbf{G}(\mathbf{x}_1, t) - \mathbf{G}(\mathbf{x}_2, t)| \le\ L|\mathbf{x}_1 - \mathbf{x}_2|. \tag{3.53}$$

"Uniformly" refers to the fact that the same Lipschitz constant works for all t in the neighborhood $\mathcal{J}$. The generalizations with this weaker hypothesis are not especially important in themselves, but proving them offers practice with proofs that has some value.

18. (a) Find the general solution of the equations

$$tx' \pm x = 0$$

for $t > 0$.

Discussion: This may be solved either as a first-order linear equation or as a separable equation. Note that writing this equation in standard form yields $x' = \pm x/t$, where the RHS is singular at $t = 0$. The remainder of this exercise illustrates some nasty consequences of such a singularity.

(b) Deduce from your solution in Part (a) that the IVP

$$tx' + \ x = 0, \qquad x(0) = b$$

has no solutions.

(c) Deduce from your solution in Part (a) that the IVP

$$tx' - x = 0, \qquad x(0) = b$$

has no solutions if $b \neq 0$ and has infinitely many solutions if $b = 0$.

19. *Introduction:* Here is a physical situation that relates to our nonuniqueness example, (3.4). Consider a partially filled bucket that has a hole in its bottom. Under certain simplifying assumptions, the height $h(t)$ of the water in the bucket satisfies

$$\frac{dh}{dt} = -C\sqrt{h}.$$

We refer to Section 4.2 of [44] for details, but briefly, the derivation is (i) dh/dt is proportional to the speed v with which the water emerges from the bucket, and (ii) if friction is neglected, the kinetic energy (essentially, this means v^2) of the emerging water is proportional to the loss of potential energy (i.e., h). Without loss of generality, we may scale time to make $C = 1$ in this equation.

(a) Apply separability to solve the IVP

$$\frac{dx}{dt} = -\sqrt{x}, \qquad x(0) = b, \tag{3.54}$$

where $b > 0$, for $0 \le t < \infty$.

Hint: Observe that after some finite time, $x(t)$ reaches zero. On physical grounds one knows that $x(t) \equiv 0$ for all later times; this continuation is $\mathcal{C}^1$ and satisfies the equation.

(b) Argue that the solution of (3.54) is unique.

Hint: Equation (3.54) is meaningful only for nonnegative functions. Show that if $x(y)$ and $y(t)$ are both $\mathcal{C}^1$ nonnegative solutions of (3.54), then

$$\frac{d}{dt}[x(t) - y(t)]^2 = 2[x(t) - y(t)]\,[-\sqrt{x(t)} + \sqrt{y(t)}].$$

Argue that the RHS of this equation is nonpositive by considering $x \le y$ and $y \le x$ as two separate cases. Conclude that $[x(t)-y(t)]^2 \le [x(0)-y(0)]^2 = 0$.

Remark: Reinterpreting our nonuniqueness example from Section 1 in the present context, the IVP (in forward time) for (3.4) may be articulated as, "Suppose you come into the room and see that the bucket is empty; how full was it an hour ago?"

20. Construct a continuous function on $\mathbb{R}$ that is not Lipschitz over any open interval.

Hint: This exercise is based on imitating (3.55) in the Pearls. Let $\{q_k\}$ be an enumeration of the rational numbers. Form an infinite sum of terms involving $\sqrt{|x - q_k|}$ with a smooth denominator that keeps each term bounded.

3.6 Pearls of Wisdom

3.6.1 Miscellaneous

Although we have assumed Lipschitz continuity of $\mathbf{F}$ in proving the existence of solutions of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, this is not necessary. In fact, mere continuity of $\mathbf{F}$ suffices.

This is proved, for example, in Birkhoff–Rota [9], Chapter 6, Section 14. (Regarding uniqueness, as we saw in Section 3.1, continuity alone is not sufficient.)

An ODE in which the RHS is not continuous must be approached with caution, but it may nevertheless behave perfectly decently. We illustrate this claim with scalar ODEs. Consider

$$f_1(x) = \begin{cases} -1 & \text{if } x > 0 \\ -2 & \text{if } x \le 0, \end{cases} \qquad f_2(x) = \begin{cases} -1 & \text{if } x > 0 \\ 2 & \text{if } x \le 0. \end{cases}$$

It is natural to call the continuous piecewise differentiable function

$$x(t) = \begin{cases} 1 - t & \text{if } t \le 1 \\ -2(t-1) & \text{if } t > 1 \end{cases}$$

a solution of the ODE $x' = f_1(x)$ with initial condition $x(0) = 1$ even though x is not differentiable at $t = 1$, where it crosses zero; moreover, this generalized solution is unique. By contrast, there is no way to continue the solution of $x' = f_2(x)$ with initial condition $x(0) = 1$ past $t = 1$.

Interpreting a discontinuous ODE is simpler if the discontinuous behavior in a nonautonomous equation $\mathbf{x}' = \mathbf{G}(\mathbf{x}, t)$ is isolated in the t-dependence of $\mathbf{G}$. For instance, in (3.47), near the discontinuity of $S(t)$ at $t = \pi$, the obvious solution is obtained by solving the equation $x'' + (\kappa + \varepsilon)x = 0$ for $t \le \pi$ and then using $x(\pi), x'(\pi)$, i.e., this solution at the final time, as initial conditions to solve $x'' + (\kappa - \varepsilon)x = 0$ for the next range of t, etc.

Let's construct a more dramatic example than $|x|$ that a Lipschitz function need not be continuously differentiable. The starting point in the example is the function $|x|/\sqrt{1+x^2}$, which is Lipschitz continuous with constant 1 and fails to be differentiable at $x = 0$; we include the denominator so that the function is bounded over all x. Let $\{q_k\}$ be an enumeration of the rational numbers and let

$$F(x) = \sum_{k=1}^{\infty} k^{-2} \, \frac{|x - q_k|}{\sqrt{1 + (x - q_k)^2}}. \tag{3.55}$$

Then F is Lipschitz continuous (with Lipschitz constant $L = \sum_1^\infty k^{-2}$) but is not differentiable on any open interval.

Let's compare use of the contraction-mapping formalism to Picard iteration in proving Theorem 3.2.1. The latter has two advantages: (i) It avoids the abstractness of the contraction-mapping principle. (ii) The iteration may converge over a larger interval of t than what's needed to guarantee that the operator $\mathfrak{T}$ in (3.23) is

a contraction (cf. Exercise 16). Two points favoring the former:[19] (i) In our view, the contraction-mapping formalism actually clarifies the proof of Theorem 3.2.1 by guiding you to exactly what is needed to guarantee that the iteration may be continued indefinitely and that it converges, i.e., Claims 1 and 2 in Section 3.2.5. (ii) In general, determining the exact range of t for which Picard iteration converges is a nightmarish problem; it is simpler to estimate by other means the range of t in which the IVP actually has a solution, which is the approach taken in Chapter 4.

3.6.2 Resonance

If an ODE with oscillatory solutions is subjected to a periodic perturbation, the perturbation may have a large effect over time if its frequency matches the natural frequency of oscillations of the ODE. This phenomenon is called *resonance*. We saw an example of resonance in Exercise 1.13. In that problem, the periodic disturbance occurred in an inhomogeneous term not involving the dependent variable x. In equation (3.46), which also exhibits resonance, the periodic disturbance occurs in a coefficient of x. This situation is known as *parametric resonance*, a term that comes from regarding the coefficients in a linear ODE as "parameters." (Of course, this viewpoint is slightly oxymoronic, since parameters are supposed to be constant, but hey, we didn't make up the term.)

Physically, (3.46) may be viewed as a spring–mass system in which the spring constant varies with time. Figure 3.2 illustrates how, if appropriately timed, such a perturbation can amplify oscillations of the system. Every child who can pump a swing[20] understands this at a gut level.

[19] These points apply to the context of Theorem 3.2.1. In more general contexts, such as *stochastic differential equations*, Picard iteration may be the preferable method.

[20] A minor clarification: pumping a swing is an example of parametric resonance in a *nonlinear* problem: a pendulum.

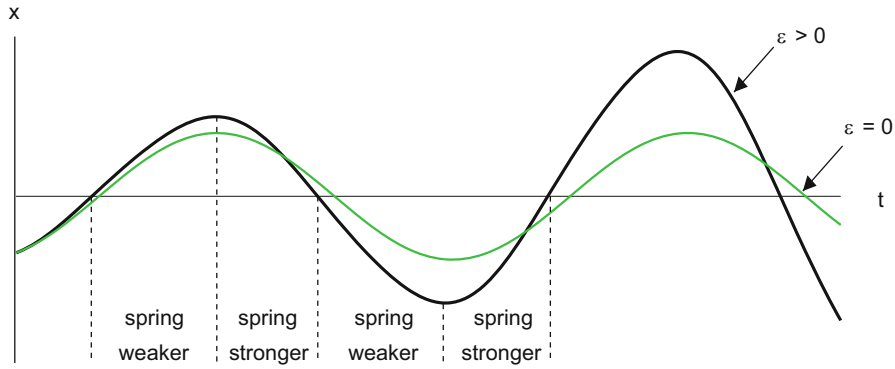

Figure 3.2: *Unperturbed ($\varepsilon = 0$, shown in green) and perturbed ($\varepsilon = 0.05$, shown in black) solutions of the Mathieu equation (3.46) starting from the same initial conditions. (Cf. Exercise 13.) Resonance occurs in the perturbed system because the spring is weaker when x moves away from equilibrium and stronger when x is pulled back toward equilibrium.*

Chapter 4

Nonlinear Systems: Global Theory

Theorem 3.2.1 guarantees that a solution to the initial value problem exists for what might be an extremely short time. Typically, the ODEs that arise in physical applications possess solutions for much longer times than can be deduced with the contraction-mapping principle, and in this chapter we introduce methods for demonstrating this behavior. The technique is based on extending a short-time solution obtained from Theorem 3.2.1. The main tool for such extensions, Theorem 4.1.2, is proved in Section 4.1. In Section 4.2, this theorem is used to derive two theoretical results that guarantee global existence.

In Sections 4.3 and 4.4 we introduce techniques, including *nullclines*, for verifying the hypotheses of these global existence theorems, and we illustrate the techniques by applying them to specific ODEs drawn from various fields. These equations will reoccur frequently in later chapters.

We regard the introduction of meaningful applications to illustrate the theory as one of the attractive features of this book. In the present chapter we consider the ODEs, without motivation, from a purely mathematical point of view; this analysis completes the theoretical treatment of IVPs begun in Chapters 2 and 3. (In Chapter 5 we introduce models in their original form, including interpretation of the variables and underlying physical assumptions. Central to that chapter, we study scaling as a systematic technique to simplify the original equations to forms more convenient for analysis, as considered in the present chapter.)

In the last two sections of this chapter we show that the solution of an IVP depends continuously (Section 4.5) and even differentiably (Section 4.6) on its initial conditions. These results are a fundamental part of the theory.

An appendix is devoted to Euler's method, the simplest numerical approximation for solutions of an IVP. In particular, we prove that as the step size tends to zero, the

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8_4

approximations converge to the true solution. This proof closely mimics the proof in Section 4.5 that the solution of an IVP depends continuously on its initial data.

4.1 The Maximal Interval of Existence

Our first result asserts that there is a maximal interval for which the solution of an IVP exists, a sort of "gold standard" for solutions.

Proposition 4.1.1. *Let $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$ be locally Lipschitz on $\mathcal{U} \subset \mathbb{R}^d$. Given $\mathbf{b} \in \mathcal{U}$, there is a solution $\mathbf{x}_* : (-\alpha_*, \beta_*) \to \mathcal{U}$ of the IVP*

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0) = \mathbf{b} \tag{4.1}$$

that is maximal in the following sense: if another function $\mathbf{x}$ solves (4.1) for t in some open interval $\mathcal{I}$, then

$$\textit{(i)}\ \mathcal{I} \subset (-\alpha_*, \beta_*) \qquad \textit{and} \qquad \textit{(ii)}\ \mathbf{x}(t) = \mathbf{x}_*(t)\ \textit{for}\ t \in \mathcal{I}. \tag{4.2}$$

Remark: It often happens that either α_* or β_*, or both, equals infinity. Note that the maximal interval of existence is always open, even if α_* or β_* is finite.

Proof. We focus only on β_* and $t \geq 0$, leaving the analogous treatment of α_* and $t \leq 0$ for the dedicated reader. Let

$$\beta_* = \sup\{\beta : \text{ IVP (4.1) is solvable for } 0 \leq t < \beta\}.$$

Of course, by Theorem 3.2.1, $\beta_* > 0$. For $n = 1, 2, \ldots$, choose solutions $\mathbf{x}_n$ of (4.1) that exist for times $t \in [0, \beta_n)$, where $\beta_n \to \beta_*$, finite or infinite. To define $\mathbf{x}_*$, given $t \in [0, \beta_*)$ choose any n such that $\beta_n > t$ and let

$$\mathbf{x}_*(t) = \mathbf{x}_n(t). \tag{4.3}$$

By Theorem 3.3.4, the uniqueness result, the definition (4.3) does not depend on the choice of n, and moreover, $\mathbf{x}_*$ is a solution of (4.1). It is readily checked *(do so!)* that every solution $\mathbf{x}$ of (4.1) on some interval $\mathcal{I}$ satisfies properties (i) and (ii) of (4.2). □

Although it may not be apparent, the following result is extremely useful in extending solutions to larger times. Of course there is an analogous result for negative time.

Theorem 4.1.2. *Suppose, regarding the maximal solution $\mathbf{x}_* : (-\alpha_*, \beta_*) \to \mathbb{R}^d$, that $\beta_* < \infty$. Then for every compact set $\mathcal{K} \subset \mathcal{U}$, there is an $\varepsilon > 0$ such that $\mathbf{x}_*(t) \notin \mathcal{K}$ for $\beta_* - \varepsilon < t < \beta_*$.*

Proof. By Corollary 3.3.3, there is a compact set $\mathcal{K}'$ and a $\delta > 0$ such that for all $\mathbf{x} \in \mathcal{K}$, the closed ball $\overline{B(\mathbf{x}, \delta)}$ is contained in $\mathcal{K}'$. Let $M = \max_{\mathcal{K}'} |\mathbf{F}(\mathbf{x})|$, let L be a Lipschitz constant for $\mathbf{F}$ on $\mathcal{K}'$, and choose $\varepsilon < \min\{\delta/M, 1/L\}$.

We claim that $\mathbf{x}_*(t) \notin \mathcal{K}$ if $t > \beta_* - \varepsilon$. Suppose to the contrary that there is a time $t_0 > \beta_* - \varepsilon$ such that $\mathbf{x}_*(t_0) \in \mathcal{K}$. It follows from Corollary 3.2.8 that the IVP

$$\mathbf{y}' = \mathbf{F}(\mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{x}_*(t_0)$$

has a solution on $(t_0 - \varepsilon, t_0 + \varepsilon)$. Applying Lemma 3.2.9, we conclude that the original solution $\mathbf{x}_*$ may be defined on $[0, t_0 + \varepsilon)$. But $t_0 + \varepsilon > \beta_*$, which contradicts the hypothesis that β_* was maximal. □

4.2 Two Sufficient Conditions for Global Existence

4.2.1 Linear Growth of the RHS

Our first result[1] gives existence for all times, positive and negative.

Theorem 4.2.1. *If $\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^d$ is locally Lipschitz and if there exist nonnegative constants B, K such that*

$$|\mathbf{F}(\mathbf{x})| \le K|\mathbf{x}| + B, \qquad \mathbf{x} \in \mathbb{R}^d, \tag{4.4}$$

then the solution $\mathbf{x}(t)$ of (4.1) exists for all time, $-\infty < t < \infty$, and moreover,

$$|\mathbf{x}(t)| \le |\mathbf{b}|e^{K|t|} + \frac{B}{K}(e^{K|t|} - 1), \qquad -\infty < t < \infty. \tag{4.5}$$

Proof. This proof will use the generalization of Gronwall's lemma given in Exercise 3.8(a). We consider only forward time, $t \ge 0$; negative time can be handled with trivial modifications of the argument. Suppose (4.1) has a solution for $t \in [0, \beta)$, which of course satisfies the integral equation

$$\mathbf{x}(t) = \mathbf{b} + \int_0^t \mathbf{F}(\mathbf{x}(s))\, ds, \qquad 0 \le t < \beta.$$

Defining $g(t) = |\mathbf{x}(t)|$, we deduce that

$$g(t) \le |\mathbf{b}| + \int_0^t [Kg(s) + B]\, ds, \qquad 0 \le t < \beta.$$

Hence by the generalized Gronwall lemma, $\mathbf{x}$ satisfies the estimate (4.5) for its entire domain of existence, $0 \le t < \beta$.

[1] We alert you one final time: the same symbol $\mathbf{x}$ may simply denote a point in $\mathbb{R}^d$ (as in (4.4)) or may denote a vector-valued function of time (as in (4.5)).

Now let $\mathbf{x}_*, \beta_*$ be the maximal solution of (4.1), and suppose $\beta_* < \infty$. According to (4.5), $\mathbf{x}_*(t)$ belongs to the compact ball

$$\mathcal{K} = \{\mathbf{z} \in \mathbb{R}^d : |\mathbf{z}| \le |\mathbf{b}|e^{K\beta_*} + \frac{B}{K}(e^{K\beta_*} - 1)\}$$

for all $t \in [0, \beta_*)$. This estimate contradicts Theorem 4.1.2, so we must have β_* infinite. □

This result may be easily extended to nonautonomous equations that satisfy a linear-growth estimate. (See Exercise 5(a).)

4.2.2 Trapping Regions[2]

Our second result, which gives global existence in forward time only, is more widely applicable but also requires more explanation.

(a) An introductory example

Rewriting Duffing's equation (1.28) as a first-order system, we obtain

$$\begin{aligned} x' &= y \\ y' &= -\beta y + x - x^3. \end{aligned} \tag{4.6}$$

Let us repeat the calculation from Section 1.4.1 that the energy

$$E(x, y) = y^2/2 - x^2/2 + x^4/4 \tag{4.7}$$

decreases along trajectories of (4.6). Indeed, by the chain rule,

$$\frac{dE}{dt} = \frac{\partial E}{\partial x}\frac{dx}{dt} + \frac{\partial E}{\partial y}\frac{dy}{dt} = \langle \nabla E, \mathbf{F} \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the inner product on $\mathbb{R}^2$. Obtaining ∇E from (4.7) and $\mathbf{F}$ from (4.6), we calculate that

$$\langle \nabla E, \mathbf{F} \rangle = -\beta y^2 \le 0. \tag{4.8}$$

For a constant E_0, consider the sublevel set

$$\mathcal{K} = \{(x, y) \in \mathbb{R}^2 : E(x, y) \le E_0\}, \tag{4.9}$$

[2]Trapping regions represent a first hint of a shift toward more geometric thinking in this book. In this connection, you may be amused by the aphorism, "Geometry is the art of reasoning well from badly drawn figures." The oldest citation that we can give for this is from an article by Poincaré in 1895 [64], but apparently the quotation was old even at that time, for Poincaré introduces it with the remark, "It is worth repeating . . . ."

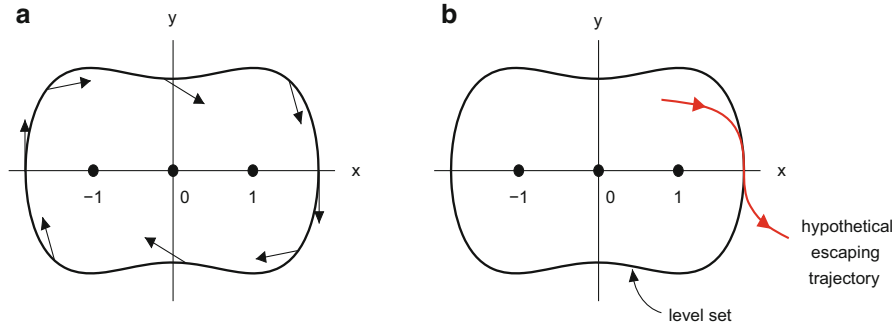

Figure 4.1: *(a) The level set $E(x, y) = 1$, where the energy $E(x, y)$ is given by (4.7), and the flow direction for Duffing's equation (4.6) at selected points on the curve, assuming $\beta = 1$. The flow is strictly inward except for $y = 0$, where it is tangential. (b) Hypothetical escaping trajectory for the Duffing system. Were such a trajectory possible, it would have to cross the level set tangentially at one of the two points where the flow is not strictly inward (i.e., along the x-axis).*

whose boundary is the level set

$$\partial\mathcal{K} = \{(x, y) \in \mathbb{R}^2 : E(x, y) = E_0\}. \tag{4.10}$$

Provided[3] that $E_0 > 0$, formula (4.10) defines a smooth closed curve (see Figure 4.1). Now ∇E is normal to the level set; since ∇E points in the direction of increasing E, the *inward* normal is given by $\mathbf{N} = -\nabla E$. Thus, we may rewrite (4.8) as

$$\langle \mathbf{N}_\mathbf{x}, \mathbf{F}(\mathbf{x}) \rangle \geq 0 \qquad (\mathbf{x} \in \partial\mathcal{K}).$$

In words, the direction of the flow of the ODE at $\partial\mathcal{K}$ is inward, or at worst tangential or zero.

As we now show, if an inequality of this type holds on the boundary of a region, then the solution of the IVP (for positive times) is trapped inside that region.

(b) Statement and discussion of the result

Consider the IVP for an ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, where $\mathbf{F}$ is defined on an open subset $\mathcal{U}$ of $\mathbb{R}^d$. Let $\mathcal{K}$ be a closed subset of $\mathcal{U}$ whose boundary is a $\mathcal{C}^1$ surface (see Section B.3.3 for definitions), and for $\mathbf{x} \in \partial\mathcal{K}$, let $\mathbf{N}_\mathbf{x}$ be an inward normal to $\partial\mathcal{K}$ at $\mathbf{x}$. We shall

[3]The topology of the set (4.9) changes if $E_0 < 0$. For simplicity, we sidestep this complication.

call $\mathcal{K}$ a *trapping region*[4] for $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ if

$$(\forall \mathbf{x} \in \partial\mathcal{K})\ \langle \mathbf{N}_\mathbf{x}, \mathbf{F}(\mathbf{x})\rangle \geq 0. \tag{4.11}$$

Theorem 4.2.2. *Suppose that $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$ is $\mathcal{C}^1$ and that $\mathcal{K}$ is a compact trapping region for $\mathbf{x}' = \mathbf{F}(\mathbf{x})$. If the initial data $\mathbf{b}$ lies in the interior of $\mathcal{K}$, then the solution $\mathbf{x}$ to equation (4.1) exists for all positive time and moreover lies in the interior of $\mathcal{K}$.*

If the inequality (4.11) is replaced by strict inequality,

$$(\forall \mathbf{x} \in \partial\mathcal{K})\ \langle \mathbf{N}_\mathbf{x}, \mathbf{F}(\mathbf{x})\rangle > 0, \tag{4.12}$$

only a few lines suffice[5] to prove this result:

Proof of Theorem 4.2.2 assuming (4.12). By Theorem 4.1.2, the solution may cease to exist only if it first leaves $\mathcal{K}$. If, coming from inside $\mathcal{K}$, the trajectory $\mathbf{x}(t)$ reaches a point on $\partial\mathcal{K}$, then at that point the (outward) normal velocity must be nonnegative, and this contradicts the trapping hypothesis (4.12). □

If we have only the weaker hypothesis (4.11), our proof must rule out the possibility illustrated in Figure 4.1(b) for Duffing's equation (4.6): could a trajectory escape tangentially from $\mathcal{K}$ at a point where the inner product (4.8) vanishes?[6] This proof is somewhat technical and does not contain fundamental new ideas.[7] You may safely postpone reading it, but since the full result is useful, we will feel free to invoke it below in studying specific examples.

(c) Proof of Theorem 4.2.2. By Theorem 4.1.2, the solution may cease to exist only if it first leaves $\mathcal{K}$. Let

$$t_* = \sup\{t : (\forall s \leq t)\ \mathbf{x}(s) \in \text{Int } \mathcal{K}\}. \tag{4.13}$$

[4]Some authors reserve the word "region" to describe an open set. Note that we are not following that convention here.

[5]If you feel that these remarks are too sketchy to constitute a real proof, we urge you to revisit them after reading the proof of the full result.

[6]For Duffing's equation, we can argue that such an escape is not possible, since the function $E(x, y)$ is nonincreasing along orbits. However, this argument uses the fact that $E(x, y)$ is defined in a neighborhood of $\partial\mathcal{K}$, while we want to prove the theorem using only information derived from the fact that (4.11) holds on $\partial\mathcal{K}$.

[7]Indeed, the argument is only a minor extension of what you already were asked to do in Exercise 1.16.

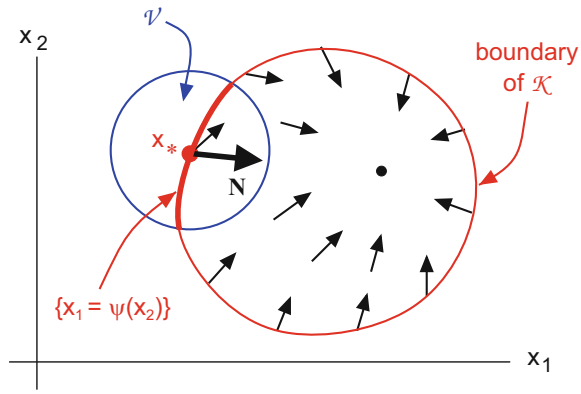

Figure 4.2: *Schematic of a trapping region $\mathcal{K}$ and the data of (4.14). Remark: Note that the vector field* **F** *has an equilibrium point inside the trapping region; in two dimensions, this cannot be avoided if $\mathcal{K}$ is simply connected.*

By the local existence theorem, $t_* > 0$. We suppose $t_* < \infty$ and look for a contradiction. Of course at the supremum (4.13), $\mathbf{x}(t_*) \in \partial\mathcal{K}$. For brevity we write $\mathbf{x}_* = \mathbf{x}(t_*)$.

Since $\partial\mathcal{K}$ is a $\mathcal{C}^1$ surface, near the hypothetical exit point $\mathbf{x}_*$, one of the coordinates, say x_1, may be expressed as a function of the others. That is, there exist a neighborhood $\mathcal{V} \subset \mathcal{U}$ of $\mathbf{x}_*$ and a function $\psi : \mathcal{V} \to \mathbb{R}$, independent of x_1, such that

$$\partial\mathcal{K} \cap \mathcal{V} = \{\mathbf{x} \in \mathcal{V} : x_1 = \psi(\tilde{\mathbf{x}})\}, \tag{4.14}$$

where $\tilde{\mathbf{x}}$ is shorthand for $(x_2, \ldots, x_d)$. Moreover, reversing signs of both x_1 and ψ if necessary, we may assume that $x_1 > \psi(\tilde{\mathbf{x}})$ on the interior of $\mathcal{K}$, or $\mathbf{N} = (1, -\tilde{\nabla}\psi)$ is an *inward* normal along $\partial\mathcal{K}$. (Cf. Figure 4.2.)

By continuity, there is an interval $[t_* - \delta, t_*]$ such that $\mathbf{x}(t) \in \mathcal{V}$ for t in this interval. Let

$$g(t) = x_1(t) - \psi(\tilde{\mathbf{x}}(t)), \qquad t_* - \delta \le t \le t_*. \tag{4.15}$$

Then $g(t) > 0$ for $t_* - \delta \le t < t_*$, while $g(t_*) = 0$. On the other hand, we claim that

$$g'(t) \ge -Kg(t), \qquad t_* - \delta \le t \le t_* \tag{4.16}$$

for some constant K. Given (4.16), it follows that $(d/dt)[e^{Kt}g] \ge 0$, so

$$e^{Kt_*}g(t_*) \ge e^{K(t_*-\delta)}g(t_* - \delta).$$

But $g(t_* - \delta) > 0$, and hence this inequality implies that $g(t_*) > 0$, which is a contradiction that will prove Theorem 4.2.2.

It remains to prove the claim, (4.16). Applying the chain rule to (4.15), we

calculate that $g'(t) = G(\mathbf{x}(t))$, where

$$G(\mathbf{x}) = F_1(\mathbf{x}) - \sum_{j=2}^{d} \frac{\partial \psi}{\partial x_j}(\tilde{\mathbf{x}}) F_j(\mathbf{x}). \tag{4.17}$$

This function need not be $\mathcal{C}^1$, because ψ may not have the requisite smoothness, but it is continuously differentiable *with respect to the first component*, x_1. Also, if $\mathbf{x} \in \partial\mathcal{K}$, then

$$G(\mathbf{x}) = \langle \mathbf{N}_{\mathbf{x}}, \mathbf{F}(\mathbf{x}) \rangle \geq 0. \tag{4.18}$$

To estimate G at a general point $(x_1, \tilde{\mathbf{x}})$ in $\mathcal{K} \cap \mathcal{V}$, we add and subtract G evaluated at a nearby point $(\psi(\tilde{\mathbf{x}}), \tilde{\mathbf{x}})$ on $\partial\mathcal{K}$:

$$G(x_1, \tilde{\mathbf{x}}) = [G(x_1, \tilde{\mathbf{x}}) - G(\psi(\tilde{\mathbf{x}}), \tilde{\mathbf{x}})] + G(\psi(\tilde{\mathbf{x}}), \tilde{\mathbf{x}}). \tag{4.19}$$

Invoking (4.18) to drop the second term, we have

$$G(x_1, \tilde{\mathbf{x}}) \geq G(x_1, \tilde{\mathbf{x}}) - G(\psi(\tilde{\mathbf{x}}), \tilde{\mathbf{x}}). \tag{4.20}$$

We apply the fundamental theorem of calculus to rewrite the RHS of (4.20) as

$$G(x_1, \tilde{\mathbf{x}}) - G(\psi(\tilde{\mathbf{x}}), \tilde{\mathbf{x}}) = \int_{\psi(\tilde{\mathbf{x}})}^{x_1} \frac{\partial G}{\partial x_1}(s, \tilde{\mathbf{x}})\, ds.$$

Since $\partial G / \partial x_1$ is continuous, by compactness it is bounded on $\mathcal{K} \cap \overline{\mathcal{V}}$, say by the constant K. Therefore,

$$G(x_1, \tilde{\mathbf{x}}) - G(\psi(\tilde{\mathbf{x}}), \tilde{\mathbf{x}}) \geq -K[x_1 - \psi(\tilde{\mathbf{x}})],$$

which we may substitute into the RHS of (4.20). Thus,

$$g'(t) = G(\mathbf{x}(t)) \geq -K[x_1(t) - \psi(\tilde{\mathbf{x}}(t))] = -Kg(t),$$

as claimed in (4.16). The proof of Theorem 4.2.2 is now complete. □

(d) Various generalizations of the theorem

Theorem 4.2.2 may be generalized to trapping regions whose boundary is only piecewise smooth, which provides a *much* more versatile tool. Here is such a result for two-dimensional problems, which we ask you to prove in Exercise 2. Note that the trapping condition (4.11) need not be explicitly imposed at corner points[8] of the boundary; information derived from continuity suffices to handle the corners.

[8]If this term is unclear, see Section B.3.3(b), where the distinction between regular points and corner points on the boundary is defined.

Theorem 4.2.3. *Suppose that* $\mathbf{F} : \mathcal{U} \to \mathbb{R}^2$ *is* $\mathcal{C}^1$ *on a domain* $\mathcal{U} \subset \mathbb{R}^2$ *and that* $\mathcal{K} \subset \mathcal{U}$ *is a compact region with a piecewise smooth boundary such that* (4.11) *holds at all regular points of* $\partial\mathcal{K}$. *If the initial data* $\mathbf{b}$ *lies in the interior of* $\mathcal{K}$, *then the solution* $\mathbf{x}$ *to equation* (4.1) *exists for all positive time and moreover lies in the interior of* $\mathcal{K}$.

Of course analogous results hold in dimensions higher than two, but it is rather tedious to deal carefully with all possible cases, and not much is learned in doing so. We do not formulate any such generalization.

With a minor revision of the proof of Theorem 4.2.2, you may show that even if a trapping region is not compact, the following conclusion still holds. (*Check this result! We will use it below.*)

Corollary 4.2.4. *Suppose that* $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$ *is* $\mathcal{C}^1$ *and that* $\mathcal{K}$ *is a trapping region for* $\mathbf{x}' = \mathbf{F}(\mathbf{x})$. *If the initial data* $\mathbf{b}$ *lies in the interior of* $\mathcal{K}$, *then the solution* $\mathbf{x}$ *to equation* (4.1) *lies in the interior of* $\mathcal{K}$ *for as long as this solution continues to exist.*

Two further generalizations: (i) If the initial data $\mathbf{b}$ of an IVP belongs to the *boundary* of a compact trapping region, global existence may still be deduced. (ii) Theorem 4.2.2 may be extended to nonautonomous equations. In practice, neither of these results turns out to be terribly useful, and we do not pursue them.

4.3 Level Sets and Trapping Regions

4.3.1 Introduction via Duffing's Equation

In Section 4.2.2(a), we observed that sublevel sets (4.9) of the energy function (4.7) are trapping regions for Duffing's equation (4.6). This fact provides an easy global existence proof for the IVP for this equation. For every E_0, the set (4.9) is a compact trapping region. Given initial conditions $\mathbf{b} \in \mathbb{R}^2$, choose E_0 large enough that $\mathbf{b} \in \mathcal{K}$. By invoking Theorem 4.2.2, we obtain existence for all positive time.

For many other ODEs as well, level sets of some auxiliary function(s) may be used to construct trapping regions; we present two such examples.

4.3.2 The Chemostat

The chemostat is described by the scaled ODEs

$$\begin{aligned} x' &= \frac{y}{y+1}x - \rho x, \\ y' &= -\frac{y}{y+1}x - \rho(y-\sigma), \end{aligned} \tag{4.21}$$

where ρ, σ are positive constants. The variables x and y are concentrations, so they are nonnegative. The linear terms $-\rho x$ and $-\rho y$ represent decay, and the constant term $\rho\sigma$ in the second equation represents replenishment of y. The assumptions leading to the nonlinear terms will be explained in Section 5.5. Even without understanding the basis for these terms, we can see that they tend to increase x at the expense of y. Since these terms differ only by a minus sign, we may add the equations and deduce a third ODE

$$x' + y' = -\rho(x + y) + \rho\sigma \tag{4.22}$$

that will be useful in analyzing global existence.

Let's seek a triangular trapping region of the form

$$\mathcal{K} = \{(x, y) \in \mathbb{R}^2 : x \geq 0,\ y \geq 0,\ x + y \leq A\}, \tag{4.23}$$

where A is a constant. Along the sloping face of $\partial\mathcal{K}$, the inward normal is given by $\mathbf{N} = (-1, -1)$, and we observe from (4.22) that $\langle \mathbf{N}, \mathbf{F} \rangle = -x' - y' = \rho(A - \sigma)$. In particular, provided $A \geq \sigma$, we have $\langle \mathbf{N}, \mathbf{F} \rangle \geq 0$ along this face, which shows that flow is inward here. Regarding the other two sides, you may check that the flow of (4.21) is inward along the x-axis and is tangential along the y-axis. Therefore, $\mathcal{K}$ is a compact trapping region.

For every initial condition $\mathbf{b}$ in the first quadrant, the constant A in (4.23) may be chosen large enough that $\mathbf{b}$ belongs to the trapping region $\mathcal{K}$. Therefore, we may invoke Theorem 4.2.3 to obtain global existence for the IVP for (4.21).

The lesson to take away from this example is that we have used level sets of the linear function $L(x, y) = x + y$ in constructing trapping regions for (4.21).

4.3.3 The Torqued Pendulum and ODEs on Manifolds[9]

Consider a pendulum, as illustrated in Figure 4.3, that is subjected to a "torque" μ, which tends to twist the unperturbed pendulum away from its stable, straight-down equilibrium. If friction is modeled by linear damping, then after appropriate scaling,

[9]Although we use the general term "manifold" here, in fact we need only a couple of special cases (like the circle S^1) with which you are probably already familiar. A manifold is a topological space in which each point has a neighborhood homeomorphic to a ball in Euclidean space, subject to some compatibility conditions. If you want precise definitions, you may find these in Section 2.7 of [63] or look online, but we expect that most readers will not need to consult other sources to read the present section.

this problem may be described by the first-order system[10]

$$\begin{aligned} x' &= y, \\ y' &= -\sin x - \beta y + \mu, \end{aligned} \tag{4.24}$$

where $\beta > 0$ and μ are constants. Without loss of generality, it suffices to consider the case $\mu \geq 0$. *(Why?)*

We seek a trapping region derived from the total energy, kinetic plus potential, of the pendulum, which is given by

$$E(x, y) = y^2/2 - \cos x. \tag{4.25}$$

Thus, given a constant E_0, we consider the sublevel set

$$\mathcal{K} = \{(x, y) \in \mathbb{R}^2 : \ y^2/2 - \cos x \leq E_0\}. \tag{4.26}$$

Substitution into (4.24) yields

$$\frac{dE}{dt} = \langle \nabla E, \mathbf{F} \rangle = -\beta y^2 + \mu y.$$

In general, dE/dt may have either sign, but if $|y| > \mu/\beta$, then $dE/dt = -\beta y(y - \mu/\beta)$ is negative. At the boundary of (4.26), $y = \pm\sqrt{2(E_0 + \cos x)}$, so $|y| > \mu/\beta$, provided E_0 is sufficiently large. Doing the calculation, we see that (4.26) is a trapping region for (4.24), provided $E_0 \geq (\mu/\beta)^2/2 + 1$. Increasing E_0 if necessary, we may also assume that $\mathcal{K}$ contains any proposed initial data.

However, our attempt to apply Theorem 4.2.2 is thwarted by the fact that $\mathcal{K}$ is not compact (see Figure 4.4a). In principle, a solution of (4.24) might stay inside $\mathcal{K}$ while x marches off to infinity in finite time. In fact, this does not happen: the RHS of (4.24) satisfies the hypothesis (4.4) of Theorem 4.2.1, so the solution exists for all $t \in \mathbb{R}$.

The trapping-region existence proof may be revived with the addition of some geometry. Note that the RHS of (4.24) is 2π-periodic in x. Therefore, rather than considering (4.24) as an ODE on the Euclidean space $\mathbb{R} \times \mathbb{R}$, we may regard it as an ODE on the cylinder[11] $S^1 \times \mathbb{R}$, where $S^1 = \mathbb{R}/2\pi\mathbb{Z}$ is the circle. If $\mathcal{K}$ is considered a subset of $S^1 \times \mathbb{R}$, then this set *is* compact (see Figure 4.4b). Moreover, as you will show in Exercise 6, Theorem 4.2.2 generalizes to trapping regions on $S^1 \times \mathbb{R}$, and thus we may obtain a different, in our view more elegant, proof of global existence in forward time for (4.24). As an added bonus, using this approach, you will find

[10]Incidentally, the equations (4.24) also describe the behavior of an electrical device known as the *Josephson junction*. See Strogatz [81], Section 4.6, for details.

[11]You have already encountered an equation on a cylinder: in using polar coordinates to study an ODE in the plane, you obtain an ODE in which r, θ belong to the manifold $(0, \infty) \times S^1$.

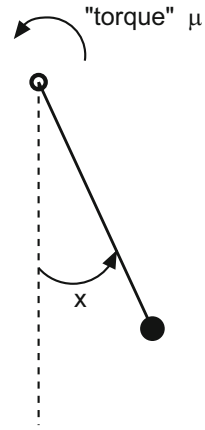

Figure 4.3: *Schematic diagram of the torqued pendulum.*

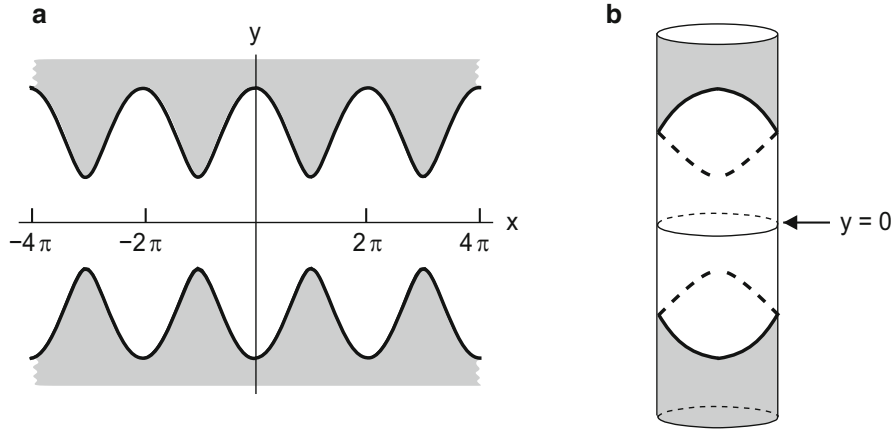

Figure 4.4: *Sketch of the trapping region $\mathcal{K}$ for the torqued pendulum, (4.26). In Panel (a), $\mathcal{K}$ is the unbounded region that lies between the two curves. In Panel (b), the x, y-plane is wrapped around a cylinder, so that values of x differing by 2π are identified.*

that the solution grows at worst linearly as $t \to \infty$. This conclusion is stronger than what can be deduced using Theorem 4.2.1.

In this book we do not attempt to extend Theorem 4.2.2 to ODEs on general manifolds, even though given the right technical background, this is not difficult. Rather, we consider ODEs only on two specific manifolds, the cylinder $S^1 \times \mathbb{R}$ and the torus $\mathbb{T}^2 = S^1 \times S^1$. For these two special cases, the generalization of Theorem 4.2.2 may be proved with ad hoc arguments using (multivalued) Euclidean coordinates, as in Exercise 6.

4.4 Nullclines and Trapping Regions

4.4.1 Nullclines in the Chemostat

The term *nullcline* refers to a curve[12] where one of the components of the velocity vector vanishes. For example, consider the chemostat (4.21). We have that x' vanishes if

$$x = 0 \qquad \text{or} \qquad y = \frac{\rho}{1-\rho}, \tag{4.27}$$

while y' vanishes if

$$x = -\rho \, \frac{(y-\sigma)(y+1)}{y}. \tag{4.28}$$

The portions of both curves that lie in the first quadrant are graphed in Figure 4.5, in brown for the x-nullclines (4.27) and in cyan[13] for the y-nullcline (4.28). Intersections of the nullclines at

$$(x, y) = (0, \sigma) \quad \text{and} \quad \left(\sigma - \frac{\rho}{1-\rho}, \frac{\rho}{1-\rho} \right) \tag{4.29}$$

are equilibrium solutions of the ODE. In the figure we show the more interesting case, for which the second equilibrium lies in the first quadrant, i.e., parameters such that $\rho < 1$ and $\sigma > \rho/(1-\rho)$.

To extract information from the nullclines with minimal pain, it is helpful to proceed in the three stages represented in Figure 4.5. Whenever you need to graph nullclines, *we urge you to follow this three-stage procedure*. Being systematic in this way reduces (slightly) the opportunities for making careless mistakes, of which there are plenty.

---

[12] "Curve" is the appropriate word for two-dimensional systems, which is the usual context in which nullclines are studied. For higher-dimensional systems, one should say the "set where ..." or "surface where ...."

[13] Whenever we plot nullclines, we will follow this color convention.

- In Figure 4.5(a), vertical lines are drawn along the x-nullclines (4.27) because the flow is vertical there, i.e., $x' = 0$. Similarly, horizontal lines have been drawn along the y-nullcline (4.28).

- Figure 4.5(b) augments the previous figure by specifying along the x-nullclines whether the flow is up or down; and along the y-nullcline whether the flow is to the right or left. Here is the thinking behind the construction of this figure. The orientation of the flow along (4.27) changes from up to down[14] whenever this curve crosses a nullcline of the other family. Alternatively put, on every segment of (4.27) that does *not* intersect (4.28), the orientation of the flow does *not* change. It may be seen from (4.21b) that y' is positive at the origin. As shown in the figure, the flow remains upward as one moves away from the origin until points where (4.28) is crossed, causing the direction to reverse itself. Similarly, for the other nullcline, start by observing from (4.21a) that far out on (4.28), near the x-axis, x' is negative. Then the other horizontal arrows along (4.28) in Figure 4.5(b) may be constructed by reversing direction whenever (4.27) is crossed.

- The nullclines partition the first quadrant into regions. Within one region the flow $\mathbf{F}(x, y)$ points into one of the four quadrants, $\{\pm x > 0, \pm y > 0\}$, and the quadrant remains the same if (x, y) moves within this region. The quadrant of the flow in each of the regions is indicated by a thick black arrow in Figure 4.5(c). Again, one can complete this figure by analyzing a special case (e.g., along the x-axis, the flow points into the second quadrant) and making appropriate reversals on crossing nullclines.

We will call the completed Figure 4.5(c) a *flow-quadrant* diagram.

Especially for two-dimensional ODEs, nullclines are an invaluable aid in sketching trajectories. We shall see that this technique is most effective when used in conjunction with information from Chapter 6 about flow near equilibria. For now we turn our attention to using nullclines to construct trapping regions.

4.4.2 An Activator–Inhibitor System

In this example x and y evolve according to the ODEs

$$\begin{aligned} (a) \quad x' &= \sigma \frac{1}{1+y} \frac{x^2}{1+x^2/\kappa^2} - x, \\ (b) \quad y' &= \rho \left[\frac{x^2}{1+x^2/\kappa^2} - y \right], \end{aligned} \tag{4.30}$$

[14] Well, *generically* one expects the orientation of the flow arrows to change, but here is a cautionary example: $x' = y^2$, $y' = -x$. Along the y-nullcline (the y-axis), all arrows are oriented in the direction of increasing x. Of course, this example is concocted with nongeneric behavior in mind: Generically, polynomials change sign at every root, whereas y^2 does not.

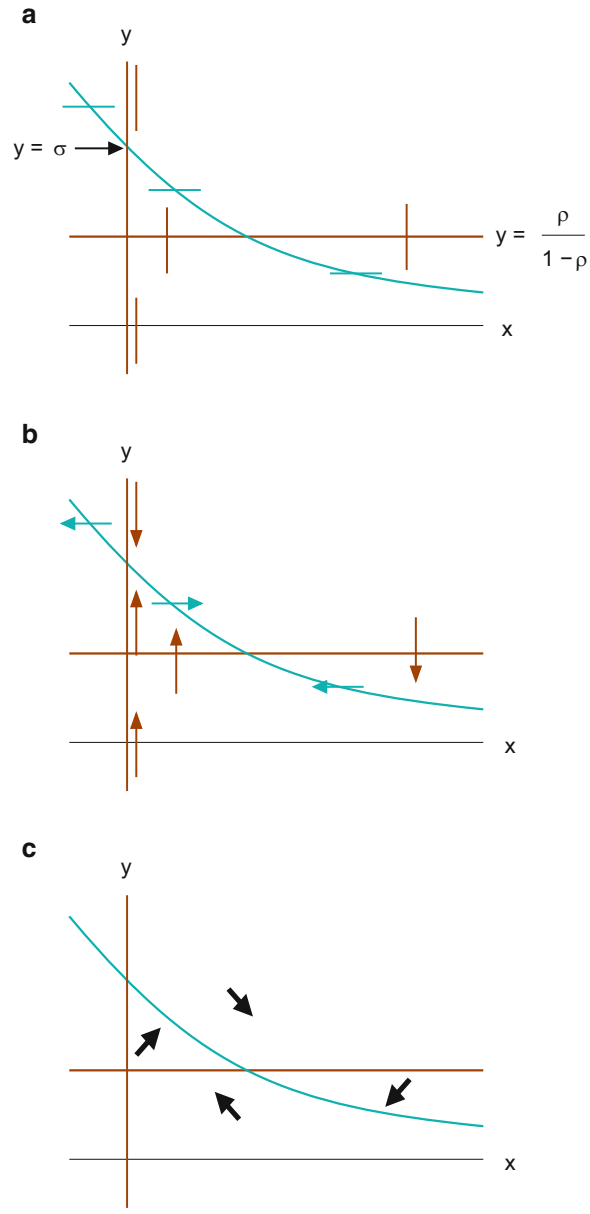

Figure 4.5: *Nullclines for the chemostat equations* (4.21) *with* $\rho = 1/2$ *and* $\sigma = 2$. *See text for a detailed description of the panels and the color conventions.*

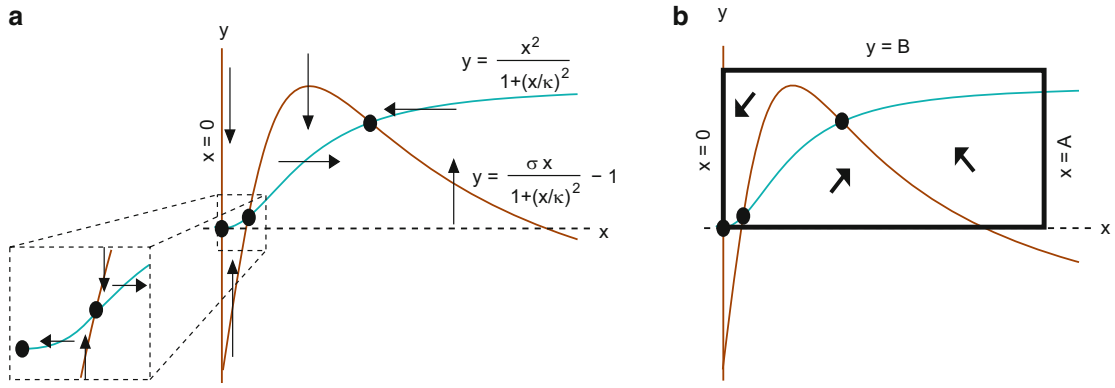

Figure 4.6: *(a) Nullclines for the activator–inhibitor equations* (4.30) *with* $\sigma = 4$ *and* $\kappa = 1$*, in which case there are three equilibria (bold dots). The inset shows a blowup of a region near the origin that contains two of the three equilibria. (b) A rectangular trapping region in the first quadrant, with flow quadrants indicated.*

where σ, ρ, κ are positive parameters; κ is often quite large. The variables x and y, which represent concentrations, are nonnegative. Linear terms in each equation describe decay. The physical basis for the nonlinear terms will be discussed in Section 5.6.

Panel (a) in Figure 4.6 shows the nullclines of (4.30) *(Check them!)*, and Panel (b) superimposes a rectangular region

$$\mathcal{K} = \{(x, y) : 0 \le x \le A,\ 0 \le y \le B\} \tag{4.31}$$

on the flow-quadrant diagram. Just from the figure, we may deduce that $\mathcal{K}$ is a trapping region. We see from Panel (a) that the flow is tangential along its left side, the y-axis, because this is an x-nullcline; and we see from the flow quadrant vectors in Panel (b) that the flow is inward along the other three sides, *provided* A and B are appropriately large. (You need to fill in the flow quadrant for the underresolved triangular region near the origin.) It is good analytical practice to determine explicit estimates for A, B to ensure that the flow is inward.

To derive global existence: given initial data $x(0) = a,\ y(0) = b$, where (a, b) lies in the first quadrant, choose A, B large enough that $\mathcal{K}$ is a trapping region and (a, b) belongs to $\mathcal{K}$, and then apply Theorem 4.2.3.

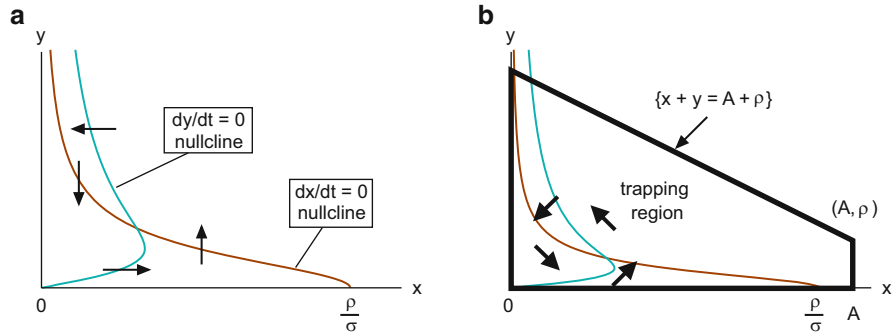

Figure 4.7: *(a) Nullclines of the Sel'kov model (4.32). (b) A trapezoidal trapping region. The diagonal line segment has slope* -1 *and connects the* y*-axis to the point* (A, ρ)*, where* $A > \rho/\sigma$*. Along this diagonal line it is not apparent from the flow quadrant diagram that the flow is inward; a calculation is needed for this.*

4.4.3 Sel'kov's Model for Glycolysis

The Sel'kov model, which will be introduced in Section 5.1.3, is given by the equations

$$\begin{aligned} x' &= \rho - \sigma x - xy^2, \\ y' &= -y + \sigma x + xy^2, \end{aligned} \tag{4.32}$$

where ρ, σ are positive constants. The variables x and y are concentrations and must be nonnegative. As in the chemostat (4.21), the sole nonlinear terms in the two equations are equal in magnitude and opposite in sign. Imitating that example, let us seek a triangular trapping region like (4.23). The flow is inward along both coordinate axes. Regarding the sloping side of (4.23), we add the equations to derive

$$x' + y' = \rho - y.$$

Unfortunately, the flow along this side is inward only if $y \geq \rho$. Thus, no region of the form (4.23) will trap solutions of (4.32).

Nullclines provide an easy fix for this minor difficulty. Nullclines for (4.32) are shown in Panel (a) of Figure 4.7, and in Panel (b) a trapezoidal region is superimposed on the flow-quadrant diagram. Regarding the right side of the trapezoid, we require that $A > \rho/\sigma$. Then, because the x-nullcline of (4.32) crosses the x-axis at $x = \rho/\sigma < A$, we see from the flow-quadrant diagram that the flow is inward along this side. We already know that the flow is inward along the other three sides. Hence every trapezoid with A sufficiently large is a trapping region, which may be used to prove global existence.

To conclude, we have found trapping regions with a combination of level sets and nullclines.

4.4.4 Van der Pol's Equation

Next we consider the van der Pol system,

$$\begin{array}{lrcl} (a) & x' & = & y, \\ (b) & y' & = & -\beta(x^2-1)y - x. \end{array} \tag{4.33}$$

Recall that the rate of change of the energy-like function $E(x,y) = (x^2+y^2)/2$ along a trajectory is given by

$$\frac{dE}{dt} = -\beta(x^2-1)y^2.$$

Thus, the flow is inward along *most* of a large (circular) level set $\{E(x,y) = E_0\}$, but not within the vertical strip $\{-1 < x < 1\}$.

To handle this difficulty, we modify a sublevel set $\{E(x,y) \le E_0\}$, as indicated in Panel (b) of Figure 4.8, by deleting slices PQRP from the top and STUS from the bottom of this disk. Taking advantage of the fact that the flow (4.33) is odd under the reflection $(x,y) \mapsto (-x,-y)$, we construct the region to be invariant under this reflection; thus, we need specify only PQR, the boundary of the upper deletion. Let QR be the line segment given by the equation

$$y = A + 2\beta x, \qquad -2 \le x \le 2, \tag{4.34}$$

where A is a large constant to be chosen below. Once A is chosen, the region is completely specified, as follows: The point R has coordinates $(2, A+4\beta)$, which determine the radius $\sqrt{2E_0}$ of the circle; i.e.,

$$2E_0 = 2^2 + (A+4\beta)^2.$$

The horizontal line starting at Q, along which $y = A - 4\beta$, meets the circle at P, which has coordinates $(-X, A-4\beta)$, where

$$(-X)^2 + (A-4\beta)^2 = 2E_0.$$

Finally, S, T, and U are located by symmetry.

We claim that provided A is sufficiently large, PQRSTUP is a trapping region. By symmetry we need to show that the flow is inward only along half of the boundary, say along PQRS. (i) Along the circular arc RS, we already know that the flow is inward. (ii) Regarding QR, by dividing the two equations in (4.33), we calculate that the flow direction has slope

$$\frac{dy}{dx} = \beta(1-x^2) - \frac{x}{y} \le \beta - \frac{x}{y}. \tag{4.35}$$

Choose A large enough that along QR, the second term on the RHS of (4.35) satisfies $|x/y| \le \beta$; then the flow direction along QR has slope $dy/dx \le 2\beta$. But QR has

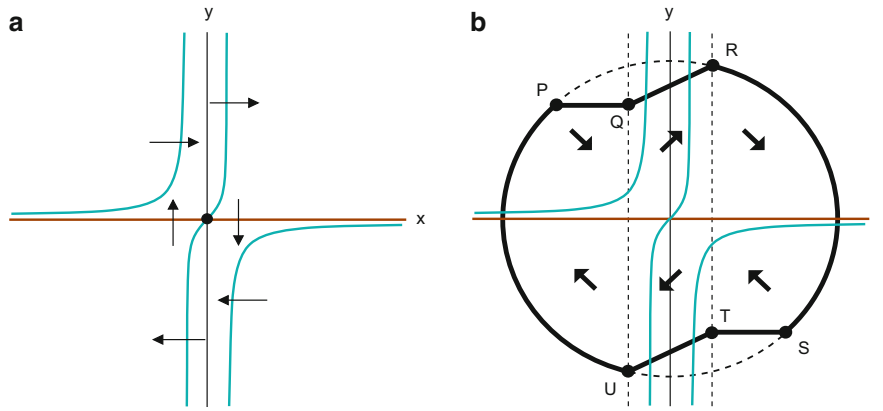

Figure 4.8: *(a) Nullclines for the van der Pol system (4.33). The y-nullclines have vertical asymptotes at $x = \pm 1$. (b) A trapping region and flow-quadrant diagram as described in the text. The dashed vertical lines are located at $x = \pm 2$.*

slope exactly 2β, so for such large A, the flow is inward along QR. (iii) Along PQ, we see from the flow-quadrant diagram Figure 4.8(b) that the flow is inward, which finishes the proof of the claim.

Using these trapping regions, we derive global existence for van der Pol's equation. As in the previous example, the successful strategy combined level sets and nullclines.

4.4.5 Michaelis–Menten Kinetics

The above examples provide an adequate introduction to nullclines and their usefulness in finding trapping regions, and more examples are given in the exercises. However, we present one more example because of its scientific interest.

Michaelis–Menten kinetics arises in modeling the concentrations (thus $x, y \geq 0$) of certain chemical species in an enzyme-mediated reaction (see Section 5.7). Applying suitable scaling, the equations take the form

$$\begin{array}{lrcl} (a) & x' & = & -x(1-y)+y, \\ (b) & \varepsilon\, y' & = & x(1-y)-(1+\kappa)y, \end{array} \tag{4.36}$$

where ε, κ are positive parameters. Typically ε is very small indeed. Reflecting this, we shall call (4.36) a *fast–slow* system, because at least away from the y-nullcline, the second variable evolves much more rapidly than x.

Global existence for (4.36) is easily demonstrated. By adding the equations, one sees that the derivative of $x + \varepsilon y$ is negative. Any triangular region bounded by the x-axis, the y-axis, and a line $\{x + \varepsilon y = A\}$, where $A > 0$, can serve as a trapping region.

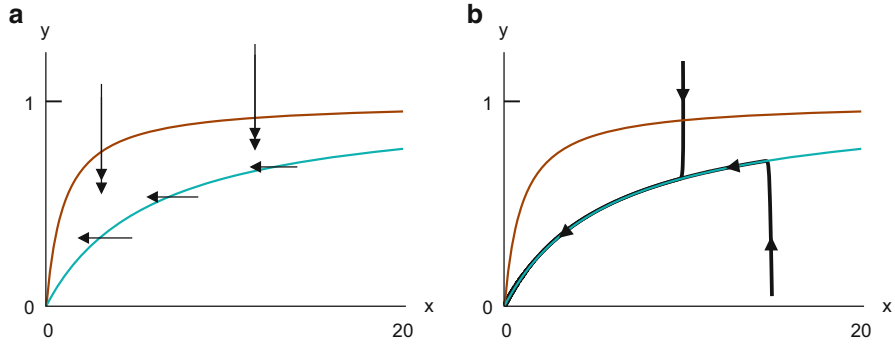

Figure 4.9: *(a) Nullclines for the scaled Michaelis–Menten equations* (4.36) *form a trapping region* ($\kappa = 5$ *in this figure). We use double arrowheads to indicate fast flow in the vertical direction (for small* ε*); if we attempted to represent lengths accurately, these vectors would be absurdly long. (b) Two trajectories for* (4.36)*, starting outside the trapping region. During a brief initial transient, motion is nearly vertical, after which trajectories hug the y-nullcline while both variables decay to zero.*

However, the region between nullclines

$$y = \frac{x}{x+1} \qquad \text{and} \qquad y = \frac{x}{x+1+\kappa}$$

is a much more interesting trapping region (see Figure 4.9(a)). The fast equation (4.36b) drives (x, y) into this trapping region, after which both variables tend slowly to zero *while staying inside the trapping region*, as sketched in Figure 4.9(b). Since ε is small, the solution hugs the y-nullcline, the lower boundary of the trapping region.

In this and other fast–slow systems, it is natural to consider the approximation[15] of setting $\varepsilon = 0$ in (4.36b). In this approximation, we may then solve (4.36b), *now an algebraic equation*, to obtain $y = x/(x+1+\kappa)$; substituting into (4.36a), we derive

$$\frac{dx}{dt} = -\frac{\kappa x}{x+1+\kappa}. \tag{4.37}$$

This equation is the (scaled) Michaelis–Menten approximation for the enzymatic reaction rate arising from (4.36).

We hope that you are worried about the violent approximation from which (4.37) is derived; indeed, the approximation changes the differential equation (4.36b) to

[15]Chemists call this approximation "letting the fast reaction go to completion" or the "quasi-steady-state" assumption. Modifying this language, we shall speak of "letting the fast equation go to equilibrium."

an algebraic equation! However, *the above argument with nullclines supports the approximation*. It indicates that after a brief initial transient, the exact solution follows the approximation rather closely.

4.5 Continuity Properties of the Solution

Your authors are fond of the "proof through pictures" style of the last two sections. We will encounter a lot more of this style of mathematics in later chapters, but for now it's back to hard analysis without the relief provided by pictures.

4.5.1 The Main Issue: Continuous Dependence on Initial Conditions

Theorem 4.5.1. *Suppose* $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$ *is locally Lipschitz, and let* $\mathbf{x}_0(t)$, $0 \le t < \beta_0$, *be a solution in forward time of* $\mathbf{x}_0' = \mathbf{F}(\mathbf{x}_0)$ *with initial condition* $\mathbf{x}_0(0) = \mathbf{b}_0$. *(i) For every positive* $T < \beta_0$, *there is a neighborhood* $\mathcal{V}$ *of* $\mathbf{b}_0$ *such that if* $\mathbf{b} \in \mathcal{V}$, *the IVP*

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0) = \mathbf{b} \tag{4.38}$$

has a solution for $0 \le t < T$. *(ii) Moreover, there is a constant* L *such that for all* $\mathbf{b} \in \mathcal{V}$,

$$|\mathbf{x}(t) - \mathbf{x}_0(t)| \le |\mathbf{b} - \mathbf{b}_0|\, e^{Lt}, \qquad 0 \le t < T. \tag{4.39}$$

In words, Conclusion (ii) asserts that if the initial data for an IVP are altered slightly, then the perturbed solution diverges from the original solution no faster than at a controlled exponential rate. Conclusion (i), which guarantees that the perturbed solution exists for nearly as long as $\mathbf{x}_0$, gives the estimate more significance.

Incidentally, $\beta_0 = \infty$ is allowed in Theorem 4.5.1, but the condition $T < \beta_0$ means that T must be finite. Such issues are clarified by examining the theorem in the context of an example, the scalar IVP

$$x' = x^3, \quad x(0) = b.$$

The solution $x_0(t) \equiv 0$ with initial condition $b = b_0 = 0$ exists for all $t \ge 0$, but for every $b \ne 0$, the IVP is solvable only for a finite interval. If T in the theorem is increased, the neighborhood $\mathcal{V}$ must be shrunk in compensation. An example in the reverse direction: it is possible for $\mathbf{x}_0$ to blow up in finite time while most nearby solutions exist for infinite times (see Exercise 4(c)).

Proof of Theorem 4.5.1. Let $\mathcal{K}_0$ be the image of $[0, T]$ under $\mathbf{x}_0$, which is a compact subset of $\mathcal{U}$. By Corollary 3.3.3, there exist a larger compact subset $\mathcal{K} \subset \mathcal{U}$ and a $\delta > 0$ such that

$$(\forall t \le T) \quad \overline{B(\mathbf{x}_0(t), \delta)} \subset \mathcal{K}. \tag{4.40}$$

By Proposition 3.3.2, $\mathbf{F}|\mathcal{K}$ is Lipschitz continuous, say with Lipschitz constant L. It is technically convenient to assume[16] $L > 0$, so that $e^{-L\varepsilon} < 1$ for every $\varepsilon > 0$.

Let $\mathcal{V}$ be the ball $B(\mathbf{b}_0, e^{-LT}\delta)$. If $\mathbf{b} \in \mathcal{V}$, let $\mathbf{x}$ be the solution in forward time of (4.38), extended to the maximal interval $0 \le t < \beta$, and define

$$t_* = \sup\{t \in [0,T] \cap [0,\beta) \,:\, (\forall s \le t)\, |\mathbf{x}(s) - \mathbf{x}_0(s)| \le \delta\}. \tag{4.41}$$

Since $|\mathbf{x}(0) - \mathbf{x}_0(0)| \le e^{-LT}\delta < \delta$, we know that $t_* > 0$. On the other hand, it follows from (4.41) that $\mathbf{x}(t)$ remains in $\mathcal{K}$ for $0 \le t < t_*$, so by Theorem 4.1.2 we have $t_* < \beta$. To derive Conclusion (i) we will show that $t_* = T$.

Let $g(t) = |\mathbf{x}(t) - \mathbf{x}_0(t)|$. From subtracting the integral equations for $\mathbf{x}$ and $\mathbf{x}_0$, we deduce that

$$g(t) \le |\mathbf{b} - \mathbf{b}_0| + \int_0^t |\mathbf{F}(\mathbf{x}(s)) - \mathbf{F}(\mathbf{x}_0(s))|\, ds. \tag{4.42}$$

If $t \le t_*$, then both $\mathbf{x}(s)$ and $\mathbf{x}_0(s)$ in the integrand belong to $\overline{B(\mathbf{x}_0(s), \delta)} \subset \mathcal{K}$, so Lipschitz continuity gives us

$$|\mathbf{F}(\mathbf{x}(s)) - \mathbf{F}(\mathbf{x}_0(s))| \le L|\mathbf{x}(s) - \mathbf{x}_0(s)|. \tag{4.43}$$

Substituting into (4.42), we have

$$g(t) \le |\mathbf{b} - \mathbf{b}_0| + L\int_0^t g(s)\, ds,$$

and hence by Gronwall's lemma,

$$g(t) \le |\mathbf{b} - \mathbf{b}_0| e^{Lt}, \qquad 0 \le t \le t_*. \tag{4.44}$$

To complete the proof we show that $t_* = T$, and thus (4.44) gives us (4.39). We see from the definition (4.41) that $t_* \le T$. But we have from (4.44) that

$$g(t_*) \le (e^{-LT}\,\delta)\, e^{Lt_*} = e^{-L(T-t_*)}\delta.$$

If t_* were strictly less than T, then we would have $g(t_*) < \delta$. By continuity, $g(t) = |\mathbf{x}(t) - \mathbf{x}_0(t)|$ would remain less than δ for some interval beyond t_*, and this would contradict the definition of t_* as a supremum. □

For use below let us formulate a refinement of this result under the stronger hypothesis that $\mathbf{F} \in \mathcal{C}^1(\mathcal{U})$. Given $\mathbf{x}_0(t)$ as in the theorem and $T < \beta_0$, choose $\delta > 0$

[16] Rigor can be such a pain! We are guarding against a triviality, since L could vanish only if $\mathbf{F}(\mathbf{x})$ were constant.

and a compact set $\mathcal{K} \subset \mathcal{U}$ for which (4.40) holds, and let

$$L = \max_{\mathbf{x}\in\mathcal{K}} \|\mathbf{DF}(\mathbf{x})\|. \tag{4.45}$$

Corollary 4.5.2. *Under these hypotheses, if* $\mathbf{b} \in B(\mathbf{b}_0, e^{-LT}\delta)$*, the IVP* (4.38) *is solvable for* $0 \le t < T + \eta$*, where* $\eta > 0$*,*

$$\mathbf{x}(t) \in \overline{B(\mathbf{x}_0(t), \delta)} \subset \mathcal{K}, \qquad 0 \le t \le T, \tag{4.46}$$

and (4.39) *holds with the constant* (4.45)*.*

Proof. A problem in adapting the proof of Theorem 4.5.1 to the present case: we cannot assume via Corollary 3.2.4 that (4.45) is a Lipschitz constant for $\mathcal{K}$, since this set is not necessarily convex. The saving point is that L need not be a Lipschitz constant for the entire set $\mathcal{K}$; it suffices if (4.43) is satisfied for all $s \in [0, T]$. By Corollary 3.2.4, given a value of s , (4.43) is satisfied for that s if

$$L \ge \max_{\mathbf{x}\in\overline{B(\mathbf{x}_0(s),\delta)}} \|\mathbf{DF}(\mathbf{x})\|,$$

and it is satisfied for all s if

$$L \ge \max_{0\le s\le T} \max_{\mathbf{x}\in\overline{B(\mathbf{x}_0(s),\delta)}} \|\mathbf{DF}(\mathbf{x})\|.$$

This inequality holds for the constant (4.45). □

Of course, the above results have analogues in backward time, which we invite you to formulate. Less trivially, the solution of the IVP for a nonautonomous equation $\mathbf{x}' = \mathbf{G}(\mathbf{x}, t)$ depends continuously on the initial condition, but we do not pursue this generalization.

4.5.2 Some Associated Formalism

Sometimes, when it is instructive to focus on how the solution of an IVP depends on the initial data, we shall use the flow notation. Specifically, we shall write

$$\boldsymbol{\varphi}(t, \mathbf{b}) = \mathbf{x}(t) \tag{4.47}$$

for the solution $\mathbf{x}(t)$ of the IVP (4.38). This *solution operator* or *flow function* is a mapping $\boldsymbol{\varphi} : \Omega \to \mathcal{U}$, where its domain is given by

$$\Omega = \{(t, \mathbf{b}) \in (-\infty, \infty) \times \mathcal{U} \ : \ t \in \text{ maximal interval of existence for (4.38)}\}. \tag{4.48}$$

It follows from Theorem 4.5.1 that $\boldsymbol{\varphi}$ is locally Lipschitz with respect to its second argument, $\mathbf{b}$, provided $\mathbf{F}(\mathbf{x})$ is locally Lipschitz. Trivially, $\boldsymbol{\varphi}$ is in fact locally Lipschitz with respect to both arguments simultaneously. *(Why is this trivial?)*

The solution operator[17] satisfies the following relation, which is known as the *semigroup property*:

Proposition 4.5.3. *If* $(s, \mathbf{b}) \in \Omega$ *and if* $(t, \boldsymbol{\varphi}(s, \mathbf{b})) \in \Omega$, *then* $(s + t, \mathbf{b}) \in \Omega$ *and*

$$\boldsymbol{\varphi}(t, \boldsymbol{\varphi}(s, \mathbf{b})) = \boldsymbol{\varphi}(s + t, \mathbf{b}). \tag{4.49}$$

This result follows easily from Lemma 3.2.9. (*Show this!*)

4.5.3 Continuity with Respect to Parameters

In Theorem 4.5.1 we proved that the solution of an IVP depends continuously on its initial data. It is also true that the solution "depends continuously on the equation." The most straightforward version of such a result addresses how the solution of a parametrized family of IVPs depends on the parameters. For simplicity, in the following theorem we address this issue only in the case of linear equations, which suffices for our needs below. Thus, suppose $A(t, \alpha_1, \ldots, \alpha_m)$ is an m-parameter family of $d \times d$ matrices, defined for $t \in (T_1, T_2)$, where this interval contains zero and, using an obvious vector notation, for $\boldsymbol{\alpha} \in \mathcal{V}$, where $\mathcal{V} \subset \mathbb{R}^m$ is open. Let $\mathbf{w}(t, \boldsymbol{\alpha})$ be the solution of

$$\mathbf{w}' = A(t, \boldsymbol{\alpha})\mathbf{w}, \qquad \mathbf{w}(0, \boldsymbol{\alpha}) = \mathbf{b}. \tag{4.50}$$

Theorem 4.5.4. *If* $A(t, \boldsymbol{\alpha})$ *is continuous on* $(T_1, T_2) \times \mathcal{V}$, *then* $\mathbf{w}(t, \boldsymbol{\alpha})$ *is continuous on this set.*

In Exercise 7 we provide hints to help you prove this result.

4.6 Differentiability Properties of the Solution

4.6.1 Dependence on Initial Conditions

In the previous section we proved that the flow $\boldsymbol{\varphi}(t, \mathbf{b})$ is Lipschitz continuous in $\mathbf{b}$. In this section we show that $\boldsymbol{\varphi}$ is in fact $\mathcal{C}^1$, provided of course that $\mathbf{F}$ is $\mathcal{C}^1$.

The above phrasing is a concise summary of the results of this section. However, let us restate the conclusion in the more discursive language of perturbation theory, since we believe that this makes the discussion more intuitive. We suppose $\mathbf{x}_0(t)$, $0 \leq t < \beta_0$, is a solution in forward time of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ with initial condition $\mathbf{x}_0(0) = \mathbf{b}_0$,

[17] A function $\boldsymbol{\varphi}$ satisfying (4.49) is sometimes called a *dynamical system*.

and we ask how the solution changes if the initial condition is perturbed. Specifically, let $\mathbf{x}(t, \varepsilon)$ be the solution of

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0, \varepsilon) = \mathbf{b}_0 + \varepsilon \mathbf{b}_1. \tag{4.51}$$

We look for an expansion[18] of this solution in powers of ε,

$$\mathbf{x}(t, \varepsilon) = \mathbf{x}_0(t) + \varepsilon \mathbf{x}_1(t) + \dots \ . \tag{4.52}$$

The size of the neglected terms, which are represented by the dots, will be estimated in Theorem 4.6.1. For the moment we proceed formally. Substituting (4.52) into (4.51), we obtain

$$\mathbf{x}_0'(t) + \varepsilon \mathbf{x}_1'(t) + \dots = \mathbf{F}(\mathbf{x}_0(t) + \varepsilon \mathbf{x}_1(t) + \dots).$$

Using a Taylor series to expand[19] the RHS of this equation in powers of ε, we calculate

$$\mathbf{x}_0'(t) + \varepsilon \mathbf{x}_1'(t) + \dots = \mathbf{F}(\mathbf{x}_0(t)) + \varepsilon \mathbf{DF}(\mathbf{x}_0(t)) \cdot \mathbf{x}_1 + \dots . \tag{4.53}$$

Equation (4.53) must hold for all values of ε, i.e., the two power series on either side of the equation define the same functions of ε. Thus the coefficients of each power of ε must be equal. Matching corresponding powers of ε in(4.53), we obtain

$$\begin{array}{lrcl} \text{(a) } \mathcal{O}(\varepsilon^0): & \mathbf{x}_0' & = & \mathbf{F}(\mathbf{x}_0), \\ \text{(b) } \mathcal{O}(\varepsilon^1): & \mathbf{x}_1' & = & \mathbf{DF}(\mathbf{x}_0(t)) \cdot \mathbf{x}_1. \end{array} \tag{4.54}$$

(The letter $\mathcal{O}$ is a mnemonic for "order.") The $\mathcal{O}(\varepsilon^0)$-equation merely repeats the ODE for our original solution. The $\mathcal{O}(\varepsilon^1)$-equation gives new information, i.e., an ODE for $\mathbf{x}_1$, which is a linear homogeneous system with time-dependent coefficients

$$\mathbf{x}_1' = A(t)\mathbf{x}_1, \tag{4.55}$$

where the coefficient matrix is given in (4.54b). Similarly, matching powers of ε in the initial conditions (4.51) gives

$$\begin{array}{lrcl} \text{(a) } \mathcal{O}(\varepsilon^0): & \mathbf{x}_0(0) & = & \mathbf{b}_0, \\ \text{(b) } \mathcal{O}(\varepsilon^1): & \mathbf{x}_1(0) & = & \mathbf{b}_1. \end{array} \tag{4.56}$$

The $\mathcal{O}(\varepsilon^0)$-equation here is nothing new, but the $\mathcal{O}(\varepsilon^1)$-equation provides an initial

[18] We may describe (4.52) as an *ansatz*, i.e., an assumed form for the solution of a problem. This term, which comes from German, is a useful one to add to your (mathematical) vocabulary.

[19] Fortunately, it suffices for our purposes to carry the expansion only through the first power. Although higher-order terms of a multivariable Taylor series can be handled efficiently with the multi-index notation (cf. Section 10.2 of [74]), it would be a distraction to introduce it here.

condition for the ODE (4.55). Theorem 3.4.1 guarantees that the IVP (4.55), (4.56b) has a unique solution $\mathbf{x}_1(t)$ for t in the same interval $[0, \beta_0)$ on which $\mathbf{x}_0$ is defined.

Here is the main result of Section 4.6. Of course analogous results hold for negative times.

Theorem 4.6.1. *Let $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$ be $\mathcal{C}^1$. In the above notation, for every $t \in [0, \beta_0)$,*

$$\lim_{\varepsilon \to 0} \frac{|\mathbf{x}(t, \varepsilon) - \mathbf{x}_0(t) - \varepsilon \mathbf{x}_1(t)|}{\varepsilon} = 0. \tag{4.57}$$

Moreover, if $T < \beta_0$, the limit is uniform for $0 \le t \le T$.

You may find the proof of this result tough going. We postpone the proof until Section 4.6.5, first exploring related ideas that among other things, make the proof easier to read.

4.6.2 The Perspective of Differentiability

Consider differentiating the flow map $\boldsymbol{\varphi} : \Omega \to \mathcal{U}$ with respect to the initial condition $\mathbf{b}$; i.e., consider

$$\frac{\partial \boldsymbol{\varphi}}{\partial b_j}(t, \mathbf{b}_0) = \lim_{\varepsilon \to 0} \frac{\boldsymbol{\varphi}(t, \mathbf{b}_0 + \varepsilon \mathbf{e}_j) - \boldsymbol{\varphi}(t, \mathbf{b}_0)}{\varepsilon}. \tag{4.58}$$

In our notation above, we have $\boldsymbol{\varphi}(t, \mathbf{b}_0) = \mathbf{x}_0(t)$, and if in (4.51) we define $\mathbf{b}_1 = \mathbf{e}_j$, then $\boldsymbol{\varphi}(t, \mathbf{b}_0 + \varepsilon \mathbf{e}_j) = \mathbf{x}(t, \varepsilon)$. Theorem 4.6.1 implies that the limit in (4.58) exists, so $\boldsymbol{\varphi}(t, \mathbf{b})$ is in fact differentiable with respect to b_j; indeed, $\partial \boldsymbol{\varphi} / \partial b_j(t, \mathbf{b}_0)$ equals the appropriate solution of the IVP (4.55), (4.56b), which we repeat as

$$\mathbf{w}_j' = A(t)\mathbf{w}_j, \qquad \mathbf{w}_j(0) = \mathbf{e}_j, \tag{4.59}$$

where $A(t) = \mathbf{DF}(\mathbf{x}_0(t))$. It is noteworthy that the partial derivatives of $\boldsymbol{\varphi}$ with respect to the various components b_j all satisfy the *same* ODE; they differ only in their initial conditions.

Let us motivate how the IVP (4.59) arises. Regarding the initial condition, at time zero, (4.58) reduces to the triviality

$$\frac{\partial \boldsymbol{\varphi}}{\partial b_j}(0, \mathbf{b}_0) = \lim_{\varepsilon \to 0} \frac{(\mathbf{b}_0 + \varepsilon \mathbf{e}_j) - (\mathbf{b}_0)}{\varepsilon} = \mathbf{e}_j. \tag{4.60}$$

For $t > 0$, the flow $\boldsymbol{\varphi}(t, \mathbf{b})$ satisfies the ODE

$$\frac{\partial}{\partial t} \boldsymbol{\varphi}(t, \mathbf{b}) = \mathbf{F}(\boldsymbol{\varphi}(t, \mathbf{b})). \tag{4.61}$$

We differentiate this equation without worrying about justification; this will be provided by the proof of Theorem 4.6.1. Specifically, take the derivative of (4.61) with

respect to b_j using the chain rule,

$$\frac{\partial}{\partial b_j}\frac{\partial \boldsymbol{\varphi}}{\partial t}(t,\mathbf{b}) = \mathbf{DF}(\boldsymbol{\varphi}(t,\mathbf{b}))\,\frac{\partial \boldsymbol{\varphi}}{\partial b_j}(t,\mathbf{b}), \tag{4.62}$$

interchange the order of the t and b derivatives to obtain

$$\frac{\partial}{\partial t}\frac{\partial \boldsymbol{\varphi}}{\partial b_j}(t,\mathbf{b}) = \mathbf{DF}(\boldsymbol{\varphi}(t,\mathbf{b}))\,\frac{\partial \boldsymbol{\varphi}}{\partial b_j}(t,\mathbf{b}),$$

and set $\mathbf{b} = \mathbf{b}_0$ to argue that $\partial\boldsymbol{\varphi}/\partial b_j(t,\mathbf{b}_0)$ should satisfy the ODE in (4.59).

4.6.3 Examples

The IVP (4.59) provides a beautiful characterization of $\partial\boldsymbol{\varphi}/\partial b_j$, which is used frequently in theoretical analysis of ODEs. Unfortunately, cases in which (4.59) can be solved explicitly are the exception rather than the rule.

One important special case in which the IVP can be solved occurs when $\mathbf{b}_0$ is an equilibrium of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$. In this case $\boldsymbol{\varphi}(t,\mathbf{b}_0) \equiv \mathbf{b}_0$, so the coefficient matrix in (4.59) is independent of time, $A = \mathbf{DF}(\mathbf{b}_0)$. In other words, the system (4.59) has constant coefficients. Incidentally, this approximation, which is known as the *linearization* of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ at the equilibrium, figures heavily in the qualitative theory of ODEs.

In case you would find it helpful to see the theorem in action, here is a less trivial example in which an explicit solution of (4.59) is possible. Consider the IVP for the Lotka–Volterra equations,

$$\begin{array}{llcll} (a) & x' & = & x - xy, & x(0) = b_1, \\ (b) & y' & = & \rho(xy - y), & y(0) = b_2. \end{array} \tag{4.63}$$

If $b_2 = 0$, then (4.63) has the explicit solution $\boldsymbol{\varphi}(t,(b_1,0)) = (b_1 e^t, 0)$; i.e., without predators, the prey grow exponentially. If a small population of predators were introduced, how would their numbers evolve? For small b_2, we have the approximation (for the predator population)

$$\varphi_2(t,(b_1,b_2)) \approx b_2 \cdot \frac{\partial \varphi_2}{\partial b_2}(t,(b_1,0)).$$

In Exercise 12, we ask you to solve the appropriate version of (4.59) to show that

$$\frac{\partial \varphi_2}{\partial b_2}(t,(b_1,0)) = \exp\left\{\rho[b_1(e^t - 1) - t]\right\}. \tag{4.64}$$

Thus, to lowest order, the number of predators grows very rapidly indeed as t increases, an exponential of an exponential. Rapid growth is hardly surprising, since

the predator's food supply is increasing without bound. Of course, the above approximation must become inaccurate as t increases. Indeed, we know from Exercise 3 in Chapter 1 that the solution of (4.63) is periodic if $b_2 > 0$, so the population must remain bounded. To interpret this discussion in the language of Theorem 4.6.1, the larger you want to make T, the smaller you must take ε in the limit (4.57).

Other examples of differentiation with respect to initial conditions, including in the above example calculating the effect of the predators on the prey, are given in the exercises.

4.6.4 The Order Notation

To prepare for the proof of Theorem 4.6.1, we introduce the order notation,[20] i.e., big-$\mathcal{O}$ and little-o. This notation makes an otherwise messy proof relatively clean.

The rigorous use of big-$\mathcal{O}$, the simpler concept, is as follows: Given a quantity that depends on a parameter, say $f(\varepsilon)$, where $0 < \varepsilon < \varepsilon_0$ and f may be either a vector or scalar quantity, we say that f is order-ε, written $f(\varepsilon) = \mathcal{O}(\varepsilon)$, if

$$(\exists C)(\exists \varepsilon_1 > 0) \text{ such that } 0 < \varepsilon < \varepsilon_1 \implies |f(\varepsilon)| \leq C\varepsilon.$$

(The formula $f(\varepsilon) = \mathcal{O}(\varepsilon)$ may also be read "f is big-$\mathcal{O}$ of ε.") The same notation is also used in several more complicated contexts. If ε can assume either sign, i.e., if $f(\varepsilon)$ is defined for $0 < |\varepsilon| < \varepsilon_0$, we write $f(\varepsilon) = \mathcal{O}(|\varepsilon|)$ to mean $|f(\varepsilon)| \leq C|\varepsilon|$, provided $|\varepsilon|$ is sufficiently small. More generally, if $\phi(\varepsilon) > 0$ is some function that tends to zero as $\varepsilon \to 0$, for example $\phi(\varepsilon) = |\varepsilon|^p$, we interpret the formula $f(\varepsilon) = \mathcal{O}(\phi(\varepsilon))$ with the obvious inequality. Also, the notation is generalized to estimate quantities that depend on multiple parameters. For example, in Theorem 4.5.1 the solution $\mathbf{x}(t) = \boldsymbol{\varphi}(t, \mathbf{b})$ depends on the d parameters of the initial data, $b_1, \ldots, b_d$, and we may paraphrase the conclusion of the theorem as

$$|\boldsymbol{\varphi}(t, \mathbf{b}) - \boldsymbol{\varphi}(t, \mathbf{b}_0)| = \mathcal{O}(|\mathbf{b} - \mathbf{b}_0|). \tag{4.65}$$

In the notation of Theorem 4.6.1, we have from Theorem 4.5.1

$$|\mathbf{x}(t, \varepsilon) - \mathbf{x}_0(t)| = \mathcal{O}(\varepsilon). \tag{4.66}$$

[20] There is an unfortunate ambiguity in the use of the symbol $\mathcal{O}$. In (4.54) and (4.56), the symbol $\mathcal{O}(\varepsilon^p)$ is merely intended as a placeholder, something to indicate the terms in a power series that are proportional to ε^p. This might be called the informal usage. By contrast, we are now going to describe a rigorous, technical meaning of this symbol. Both usages appear in the literature. Although some authors introduce separate notations to distinguish between the two meanings, we prefer to avoid this proliferation of notation. We believe that once you have been alerted to the issue, you will be able to determine from context which usage is intended.

Indeed, we say that (4.66) holds *uniformly* for $t \in [0, T]$, because the inequality

$$|\mathbf{x}(t, \varepsilon) - \mathbf{x}_0(t)| \leq e^{LT} |\mathbf{b}_1| \, \varepsilon \tag{4.67}$$

holds for all t in this interval with the same constant $C = e^{LT} |\mathbf{b}_1|$. Alternatively, we may hide the explicit constants in (4.67) and rewrite this inequality as

$$\sup_{0 \leq t \leq T} |\mathbf{x}(t, \varepsilon) - \mathbf{x}_0(t)| = \mathcal{O}(\varepsilon). \tag{4.68}$$

Little-o is a more delicate concept. If f is defined for $0 < \varepsilon < \varepsilon_0$, we say that f is little-o of ε, written $f = o(\varepsilon)$, if

$$(\forall \eta > 0)(\exists \varepsilon_1 > 0) \text{ such that } 0 < \varepsilon < \varepsilon_1 \implies |f(\varepsilon)| \leq \eta \varepsilon.$$

Of course, this definition is equivalent to[21]

$$\lim_{\varepsilon \to 0} \frac{|f(\varepsilon)|}{\varepsilon} = 0.$$

The little-o notation, like big-$\mathcal{O}$, is also used in more complicated contexts. For example, if the vector-valued function $\mathbf{F}(\mathbf{x})$ is continuously differentiable, then we may write[22]

$$\mathbf{F}(\mathbf{z}) = \mathbf{F}(\mathbf{z}_0) + \mathbf{DF}(\mathbf{z}_0) \cdot (\mathbf{z} - \mathbf{z}_0) + o(|\mathbf{z} - \mathbf{z}_0|) \tag{4.69}$$

for $\mathbf{z}$ near $\mathbf{z}_0$.

The following facts are part of the order-notation liturgy:

$$\begin{array}{l} \text{(a) If } f(\varepsilon) = o(\varepsilon), \text{ then } Cf(\varepsilon) = o(\varepsilon) \text{ for every constant } C. \\ \text{(b) If } f(\varepsilon) = o(\phi(\varepsilon)) \text{ and if } \phi(\varepsilon) = \mathcal{O}(\varepsilon), \text{ then } f(\varepsilon) = o(\varepsilon). \end{array} \tag{4.70}$$

We ask you to derive (4.70) in Exercise 10. Even though these facts may be viewed as just a reworking of familiar properties of limits, it is worth your while to *do the exercise before reading the proof of Theorem 4.6.1*. Order notation allows you to focus on higher-level issues in the proof than repeatedly rederiving properties of limits. Incidentally, (4.70) can be generalized in numerous ways, but this limited version suffices for our needs.

---

[21] Incidentally, big-$\mathcal{O}$ may be similarly characterized:

$$f(\varepsilon) = \mathcal{O}(\varepsilon) \iff \limsup_{\varepsilon \to 0} |f(\varepsilon)/\varepsilon| < \infty.$$

[22] We use $\mathbf{z}$ rather than $\mathbf{x}$ in order to reserve the latter for the functions defined by (4.54), (4.56).

To return to Theorem 4.6.1, note that our desired conclusion simply asserts that

$$\sup_{0\le t\le T} |\mathbf{x}(t,\varepsilon) - \mathbf{x}_0(t) - \varepsilon \mathbf{x}_1(t)| = o(\varepsilon). \tag{4.71}$$

4.6.5 Proof of Theorem 4.6.1

Given $T < \beta_0$, choose $\delta > 0$ and a compact set $\mathcal{K} \subset \mathcal{U}$ for which (4.40) holds, and let L be given by (4.45). For all sufficiently small ε, the perturbed initial condition $\mathbf{b}_0 + \varepsilon \mathbf{b}_1$ belongs to $B(\mathbf{b}_0, e^{-LT}\delta)$, so by Corollary 4.5.2,

$$\mathbf{x}(t,\varepsilon) \in \overline{B(\mathbf{x}_0(t), \delta)} \subset \mathcal{K}, \qquad 0 \le t \le T. \tag{4.72}$$

Forming a linear combination of the integral relations

$$\begin{aligned}
\mathbf{x}(t,\varepsilon) &= \mathbf{b}_0 + \varepsilon \mathbf{b}_1 + \int_0^t \mathbf{F}(\mathbf{x}(s,\varepsilon))\, ds, \\
\mathbf{x}_0(t) &= \mathbf{b}_0 + \int_0^t \mathbf{F}(\mathbf{x}_0(s))\, ds, \\
\mathbf{x}_1(t) &= \mathbf{b}_1 + \int_0^t A(s)\mathbf{x}_1(s)\, ds,
\end{aligned}$$

we deduce that $g(t,\varepsilon) = |\mathbf{x}(t,\varepsilon) - \mathbf{x}_0(t) - \varepsilon \mathbf{x}_1(t)|$ satisfies

$$g(t,\varepsilon) \le \int_0^t |\mathbf{F}(\mathbf{x}(s,\varepsilon)) - \mathbf{F}(\mathbf{x}_0(s)) - \varepsilon A(s)\mathbf{x}_1(s)|\, ds.$$

Add and subtract $A(s)\{\mathbf{x}(s,\varepsilon) - \mathbf{x}_0(s)\}$ in the integral and use the triangle inequality to obtain

$$g(t,\varepsilon) \le \mathcal{I}_1(t,\varepsilon) + \mathcal{I}_2(t,\varepsilon), \tag{4.73}$$

where

$$\begin{aligned}
&\text{(a) } \mathcal{I}_1(t,\varepsilon) = \int_0^t |\mathbf{F}(\mathbf{x}(s,\varepsilon)) - \mathbf{F}(\mathbf{x}_0(s)) - A(s)\{\mathbf{x}(s,\varepsilon) - \mathbf{x}_0(s)\}|\, ds, \\
&\text{(b) } \mathcal{I}_2(t,\varepsilon) = \int_0^t |A(s)\{\mathbf{x}(s,\varepsilon) - \mathbf{x}_0(s) - \varepsilon \mathbf{x}_1(s)\}|\, ds.
\end{aligned} \tag{4.74}$$

The integrand in $\mathcal{I}_2(t,\varepsilon)$ is bounded by $\|A(s)\|\, g(s,\varepsilon)$, and since $\mathbf{x}_0(s) \in \mathcal{K}$, we have $\|A(s)\| = \|\mathbf{DF}(\mathbf{x}_0(s))\| \le L$. Therefore,

$$g(t,\varepsilon) \le \mathcal{I}_1(t,\varepsilon) + L \int_0^t g(s,\varepsilon)\, ds.$$

Thus, by Gronwall's inequality,

$$g(t,\varepsilon) \le e^{Lt} \sup_{0\le s\le T} \mathcal{I}_1(s,\varepsilon),$$

and taking the supremum over t yields

$$\sup_{0\le t\le T} g(t,\varepsilon) \le e^{LT} \sup_{0\le s\le T} \mathcal{I}_1(s,\varepsilon). \tag{4.75}$$

The following lemma gives control over the RHS of (4.75). (In Exercise 25 we ask you to prove the lemma, based on showing that the pointwise estimate (4.69) is uniform over an appropriate compact set.)

Lemma 4.6.2. *As ε tends to zero,*

$$\sup_{0\le s\le T} \mathcal{I}_1(s,\varepsilon) = o(\varepsilon). \tag{4.76}$$

Applying the lemma to (4.75), we may rewrite this equation as

$$\sup_{0\le t\le T} g(t,\varepsilon) = C\, o(\varepsilon),$$

where $C = e^{LT}$. Thus, our desired conclusion (4.71) follows from (4.70a). This completes the proof.

4.6.6 Tying Up Loose Ends

We claim that the flow map $\boldsymbol{\varphi}(t,\mathbf{b})$ is $\mathcal{C}^1$. By Theorem 4.6.1, the partial derivatives $\partial\boldsymbol{\varphi}/\partial b_j$ exist, and of course $\partial\boldsymbol{\varphi}/\partial t$ also exists. We need to show that these partial derivatives are continuous. Trivially, from the ODE $\partial\boldsymbol{\varphi}/\partial t = \mathbf{F}(\boldsymbol{\varphi}(t,\mathbf{b}))$, the t-derivative is continuous. The derivative with respect to b_j was characterized by the IVP (4.59), which we rewrite indicating the dependence of the coefficient matrix on $\mathbf{b}$:

$$\frac{\partial}{\partial t}\frac{\partial\boldsymbol{\varphi}}{\partial b_j} = \mathbf{DF}(\boldsymbol{\varphi}(t,\mathbf{b}))\frac{\partial\boldsymbol{\varphi}}{\partial b_j}, \qquad \frac{\partial\boldsymbol{\varphi}}{\partial b_j}(0) = \mathbf{e}_j. \tag{4.77}$$

It follows from Theorem 4.5.4 that $\partial\boldsymbol{\varphi}/\partial b_j$ is continuous. This proves the claim.

In Section 4.6.2 we motivated the formula (4.59) for $\partial\boldsymbol{\varphi}/\partial b_j$, modulo an interchange of the order of differentiation that needs to be justified. According to Theorem B.3.2 in Appendix B, it suffices to show that one of the mixed partials exists and is continuous. It follows from (4.77) that $\partial^2\boldsymbol{\varphi}/\partial t\partial b_j$ is continuous, the RHS of this ODE being a product of continuous functions.

4.6.7 Generalizations

There are more results about the differentiability of the solution of IVPs than either you or we care to explore fully, but a few highlights need to be mentioned. We refrain

from giving formal statements of these results in the hopes of making the text more readable. A careful treatment of such results is given in Section 1.7 of [15].

First let us generalize Theorem 4.6.1 to nonautonomous IVPs. It can be shown that the flow operator $\boldsymbol{\varphi}(t, t_0, \mathbf{b})$ obtained by solving

$$\mathbf{x}' = \mathbf{G}(\mathbf{x}, t), \qquad \mathbf{x}(t_0) = \mathbf{b} \tag{4.78}$$

is $\mathcal{C}^1$ with respect to all variables, provided $\mathbf{G}$ is $\mathcal{C}^1$. Moreover, $\partial\boldsymbol{\varphi}/\partial b_j(t, t_0, \mathbf{b})$ satisfies a linear homogeneous IVP

$$\frac{d\mathbf{w}}{dt} = \mathbf{DG}(\boldsymbol{\varphi}(t, t_0, \mathbf{b}), t)\mathbf{w}, \qquad \mathbf{w}(t_0) = \mathbf{e}_j, \tag{4.79}$$

where $\mathbf{DG}$ denotes the $d \times d$ matrix of partial derivatives $\partial\mathbf{G}/\partial x_j$, not including the t derivative.

Next, we consider "differentiability with respect to the equation" through a parametrized family of IVPs, say

$$\mathbf{x}' = \mathbf{G}(\mathbf{x}, \boldsymbol{\alpha}), \qquad \mathbf{x}(0) = \mathbf{b}, \tag{4.80}$$

where $\boldsymbol{\alpha} \in \mathcal{V} \subset \mathbb{R}^m$; to simplify the notation, we assume that (4.80) is autonomous. The flow map $\boldsymbol{\varphi}(t, \mathbf{b}, \boldsymbol{\alpha})$ is $\mathcal{C}^1$ with respect to all its arguments, and $\partial\boldsymbol{\varphi}/\partial\alpha_j$ satisfies the linear *inhomogeneous* IVP

$$\frac{d\mathbf{w}}{dt} = \mathbf{DG}(\boldsymbol{\varphi}(t, \mathbf{b}, \boldsymbol{\alpha}), \boldsymbol{\alpha})\mathbf{w} + \frac{\partial\mathbf{G}}{\partial\alpha_j}(\boldsymbol{\varphi}(t, \mathbf{b}, \boldsymbol{\alpha}), \boldsymbol{\alpha}), \quad \mathbf{w}(0) = \mathbf{0}. \tag{4.81}$$

As above, $\mathbf{DG}$ denotes the $d \times d$ matrix of derivatives of $\mathbf{G}$ with respect to x_j; the derivative with respect to the parameter is written out explicitly. In fact, this result does not require separate proof; it may be derived easily with a "reduce it to the previous case" ruse, as we discuss in Exercise 24.

Finally, our third generalization addresses higher-order derivatives: it can be shown that if an ODE is of class $\mathcal{C}^k$, then the solution operator also has k continuous derivatives with respect to all variables. In particular, under the hypotheses of Theorem 4.6.1, if $\mathbf{F} \in \mathcal{C}^2$, then

$$\mathbf{x}(t, \varepsilon) - \mathbf{x}_0(t) - \varepsilon\mathbf{x}_1(t) = \mathcal{O}(\varepsilon^2).$$

4.7 Exercises

After the core exercises, there are subsections on applying the differentiation results, on cleaning up some loose ends from previous chapters, and on computing.

4.7.1 Core Exercises

The primary purposes of the core exercises are as follows:

Proof of a minor generalization	1
Unfinished business	2, 5–7, 10
Use of trapping regions to prove global existence	3
Counterexamples to clarify the theory	4
A first look at asymptotic behavior	8, 9

1. (a) Prove the following variant of Theorem 4.1.2: If $\mathbf{F}$ is bounded on $\mathcal{U}$ and if $\beta_* < \infty$, then $\mathbf{x}_*(t)$ tends to a point on $\partial\mathcal{U}$ as $t \to \beta_*$.

   (b) Construct an IVP for a scalar ODE $x' = f(x)$, with f bounded, whose solution has a maximal interval of existence (α_*, β_*) with both endpoints finite.

2. Prove Theorem 4.2.3.

   Hint: You must show that the solution cannot leave $\mathcal{K}$. Rule out crossing $\partial\mathcal{K}$ at a regular point by the same argument used to prove Theorem 4.2.2. At a corner point, say $\mathbf{P}$, recall the notation of Section B.3.3 to represent $\partial\mathcal{K}$ near $\mathbf{P}$ as the intersection of two smooth curves, $\{\phi_1 = 0\}$ and $\{\phi_2 = 0\}$. Take limits of regular points to conclude that

   $$\langle \nabla\phi_k(\mathbf{P}), \mathbf{F}(\mathbf{P})\rangle \geq 0, \qquad k = 1, 2.$$

   Use the hypothesis that $\nabla\phi_1(\mathbf{P})$ and $\nabla\phi_2(\mathbf{P})$ are linearly independent to argue that one of these inequalities must be strict, as in (4.12). Rule out crossing $\partial\mathcal{K}$ at $\mathbf{P}$ by the simple argument used to derive Theorem 4.2.2 from (4.12).

3. Use trapping regions to analyze global existence for the following equations:

   (a) The Lorenz equations:

   $$\begin{aligned} x' &= \sigma(y - x), \\ y' &= \rho x - y - xz, \\ z' &= -\beta z + xy, \end{aligned}$$

   where σ, ρ, β are positive constants (arbitrary initial conditions).

   Hint: Deduce global existence by showing that a set of the form

   $$\mathcal{K} = \{(x, y, z) : x^2 + y^2 + (z - \rho - \sigma)^2 \leq A^2\}$$

   is a trapping region if A is sufficiently large; i.e., the trapping region is bounded by the level set of a carefully chosen quadratic function.

   (b) Equations for the evolution of two interacting species:

   $$x' = x\,(1 - x - by)\,, \qquad y' = \rho y\,(1 - y - cx)\,,$$

   where ρ, b, c are constants with $\rho > 0$ (both $x(0)$, $y(0)$ nonnegative).

Discussion: Note that each species has logistic growth that is modified by the presence of the other. If b, c are both positive, then the interaction is competitive; if they are both negative, it is symbiotic. The most interesting cases, for which there are coexistence equilibria in the (open) first quadrant, are (i) $1 < b, c$, (ii) $0 < b, c < 1$, and (iii) $-1 < b, c < 0$.

Hint: In Cases (i) and (ii), draw the nullclines to show that there are triangular trapping regions of the form (4.23). A slightly more complicated region is needed for Case (iii).

Discussion: Incidentally, we invite you to compute a few typical solutions for an equation belonging to Case (i) and for one belonging to Case (ii). You will find that they behave differently as $t \to \infty$. This behavior may seem a little mysterious at present, but the ideas from Chapter 6 will clarify it.

Challenge: A harder, but educational, problem is to prove that the solution may blow up in finite time when $b, c < -1$. *Try it!* As preparation for this harder problem, you might first try to show blowup for the simpler system $x' = xy,\ y' = xy$.

(c) A simplified activator–inhibitor system

$$\begin{aligned} (a) \quad x' &= \sigma \tfrac{1}{1+y} x^2 - x, \\ (b) \quad y' &= \rho\,[x^2 - y], \end{aligned} \tag{4.82}$$

i.e., the limit of equations (4.30) as $\kappa \to \infty$ (both $x(0),\ y(0)$ nonnegative).

Remark: It was straightforward to construct a trapping region for (4.30); you'll have to work considerably harder to do so for (4.82). But below we will use the simpler equations (4.82) in calculations about the properties of solutions of activator–inhibitor models.

(d) The "repressilator"

$$\begin{aligned} x' &= \frac{\mu}{1+y^n} - x, \\ y' &= \frac{\mu}{1+z^n} - y, \\ z' &= \frac{\mu}{1+x^n} - z, \end{aligned}$$

where μ and n are positive constants (initial values of all variables nonnegative).

Hint: To prove global existence for this system, which will be introduced and analyzed in Section 8.7.2, construct a trapping region of the form

$$\mathcal{K} = \{(x, y, z) : x \geq 0,\ y \geq 0,\ z \geq 0,\ x + y + z \leq A\};$$

i.e., prove that the flow is inward on each of the four faces of this simplex if A is sufficiently large. Note that $\mathcal{K}$ is a three-dimensional region whose boundary is only piecewise smooth. (We have shied away from actually defining piecewise smooth in higher dimensions, but however this phrase is defined, it surely applies to this set.) Hence Theorem 4.2.2 is not immediately applicable to proving global existence. If you are bothered by this gap, you can close it by mimicking Problem 2.

4. (a) Regarding Theorem 4.1.2, give an example to show that if $\beta_* = \infty$, then the solution of an IVP can stay inside a compact set for all time.

   (b) Give an example of a $\mathcal{C}^1$ function that satisfies the linear-growth estimate (4.4) but is not (globally) Lipschitz.

   (c) For the system

$$\begin{aligned} x' &= x^2 - y^2, \\ y' &= 2xy, \end{aligned} \tag{4.83}$$

   find a solution that blows up in finite time but most nearby solutions exist for all time.

   Hint: First show that with initial conditions $x(0) = 1, y(0) = 0$, the solution of (4.83) blows up in finite time. Although you can't solve (4.83) for other trajectories, you can locate the solution curves, i.e., find the orbits, as follows. Along an orbit you have the ODE

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{2xy}{x^2 - y^2}.$$

   Multiply the equation by $(x/y)^2 - 1$ and manipulate the result into the form[23]

$$\frac{d}{dx}\left(\frac{x^2}{y} + y\right) = 0,$$

   from which you may deduce that the solution curves are circles through the origin, $x^2 + (y - C)^2 = C^2$, where C is an arbitrary constant. Argue from this information that the solution of (4.83) with initial conditions $x(0) = 1, y(0) = b$, where $b \neq 0$, exists for all time.

5. (a) Prove the following generalization of Theorem 4.2.1, referring to Exercise 8(a) in Chapter 3.

   Theorem 4.7.1. *If $\mathbf{G} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ is locally Lipschitz in $\mathbf{x}$ and t and if there exist nonnegative continuous functions $B(t), K(t)$ such that*

$$|\mathbf{G}(\mathbf{x}, t)| \leq K(t)|\mathbf{x}| + B(t), \qquad (\mathbf{x}, t) \in \mathbb{R}^d \times \mathbb{R}, \tag{4.84}$$

   then the solution of the IVP

$$\mathbf{x}' = \mathbf{G}(\mathbf{x}, t), \qquad \mathbf{x}(0) = \mathbf{b}$$

[23] A scalar ODE of the form $(d/dx)f(x, y) = 0$ is called *exact*. In this example we have made the equation exact, and therefore solvable, by means of an *integrating factor*. This is another solution technique, one that we did not cover in Section 1.3; to learn more, see Section 1.9 of [10].

exists for all time, $-\infty < t < \infty$. Moreover, for every finite $T > 0$,

$$|\mathbf{x}(t)| \le |\mathbf{b}|e^{K_{\max}|t|} + \frac{B_{\max}}{K_{\max}}(e^{K_{\max}|t|} - 1), \qquad -T \le t \le T, \tag{4.85}$$

where

$$B_{\max} = \max_{|s|\le T} |B(s)|, \qquad K_{\max} = \max_{|s|\le T} |K(s)|.$$

Discussion: It will be useful below to have Theorem 4.7.1 explicitly formulated, and it is useful training to adapt the proof of Theorem 4.2.1 to handle the nonautonomous case. For a complete proof, you would need to prove extensions of Proposition 4.1.1 and Theorem 4.1.2 to nonautonomous equations. We invite you to skip this not very rewarding task and regard the extensions of those two results as given.

Incidentally, every linear equation $\mathbf{x}' = A(t)\mathbf{x} + \mathbf{g}(t)$ satisfies the above hypotheses, provided of course that $A(t)$ and $\mathbf{g}(t)$ are continuous. Thus, this theorem provides another proof of global existence for linear equations, an alternative to the approach based on Picard iteration that was outlined in Exercise 16 in Chapter 3.

(b) Use Theorem 4.7.1 to prove global existence for

$$\begin{aligned} x' &= y, \\ y' &= -(1/4 + \beta\cos t)x. \end{aligned}$$

Remark: Recall your computations from Exercise 13 in Chapter 3, in which you found the surprising fact that solutions of this equation may grow exponentially with time; from this problem you may conclude that solutions grow no faster than exponentially.

6. Extend Theorem 4.2.2 to ODEs on the cylinder $\mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}$.

Hint: First, some notation: If $\mathcal{K}$ is a subset of $\mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}$, let $\mathcal{K}_{\text{lift}} \subset \mathbb{R}^2$ be defined by[24]

$$\mathcal{K}_{\text{lift}} = \{(x, y) \in \mathbb{R}^2 : \Pi \cdot (x, y) \in \mathcal{K}\},$$

where $\Pi : \mathbb{R}^2 \to \mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}$ is the natural projection. In less formal terms, the projection "wraps the plane around the cylinder," and $\mathcal{K}_{\text{lift}}$ is the set obtained from $\mathcal{K}$ if the cylinder is "unwrapped." (Cf. Figure 4.4.)

By an ODE on $\mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}$ we mean a two-dimensional ODE

$$\begin{bmatrix} \theta' \\ y' \end{bmatrix} = \mathbf{F}(\theta, y), \tag{4.86}$$

where $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ is 2π-periodic in its first argument, i.e., $\mathbf{F}(\theta+2\pi, y) = \mathbf{F}(\theta, y)$. We propose to prove global existence for the IVP for (4.86) by applying the theory

[24]The subscript "lift" refers to the idea that $\mathbb{R}^2$ is a simply connected covering space for $\mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}$ that lies "above" $\mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}$.

of this chapter to the equivalent "lifted" planar IVP

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \mathbf{F}(x,y), \quad \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} = \mathbf{b}, \tag{4.87}$$

where x may vary over $(-\infty, \infty)$.

Suppose $\mathcal{K} \subset \mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}$ is a compact trapping region for (4.86); then $\mathcal{K}_{\text{lift}}$ is a trapping region for (4.87). By Corollary 4.2.4, if $\mathbf{b} \in \mathcal{K}_{\text{lift}}$, then the solution of (4.87) remains in $\mathcal{K}_{\text{lift}}$ for as long as it exists, say $0 \le t < t_*$. From compactness we have

$$C = \max_{(x,y)\in\mathcal{K}_{\text{lift}}} |\mathbf{F}(x,y)| = \max_{(\theta,y)\in\mathcal{K}} |\mathbf{F}(\theta,y)| < \infty.$$

Therefore, $|\mathbf{x}(t)| \le |\mathbf{b}| + Ct$ for $0 \le t < t_*$. If t_* were finite, this estimate would contradict Theorem 4.1.2, so (4.87) must have a global solution.

Remark: A similar argument works for ODEs on the torus $\mathbb{T}^2$.

7. Prove Theorem 4.5.4.

Hint: You need to show that $|\mathbf{w}(t_1, \boldsymbol{\alpha}_1) - \mathbf{w}(t_2, \boldsymbol{\alpha}_2)|$ can be made arbitrarily small by making t_1, t_2 and $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ sufficiently close to one another. It suffices to consider only positive times $t_k \le T$, where $T < T_2$, and to restrict $\boldsymbol{\alpha}_k$ to a compact subset $\mathcal{N}$ of the parameter set $\mathcal{V}$. Let

$$M = \max_{0\le t\le T} \max_{\boldsymbol{\alpha}\in\mathcal{N}} \|A(t,\boldsymbol{\alpha})\|. \tag{4.88}$$

It is convenient to abbreviate $\mathbf{w}(t, \boldsymbol{\alpha}_k)$ to $\mathbf{w}_k(t)$; i.e., you need to show that $|\mathbf{w}_1(t_1) - \mathbf{w}_2(t_2)|$ is small. Add and subtract a term $\mathbf{w}_2(t_1)$ to estimate

$$|\mathbf{w}_1(t_1) - \mathbf{w}_2(t_2)| \le |\mathbf{w}_1(t_1) - \mathbf{w}_2(t_1)| + |\mathbf{w}_2(t_1) - \mathbf{w}_2(t_2)|. \tag{4.89}$$

To estimate the second term here, first apply Gronwall's inequality to the ODE $\mathbf{w}_2' = A(t, \boldsymbol{\alpha}_2)\mathbf{w}_2$ to conclude that $|\mathbf{w}_2(t)| \le |\mathbf{b}|e^{Mt}$ and then integrate this ODE to obtain

$$|\mathbf{w}_2(t_1) - \mathbf{w}_2(t_2)| \le M|\mathbf{b}|e^{MT}|t_1 - t_2|.$$

To estimate the first term in (4.89), subtract the ODEs for $\mathbf{w}_k(t)$ and then add and subtract $A(t, \boldsymbol{\alpha}_1)\mathbf{w}_2(t)$ to deduce that

$$\left|\frac{d}{dt}(\mathbf{w}_1 - \mathbf{w}_2)\right| \le \|A(t,\boldsymbol{\alpha}_1)\|\,|\mathbf{w}_1 - \mathbf{w}_2| + \|A(t,\boldsymbol{\alpha}_1) - A(t,\boldsymbol{\alpha}_2)\|\,|\mathbf{w}_2|.$$

Integrate this inequality and apply the generalization of Gronwall's inequality in Exercise 3.8(a) to conclude that

$$|(\mathbf{w}_1 - \mathbf{w}_2)(t)| \le B\frac{e^{Mt} - 1}{M},$$

where M is defined by (4.88) and

$$B = \max_{0\le t\le T} |\mathbf{b}|e^{MT}\,\|A(t,\boldsymbol{\alpha}_1) - A(t,\boldsymbol{\alpha}_2)\|.$$

Combine these inequalities with the fact that $A(t, \boldsymbol{\alpha})$ is uniformly continuous in $\boldsymbol{\alpha}$ to complete your proof.

Remark: Theorem 4.5.4 may be generalized to a (Lipschitz continuous) nonautonomous nonlinear equation, say $\mathbf{x}' = \mathbf{G}(\mathbf{x}, t, \boldsymbol{\alpha})$. The most convenient proof of such a result uses the trick proposed in Exercise 24.

8. For the chemostat equations (4.21) with $0 < \sigma < \rho/(1-\rho)$, use nullclines to show that every solution with initial conditions in the first quadrant tends to the equilibrium $(0, \sigma)$ as $t \to \infty$.

   Hint: Recall that a monotone function has a limit as $t \to \infty$.

   Remark: This problem and the next address an issue that figures heavily in the second half of this book, i.e., the asymptotic behavior of solutions of an ODE. In these problems the issue can be resolved with ad hoc methods. Starting in Chapter 6, we introduce more effective tools for investigating such asymptotic behavior.

9. (a) Prove the following lemma.

   Lemma 4.7.2. *Suppose $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$ is continuous. If the solution $\mathbf{x}(t)$ of the ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ tends to a point $\mathbf{b}_* \in \mathcal{U}$ as $t \to \infty$, then $\mathbf{b}_*$ is an equilibrium of this equation.*

   Hint: Use the ODE to show that $\mathbf{x}'(t)$ has a limit as $t \to \infty$ and then argue that $\lim \mathbf{x}'(t) = \mathbf{0}$. Incidentally, satisfying an ODE is an essential part of this exercise; Exercise 3 in Appendix B concerns an example of a $\mathcal{C}^1$ bounded, monotone increasing function $g(t)$ (which must have a limit as $t \to \infty$) whose derivative does not converge.

   (b) For the chemostat equations (4.21) as shown in Figure 4.5, i.e., with $\sigma > \rho/(1-\rho) > 0$, show that every solution with initial conditions in the open first quadrant tends to the equilibrium $(\sigma - \rho/(1-\rho), \rho/(1-\rho))$ as $t \to \infty$.

   Hint: Let $x(t), y(t)$ be a solution of (4.21) with $x(0), y(0) > 0$. By solving (4.22), deduce that $x(t)+y(t)$ tends to σ as $t \to \infty$. Now refer to Figure 4.10, in which the first quadrant is divided into six regions by the nullclines and the line $\{x + y = \sigma\}$. First suppose that $(x(0), y(0)) \in R_1 \cup R_2 \cup R_3$. Argue that the sets R_3, $R_2 \cup R_3$, and $R_1 \cup R_2 \cup R_3$ are all trapping regions. Thus, one of the following must hold:

$$\begin{array}{ll} (\exists T) \text{ s.t. } (\forall t \in (T, \infty)) & (x(t), y(t)) \in R_3, \\ (\exists T) \text{ s.t. } (\forall t \in (T, \infty)) & (x(t), y(t)) \in R_2, \\ (\forall t \in (0, \infty)) & (x(t), y(t)) \in R_1. \end{array} \tag{4.90}$$

   In each of the cases, both components of $(x(t), y(t))$ are monotonic functions for large t. *(Why?)* Therefore, they have limits; i.e., $(x(t), y(t))$ tends to some point $\mathbf{b}_*$ in the first quadrant. By the lemma, $\mathbf{b}_*$ must be an equilibrium, and there is only one equilibrium.

   If $(x(0), y(0)) \in R_4 \cup R_5 \cup R_6$, you may truncate these sets as in Section 4.3.2,

$$\tilde{R}_k = R_k \cap \{(x, y) : x + y \le A\}, \qquad k = 4, 5, 6,$$

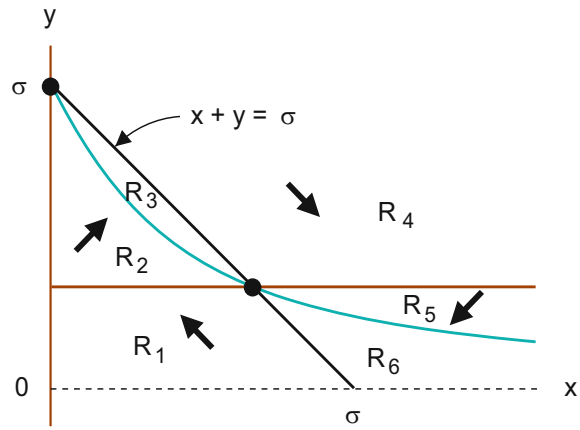

Figure 4.10: *Illustrating the regions described in Exercise 9(b).*

for an appropriate constant A and proceed similarly. Truncation is needed to show that no trajectories escape to infinity.

10. Prove the two claims regarding the order notation made in (4.70).

4.7.2 Applying the Differentiation Theorems

11. *Introduction:* This exercise is intended as a confidence builder; you verify formulas (4.79) and (4.81) in a specific example by explicit solutions. It is also useful preparation for some calculations below.

    (a) Use separability to solve the IVP

    $$\frac{dx}{dt} = \frac{\alpha + \cos t}{x}, \quad x(0) = b, \tag{4.91}$$

    where $b \neq 0$.

    (b) Differentiate your solution to compute $\partial x/\partial b$.

    (c) Define $G(x, t, \alpha) = (\alpha + \cos t)/x$ as in (4.91), calculate DG, write out the IVP (4.79), and show that your answer to Part (b) satisfies this IVP.

    (d) Differentiate your solution to Part (a) to compute $\partial x/\partial \alpha$.

    (e) Write out the IVP (4.81) and show that your answer to Part (d) satisfies this IVP.

    Remark: It might be more consistent to use the flow notation in this problem, but when explicit calculations are involved, we usually find it more intuitive to use x instead of φ. Note that what we are calling x depends on t, b, and α, but we are deliberately sloppy about not indicating all these dependencies explicitly.

12. For the Lotka–Volterra system (4.63), verify formula (4.64) for $\partial\varphi_2/\partial b_2(t,(b_1,0))$.

    Hint: Recall that $\boldsymbol{\varphi}(t,(b_1,0)) = (b_1e^t,0)$. Thus $\partial\boldsymbol{\varphi}/\partial b_j$ satisfies an ODE $\mathbf{w}' = A(t)\mathbf{w}$, where

    $$A(t) = \mathbf{DF}(\boldsymbol{\varphi}(t,(b_1,0))) = \begin{bmatrix} 1 & -b_1e^t \\ 0 & \rho(b_1e^t-1) \end{bmatrix}. \tag{4.92}$$

    As a warmup exercise, calculate from the explicit solution that $\partial\varphi_1/\partial b_1(t,(b_1,0)) = e^t$ and check that $\mathbf{w}(t) = (e^t,0)$ satisfies this ODE with initial condition $\mathbf{w}(0) = (1,0)$. Then solve the ODE with $\mathbf{w}(0) = (0,1)$ and derive (4.64). As a bonus, you may also calculate $\partial\varphi_1/\partial b_2(t,(b_1,0))$, from which you can estimate the effect of a small number of predators on the prey.

13. *Introduction:* Let $x_0(t,\mu), y_0(t,\mu)$ be the solution of the IVP for the torqued pendulum (4.24) subject to initial conditions $x(0) = 0$, $y(0) = 0$, and similarly let $x_\pi(t,\mu), y_\pi(t,\mu)$ be the solution with $x(0) = \pi$, $y(0) = 0$. If $\mu = 0$, these two initial conditions are equilibria for (4.24), so we have

    $$\begin{bmatrix} x_0(t,0) \\ y_0(t,0) \end{bmatrix} \equiv \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} x_\pi(t,0) \\ y_\pi(t,0) \end{bmatrix} \equiv \begin{bmatrix} \pi \\ 0 \end{bmatrix}.$$

    Calculate the partial derivatives of $x_0(t,\mu), y_0(t,\mu)$ and $x_\pi(t,\mu), y_\pi(t,\mu)$ with respect to μ at $\mu = 0$; i.e., determine the ODE (4.81) that these functions of t satisfy and solve the appropriate IVP.

    Food for thought: Solutions of (4.81) for the derivative of x_0, y_0 involve only decaying exponentials, while solutions of (4.81) for the derivative of x_π, y_π may include a growing exponential. How does this different behavior relate to differences between the equilibria $x = 0$ and $x = \pi$?

14. *Introduction:* Let $\boldsymbol{\varphi}(t,\mathbf{b})$ be the solution of the IVP

    $$\begin{aligned} x' &= x - y - (x^2+y^2)x, & x(0) &= b_1, \\ y' &= x + y - (x^2+y^2)y, & y(0) &= b_2. \end{aligned} \tag{4.93}$$

    With the particular initial condition $\mathbf{b}_* = (1,0)$, the IVP has the solution $\boldsymbol{\varphi}(t,\mathbf{b}_*) = (\cos t, \sin t)$, which is periodic. In this exercise we invoke polar coordinates to find the solution of the ODE (4.59) for $\partial\boldsymbol{\varphi}/\partial b_j(t,\mathbf{b}_*)$. The primary challenge of the exercise is calculational; to rephrase this more positively, the exercise offers useful practice with calculations in multivariable calculus.

    (a) Write down the ODE (4.59) for $\partial\boldsymbol{\varphi}/\partial b_j(t,\mathbf{b}_*)$.

    Discussion: Your equation will be a linear system with variable coefficients, and it's not clear how to solve this equation. Let's exploit the fact that (4.93) becomes much simpler if it is written in polar coordinates:

    $$r' = r - r^3, \qquad \theta' = 1. \tag{4.94}$$

Let $\boldsymbol{\varphi}_{\text{polar}}(t, \mathbf{c})$ be the solution of (4.94) with initial conditions $r(0) = c_1$, $\theta(0) = c_2$. (For clarity we'll write $\boldsymbol{\varphi}_{\text{cart}}$—"cart" for Cartesian—for the solution of (4.93).)

We introduce the notation

$$\boldsymbol{\Psi} : (0, \infty) \times \mathbb{R} \to \mathbb{R}^2, \qquad \boldsymbol{\Psi}(r, \theta) = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix}$$

for the transformation from polar to Cartesian coordinates. Note that the inverse $\boldsymbol{\Psi}^{-1}$ is multivalued, but we may define it uniquely in a neighborhood of $\mathbf{b}_*$, with $\boldsymbol{\Psi}^{-1}(\mathbf{b}_*) = (1, 0) = \mathbf{b}_*$. Then for $\mathbf{b}$ near $\mathbf{b}_*$ we have the representation

$$\boldsymbol{\varphi}_{\text{cart}}(t, \mathbf{b}) = \boldsymbol{\Psi} \circ \boldsymbol{\varphi}_{\text{polar}}(t, \boldsymbol{\Psi}^{-1}(\mathbf{b})). \tag{4.95}$$

This equation is valid for all t, even though the orbit circles the origin multiple times. We will differentiate (4.95) with the chain rule. The formulas will be simpler with a bit of notation: let $\mathbf{D}\boldsymbol{\varphi}_{\text{cart}}$ be the 2×2 matrix with columns $\partial\boldsymbol{\varphi}_{\text{cart}}/\partial b_j$ and define $\mathbf{D}\boldsymbol{\varphi}_{\text{polar}}$ similarly with columns $\partial\boldsymbol{\varphi}_{\text{polar}}/\partial c_j$.

(b) Show that

$$\mathbf{D}\boldsymbol{\varphi}_{\text{cart}}(t, \mathbf{b}_*) = \mathbf{D}\boldsymbol{\Psi}(\boldsymbol{\varphi}_{\text{polar}}(t, \mathbf{b}_*)) \cdot \mathbf{D}\boldsymbol{\varphi}_{\text{polar}}(t, \mathbf{b}_*), \tag{4.96}$$

where of course

$$\mathbf{D}\boldsymbol{\Psi}(r, \theta) = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}.$$

Remark: From the chain rule you would expect a third factor on the right in (4.96) from differentiation of $\boldsymbol{\Psi}^{-1}$, but $\mathbf{D}\boldsymbol{\Psi}^{-1}(\mathbf{b}_*)$ equals the identity matrix.

(c) Using the fact that $\boldsymbol{\varphi}_{\text{polar}}(t, \mathbf{b}_*) = (1, t)$, apply (4.59) to show that

$$\mathbf{D}\boldsymbol{\varphi}_{\text{polar}}(t, \mathbf{b}_*) = e^{tA}, \qquad \text{where } A = \begin{bmatrix} -2 & 0 \\ 0 & 0 \end{bmatrix}.$$

(d) Calculate $\mathbf{D}\boldsymbol{\varphi}_{\text{cart}}(t, \mathbf{b}_*)$ from (4.96).

(e) Verify that the columns of $\mathbf{D}\boldsymbol{\varphi}_{\text{cart}}(t, \mathbf{b}_*)$, i.e., $\partial\boldsymbol{\varphi}_{\text{cart}}/\partial b_j$, satisfy your equation in Part (a).

4.7.3 Some Mopping-Up Exercises

15. Prove the comparison result from Chapter 1, Theorem 1.8.1.

Hint: Make an autonomous system out of the ODE for x in the theorem,

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} f(x_1, x_2) \\ 1 \end{bmatrix}.$$

Use the function $u(t)$ to define a set

$$\mathcal{K} = \{(x_1, x_2) \in \mathbb{R}^2 : u(x_2) \leq x_1\}.$$

Show that $\mathcal{K}$ is a trapping region for this system.

16. In the context of Corollary 3.2.8, strengthen that result by showing that the IVP is solvable on $(-\eta, \eta)$, provided merely that $\eta < \delta/K$.

Hint: Imitate the proof of Theorem 4.2.1.

17. Rewrite the equation $\varepsilon x'' + x' + x = 0$ as a first-order system, draw the nullclines, and make a flow-quadrant diagram, observing the double-arrow convention (cf. Figure 4.9) for the fast-flow direction. Argue from your figure that after a brief transient, a typical trajectory hugs the y-nullcline as it decays to zero.

Discussion: In Exercise 1.14, working with explicit solutions, you showed that the approximation of setting $\varepsilon = 0$ in this equation gives decent results. Using nullclines you can understand geometrically why the approximation works. For a more complete understanding of these issues, repeat the exercise for $\varepsilon x'' - x' - x = 0$.

4.7.4 Computing Exercise

18. In a programming language of your choosing, apply Euler's method (which is introduced in the appendix of this chapter) to solve approximately the IVP for (4.93) with $\mathbf{b} = (0.1, 0)$, say for $0 \leq t \leq 0.5$. Choose various mesh sizes $h = 10^{-n/2}$, $n = 2, 3, \ldots, 10$. Compare your calculations with the exact solution by making a log-log plot of the errors in $x(0.5)$ and $y(0.5)$ as a function of h over this range.

Remark: You will find that, as suggested by Theorem 4.9.1, the error in Euler's method is roughly proportional to h as the mesh size tends to zero.

The simplest improvement over Euler's method is the unimaginatively named "improved Euler method." In this method also, one computes approximations $\mathbf{y}_n$ to $\mathbf{x}(nh)$, the solution of an IVP at integer multiples of the step size. In the improved Euler method, advancing to the next approximation is a two-step process: given $\mathbf{y}_n$, let

$$\text{(a) } \mathbf{y}_{n+1/2} = \mathbf{y}_n + (h/2)\mathbf{F}(\mathbf{y}_n), \qquad \text{(b) } \mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{F}(\mathbf{y}_{n+1/2}). \tag{4.97}$$

The only change from Euler's method is that $\mathbf{y}_{n+1/2}$, the crude initial estimate for the solution at the intermediate time $(n + 1/2)h$, is used in (4.97b). (After substitution into (4.97b), $\mathbf{y}_{n+1/2}$ is discarded.) Remarkably, the simple fudge (4.97) gives a more accurate method; as $h \to 0$, the error tends to zero as h^2. We invite you to verify this claim on the above example with your own computations.

4.7.5 PHD Exercises

19. In the context of Theorem 4.2.2, show that if a compact trapping region $\mathcal{K}$ satisfies (4.12), the distance from the solution $\mathbf{x}(t)$ to $\partial\mathcal{K}$ remains bounded away from zero as $t \to \infty$.

    Advice: You might as well assume that $\partial\mathcal{K}$ is smooth; handling the more general case of a piecewise smooth boundary would add only technical complications.

20. Prove global existence for Duffing's equation if the sign of friction is reversed,

$$\begin{aligned} x' &= y, \\ y' &= +\beta y + x - x^3, \end{aligned}$$

    where $\beta > 0$.

    Hint: You can't get anywhere on this problem using trapping regions, because the solution grows without bound. Likewise, Theorem 4.2.1 is useless, because the cubic term in the force violates (4.4). Here's a strategy that works: Calculate the rate at which energy grows. Manipulate your result to show that $dE/dt \leq \beta E$, provided (x, y) lies outside some large circle, say $x^2 + y^2 > C^2$. (In fact, $C = 2$ is sufficient.) Then extract your conclusion from this information.

21. Use the order notation to make a completely rigorous proof out of your heuristic proof from Exercise 9 of Chapter 3 that (in the notation of that problem)

$$\phi'(t) - [\mathrm{tr}A(t)]\,\phi = 0.$$

22. *Introduction:* The equations

$$\begin{aligned} x' &= x - xy, \\ y' &= \rho(\sigma + xy - y), \end{aligned} \tag{4.98}$$

    where ρ and σ are positive parameters, are like the Lotka–Volterra equations, except that even without predation, *new predators appear at a small background rate* (normalized to $\rho\sigma$). This assumption is pretty hokey when applied to foxes and rabbits, but variants of these equations arise in certain models for the evolution of a viral infection; see [62] for details. In this application, x measures the total population of virus cells in the patient's body, while y represents the body's immune mechanism; specifically, y measures the population of what are called *effector* cells.

    (a) Interpret in words each term in these equations.

    (b) Draw nullclines for these equations.

    (c) Use the nullclines to construct trapping regions to prove global existence in forward time for the IVP, assuming initial data in the first quadrant.

    Remarks: Setting $\sigma = 0$ in (4.98) yields the Lotka–Volterra equations (4.63). Recall that all orbits of (4.63) are periodic. Thus, the only trapping regions for

Lotka–Volterra are regions bounded by the periodic orbits themselves. While we were able to find explicit formulas for the orbits for the Lotka–Volterra equations, we can't do likewise for (4.98).

Hint: Try to construct a trapping region for (4.98) bounded in part by an orbit of the Lotka–Volterra equations, or at least something close to an orbit. Tolerances are small, and you have to be careful in carrying out the construction. If you find it helpful, assume that σ is small.

23. (a) Use software to solve the IVP

$$\begin{array}{lrcll} \text{(a)} & x' & = & y, & x(0)=2, \\ \text{(b)} & y' & = & -x-zy, & y(0)=0, \\ \text{(c)} & \varepsilon z' & = & -\left(z-x^2+1\right), & z(0)=1, \end{array} \tag{4.99}$$

for $t \in [0,5]$, say for $\varepsilon = 10^{-3}, 10^{-6}, 10^{-9}$.

Warning: Depending on what method you choose, you may encounter trouble for very small ε.

(b) Show that the fast–slow approximation of setting $\varepsilon = 0$ reduces the three-dimensional problem to the van der Pol system.

(c) Compare your solution in Part (a) (in cases in which you were able to get a solution) with solutions of the van der Pol equation.

Remark: In the language of Section 10.3, (4.99) is a *stiff* ODE.

24. *Introduction:* The differentiability of the solution of the IVP (4.80) with respect to the parameters can be proved with minimal effort using a trick of augmenting the system with m "fake" variables corresponding to the parameters $\alpha_1, \ldots, \alpha_m$. Specifically consider an auxiliary variable $\mathbf{y} \in \mathbb{R}^m$, and let $(\mathbf{x}, \mathbf{y})$ evolve according to the system

$$\begin{array}{rcll} \mathbf{x}' & = & \mathbf{G}(\mathbf{x},\mathbf{y}), & \mathbf{x}(0)=\mathbf{b}, \\ \mathbf{y}' & = & \mathbf{0}, & \mathbf{y}(0)=\boldsymbol{\alpha}. \end{array}$$

Apply Theorem 4.6.1 to this system to show that the solution of (4.80) is continuously differentiable with respect to α and to derive (4.81).

25. (a) *Introduction:* If $\mathbf{F} \in \mathcal{C}^1(\mathcal{U})$, let $\Delta(\mathbf{z}, \mathbf{z}_0) = \mathbf{F}(\mathbf{z}) - \mathbf{F}(\mathbf{z}_0) - \mathbf{DF}(\mathbf{z}_0) \cdot (\mathbf{z} - \mathbf{z}_0)$ be the error in the first-order Taylor series approximation for $\mathbf{F}(\mathbf{z})$ based at $\mathbf{z}_0$. In this notation, (4.69) may be rephrased as

$$\Delta(\mathbf{z}, \mathbf{z}_0) = o(|\mathbf{z} - \mathbf{z}_0|).$$

Given a compact set $\mathcal{K} \subset \mathcal{U}$, show that the above estimate is uniform for $\mathbf{z}_0 \in \mathcal{K}$; i.e., show that for every $\eta > 0$, there is a $\delta > 0$ such that for all $\mathbf{z}_0 \in \mathcal{K}$ and all $\mathbf{z}$ with $|\mathbf{z} - \mathbf{z}_0| < \delta$,

$$|\Delta(\mathbf{z}, \mathbf{z}_0)| \leq \eta |\mathbf{z} - \mathbf{z}_0|. \tag{4.100}$$

Hint: By Corollary 3.3.3, there exist a larger compact set $\mathcal{K}' \subset \mathcal{U}$ and a δ_0 such that for every $\mathbf{z}_0 \in \mathcal{K}$, the ball $\overline{B(\mathbf{z}_0, \delta_0)}$ is contained in $\mathcal{K}'$. Apply calculus to conclude that if $|\mathbf{z} - \mathbf{z}_0| < \delta_0$, then

$$\Delta(\mathbf{z}, \mathbf{z}_0) = \left\{ \int_0^1 \left[\mathbf{DF}(\mathbf{z}_0 + s(\mathbf{z} - \mathbf{z}_0)) - \mathbf{DF}(\mathbf{z}_0) \right] ds \right\} \cdot (\mathbf{z} - \mathbf{z}_0).$$

Use the fact that **DF** is uniformly continuous on $\mathcal{K}'$ to complete the argument.

(b) Prove Lemma 4.6.2.

Hint: Since $A(s) = \mathbf{DF}(\mathbf{x}_0(s))$, the definition (4.74) of $\mathcal{I}_1(t, \varepsilon)$ may be rewritten as

$$\mathcal{I}_1(t, \varepsilon) = \int_0^t \Delta(\mathbf{x}(s, \varepsilon), \mathbf{x}_0(s)) ds.$$

Apply Part (a) with $\mathcal{K}$ equal to the image of $[0, T]$ under the base solution $\mathbf{x}_0$, $\mathbf{z} = \mathbf{x}(s, \varepsilon)$, and $\mathbf{z}_0 = \mathbf{x}_0(s)$; specifically, show that

$$\sup_{0 \le s \le T} |\Delta(\mathbf{x}(s, \varepsilon), \mathbf{x}_0(s))| = o(\phi(\varepsilon)),$$

where

$$\phi(\varepsilon) = \sup_{0 \le s \le T} |\mathbf{x}(s, \varepsilon) - \mathbf{x}_0(s)|. \tag{4.101}$$

Show that $\phi(\varepsilon) = \mathcal{O}(\varepsilon)$ and invoke (4.70) to finish the proof.

4.8 Pearls of Wisdom

While Theorem 4.5.1 gives control over the solution of an IVP for every finite time, infinite times are beyond our reach: the limit of $\boldsymbol{\varphi}(t, \mathbf{b})$ as $t \to \infty$ may be discontinuous in $\mathbf{b}$. For example, the solution of the IVP for Duffing's equation

$$\begin{aligned} x' &= y, & x(0) &= b, \\ y' &= -\beta y + x - x^3, & y(0) &= 0, \end{aligned}$$

converges to $(1, 0)$ as $t \to \infty$ if $b > 0$ and to $(-1, 0)$ if $b < 0$, at least provided that $|b|$ is not too large. (You could probably prove this now; in any case, you will see it proved in Chapter 6.)

As we observed in Section 4.6.3 (but it bears repeating), if $\mathbf{b}_0$ is an equilibrium of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, we may approximate $\boldsymbol{\varphi}(t, \mathbf{b}_0 + \varepsilon \mathbf{b}_1)$, the solution of an IVP with initial conditions near $\mathbf{b}_0$, by $\mathbf{b}_0 + \varepsilon \mathbf{w}(t)$, where $\mathbf{w}(t)$ solves the linear constant-coefficient IVP

$$\mathbf{w}' = A\mathbf{w}, \qquad \mathbf{w}(0) = \mathbf{b}_1$$

with the coefficient matrix $A = \mathbf{DF}(\mathbf{b}_0)$. This approximation, known as the *linearization* of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ at the equilibrium, typically determines the qualitative behavior of solutions of the full nonlinear equation near the equilibrium. (Cf. Chapter 6.)

There is an efficient procedure for handling the linearized equations (4.59) in numerical solutions. Given a d-dimensional autonomous ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, consider the greatly enlarged system of dimension $d + d^2$ with unknowns the d-dimensional vector $\mathbf{x}(t)$ and a $d \times d$ matrix $X(t)$:

$$\begin{array}{rcll} \mathbf{x}' & = & \mathbf{F}(\mathbf{x}), & \mathbf{x}(0) = \mathbf{b}, \\ X' & = & \mathbf{DF}(\mathbf{x})X, & X(0) = I, \end{array} \tag{4.102}$$

where I is the identity matrix. Then the jth column of $X(t)$ reproduces (4.59); thus, the jth column of $X(t)$ equals $\partial\boldsymbol{\varphi}/\partial b_j(t, \mathbf{b})$. To conclude: although in discussing the theory, it is more natural to first solve the ODE and then differentiate with respect to initial conditions, in computations it works better to attack both issues at the same time.

4.9 Appendix: Euler's Method

4.9.1 Introduction

Since it is rarely possible to produce explicit solutions of ODEs, we often resort to *numerical methods*[25] in order to obtain approximate solutions. In this section we introduce the simplest numerical method, known as *Euler's method*. We do not propose to actually use this method as a practical source of information about solutions of ODEs, since software employs methods that are far more accurate than this, and their automated control of step size makes them a joy to use. Rather, we study Euler's method for cultural reasons, namely, it provides useful insight into numerical methods in general, and its simplicity allows the conceptual issues to come through more easily.

Euler's method is an iterative process for approximating the solution of an IVP, say

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0) = \mathbf{b}. \tag{4.103}$$

For simplicity, let's consider a set of evenly spaced t-values: given $h > 0$, we calculate the approximations $\mathbf{y}_n$ for $\mathbf{x}(nh)$ recursively according to the rule

$$\mathbf{y}_0 = \mathbf{b}; \qquad \mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{F}(\mathbf{y}_n), \quad n = 0, 1, \ldots . \tag{4.104}$$

Although $\mathbf{y}_n$ depends on h, here we follow the usual convention of not indicating this dependence.

As an illustration, consider the time-honored scalar IVP $x' = x$, $x(0) = 1$, which has exact solution $x(t) = e^t$. Let's approximate the solution on the interval $t \in [0, 1]$

[25] Perturbation methods, some of which are discussed in Sections 7.5 and 7.6 (among other places), offer another valuable way of approximating solutions.

with Euler's method, say using a step size of $h = 1/N$, where N is a positive integer. Starting from $y_0 = 1$, we use (4.104) to generate the subsequent iterates recursively:

$$y_{n+1} = y_n + hy_n = \left(1 + \frac{1}{N}\right) y_n, \qquad n = 0, 1, 2, \ldots,$$

so $y_n = (1 + 1/N)^n$. To test the accuracy of the approximation, consider y_n with $n = N$, which should approximate x at time $t = Nh = 1$: as $h \to 0$ (and thus $N \to \infty$), we have

$$y_N = (1 + 1/N)^N \to e = x(1),$$

as desired.

More generally, y_n provides an approximation for e^t for all t, but the formulation of this behavior is made awkward by two issues: (i) the number of steps needed to reach time t scales up as $1/h$ as $h \to 0$ and (ii) any specific time t need not belong to the set of grid points $\{nh : n = 0, 1, \ldots\}$ for which the approximations are computed. Thus, on the time interval $0 \le t \le T$, the convergence result for this example takes the somewhat clumsy form

$$\lim_{h \to 0} \max_{0 \le n \le T/h} |e^{nh} - y_n| = 0.$$

A convergence result for the general case is given in Theorem 4.9.1 below.

4.9.2 Theoretical Basis for the Approximation

We offer three motivations[26] for Euler's method. All three motivations begin with the limited goal of understanding the first step in (4.104), which we may rephrase as

$$x(h) \approx x(0) + hF(x(0)). \tag{4.105}$$

(For simplicity, we assume temporarily that we are solving a scalar equation.)

Motivation 1: *(Tangent line)* Interpreting the derivative geometrically (see Figure 4.11), we see from the ODE that the slope of the solution curve through $(0, x(0))$ equals $F(x(0))$. Thus, we may estimate $x(h)$ by following the tangent line, resulting in the approximation (4.105).

Motivation 2: *(Finite differences)* Using the difference quotient approximation

$$\frac{x(h) - x(0)}{h} \approx x'(0) = F(x(0)),$$

[26]For Euler's method, all three motivations produce the same formula, but for advanced numerical methods, different approximation formulas may result from starting with one or another of these three points of view.

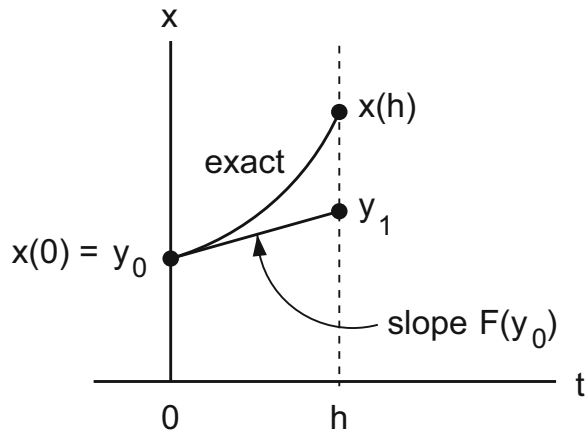

Figure 4.11: *Schematic illustration of one iteration of Euler's method. Note the discrepancy between the exact solution $x(h)$ and its Euler's method approximation y_1.*

we again obtain (4.105).

Motivation 3: *(Integral equation)* Reformulating the IVP as an integral equation,

$$x(h) \;=\; x(0) + \int_0^h F(x(s))\, ds,$$

we derive (4.105) from a one-term Riemann-sum approximation for the integral.

The continuation of Euler's method may seem like an act of desperation. It is extremely unlikely that the point (h, y_1) will lie on the exact solution curve (see Figure 4.11). Nevertheless, it is the best information we have about the solution. Therefore, we will use that point as the starting point for another iteration of Euler's method; i.e., we let y_2 equal the Euler approximation to the solution of $x' = F(x)$ through the point (h, y_1). All subsequent steps are derived similarly. One may well wonder about an approximation in which each step is based on increasingly faulty information, especially since as $h \to 0$, more and more steps are required to advance a finite time. However, as we show in the next subsection, the accumulated error in the numerical solution actually tends to zero with h.

4.9.3 Convergence of the Numerical Solution

If $\mathbf{F}$ is defined everywhere, then the definition (4.104) of $\mathbf{y}_n$ remains meaningful for arbitrarily large n, even if the solution $\mathbf{x}$ that is being approximated blows up in finite time. However, if $\mathbf{F}$ is defined only on a subset $\mathcal{U} \subset \mathbb{R}^d$, then some iterate $\mathbf{y}_n$ may lie outside $\mathcal{U}$, so the iteration would halt. This possibility is addressed in Conclusion (i) of the following theorem, which has much in common with Theorem 4.5.1.

Theorem 4.9.1. *Suppose $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$ is locally Lipschitz, and let $\mathbf{x}(t)$, $0 \le t < \beta$, be a solution in forward time of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ with initial condition $\mathbf{x}(0) = \mathbf{b}$. (i) For every positive $T < \beta$, there exists a positive constant h_0 such that if $h < h_0$, then the iterates $\mathbf{y}_n$ are defined for all n such that $nh \le T$. (ii) There are constants C, L such that if $h < h_0$, then*

$$|\mathbf{x}(nh) - \mathbf{y}_n| \le Che^{Lnh}, \qquad \text{for } 0 \le nh \le T.$$

Remark: Note that Conclusion (ii) implies the uniform error estimate

$$|\mathbf{x}(nh) - \mathbf{y}_n| \le Che^{LT}.$$

Proof. Choose a compact subset $\mathcal{K} \subset \mathcal{U}$ and a constant $\delta > 0$ such that

$$(\forall t \in [0, T]) \ \ \overline{B(\mathbf{x}(t), \delta)} \subset \mathcal{K} \subset \mathcal{U}.$$

We define the constants: let $C = \max_{\mathcal{K}} |\mathbf{F}|$, let L be a Lipschitz constant for $\mathbf{F}|\mathcal{K}$, and let $h_0 = e^{-LT}\delta/C$. We compute $\mathbf{y}_n$ for as many iterations as $nh \le T$ and $\mathbf{y}_n \in \mathcal{K}$, say $n \le N$. Note that $\mathbf{y}_N \in \mathcal{K}$, so that it is possible to calculate at least one more iterate, $\mathbf{y}_{N+1}$.

The solution $\mathbf{x}$ satisfies the integral equation

$$\mathbf{x}(t) = \mathbf{b} + \int_0^t \mathbf{F}(\mathbf{x}(s))\, ds.$$

In order to derive an analogous integral equation for the approximate solution, we construct a piecewise constant function on $[0, (N+1)h]$ as follows: for $t < (N+1)h$, let

$$\mathbf{y}^{(h)}(t) = \mathbf{y}_n \quad \text{for} \quad nh \le t < (n+1)h,$$

and at the right-hand endpoint define $\mathbf{y}^{(h)}((N+1)h) = \mathbf{y}_{N+1}$. Note that

$$\int_{nh}^{(n+1)h} \mathbf{F}(\mathbf{y}^{(h)}(s))\, ds = h\mathbf{F}(\mathbf{y}_n),$$

the integrand being constant. Therefore, at the grid points,

$$\mathbf{y}^{(h)}(nh) = \mathbf{b} + \int_0^{nh} \mathbf{F}(\mathbf{y}^{(h)}(s))\, ds\,, \qquad n = 0, 1, 2, \ldots, N+1.$$

Moving off grid points, we obtain the desired integral equation

$$\mathbf{y}^{(h)}(t) = \mathbf{b} + \int_0^t \mathbf{F}(\mathbf{y}^{(h)}(s))\, ds - \int_{nh}^t \mathbf{F}(\mathbf{y}^{(h)}(s))\, ds,$$

where n is the largest integer such that $nh \le t$.

For $0 \le t \le \min\{T, (N+1)h\}$, let $g(t) = |\mathbf{x}(t) - \mathbf{y}^{(h)}(t)|$. Subtracting the integral equations for $\mathbf{x}(t)$ and $\mathbf{y}^{(h)}(t)$, we deduce that

$$g(t) \le \int_0^t |\mathbf{F}(\mathbf{x}(s)) - \mathbf{F}(\mathbf{y}^{(h)}(s))|\, ds + \int_{nh}^t |\mathbf{F}(\mathbf{y}^{(h)}(s))|\, ds,$$

where again n is the largest integer such that $nh \le t$. Note that in the integrands, we have $\mathbf{x}(s)$, $\mathbf{y}^{(h)}(s) \in \mathcal{K}$. By Lipschitz continuity, the first term here satisfies

$$\int_0^t |\mathbf{F}(\mathbf{x}(s)) - \mathbf{F}(\mathbf{y}^{(h)}(s))|\, ds \le L \int_0^t g(s) ds,$$

and by the definition of C, the second satisfies

$$\int_{nh}^t |\mathbf{F}(\mathbf{y}^{(h)}(s))|\, ds \le Ch.$$

Thus, by Gronwall's inequality (extended to piecewise continuous functions),

$$g(t) \le Che^{Lt}, \qquad 0 \le t \le \min\{T, (N+1)h\}. \tag{4.106}$$

Regarding Conclusion (i): If $(N+1)h < T$, then taking $t = (N+1)h$ in (4.106), we see that

$$|\mathbf{y}_{N+1} - \mathbf{x}((N+1)h)| \le Che^{L(N+1)h} < Che^{LT} < \delta,$$

so $\mathbf{y}_{N+1} \in \mathcal{K}$; i.e, the iteration would continue. Inequality (4.106) verifies Conclusion (ii). □

By (4.106), the errors produced by Euler's method may be expected to be on the order of h to the first power. This apparently good news is actually bad news. Contrast this error estimate with, for example, the error in the fourth-order Runge–Kutta algorithm (RK4), which is of order h^4. Thus, if the step size is halved, the error in Euler's method is merely halved, while the error in the RK4 method[27] is decreased by a factor of 16. To achieve high accuracy with Euler's method, the step size must be chosen painfully small; this wastes computational time and also raises round-off issues [65].

Much effort has gone into devising highly accurate numerical methods for solving ODEs. In Exercise 18, we describe one simple improvement over Euler's method, but for serious further study see the cautionary examples in Section 10.3 and the references in that section.

[27] Typically, in software h is chosen automatically, so the convergence rate is not readily apparent to the user.

Chapter 5

Nondimensionalization and Scaling

The preceding chapter had some pretty heavy analysis, and the next has even more. In what may be welcome relief, the present chapter pushes in an orthogonal direction: it focuses on *nondimensionalization* and *scaling*, which are techniques for simplifying ODEs that arise in applications. In the hands of a skilled user, they can provide insights far beyond any reasonable expectation. This power is illustrated by the following anecdote about G.I. Taylor, a distinguished twentieth-century British physicist/applied mathematician. Using this kind of analysis, he estimated the power of one of the early tests of the atomic bomb [86], based on *only a series of photographs from the cover of a popular magazine of the fireball following the explosion*! His estimate was so accurate that it led to suspicions about a possible security leak! (Cf. Chapter 14 of [6].)

In Section 5.1 we briefly introduce the types of applications we consider in this text. In Sections 5.2–5.7 we apply the scaling methodology to a variety of ODEs. (For brevity, we shorten "nondimensionalization and scaling" to "scaling.") Some general comments and advice on the use of scaling are collected in the Pearls.

The only effective way to learn scaling is through specific examples. For that reason, in this chapter we have applied the technique to quite a few different ODEs. It's possible that you will grow weary of an uninterrupted diet of such examples. If so, here is a suggestion: at a bare minimum read through the end of Section 5.3, read Section 5.6, and read Section 5.9.1; then return to read other sections as need or interest dictates. However, we add one comment: these ideas are both elegant and powerful; don't give them short shrift.

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8_5

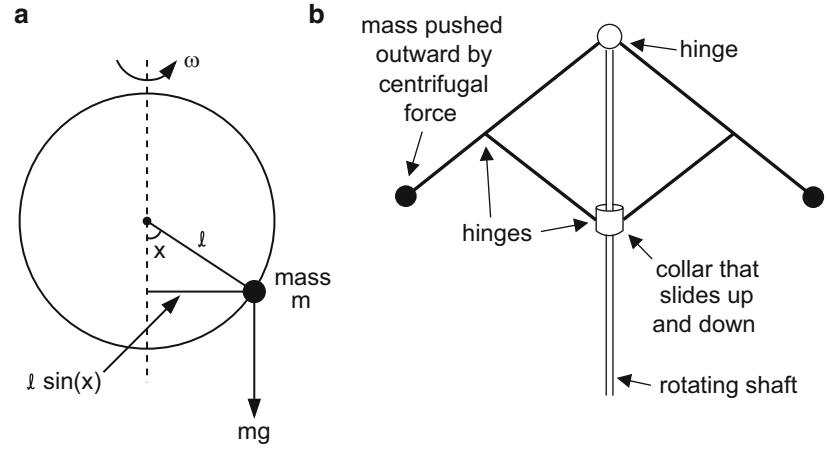

Figure 5.1: *(a) Schematic of a bead on a rotating wire hoop, described by* (5.1)*. (b) A less contrived, mathematically equivalent, system: one of the components of Watt's centrifugal governor for steam engines. (See Meiss [54] pp. 159–160, Exercise 3, for more information.)*

5.1 Classes of ODEs in Applications

The applications in this text may be grouped into three broad classes: mechanical, electrical, and "bathtub" models. Our coverage is slanted toward the physical and biological sciences;[1] the social sciences, even economics, are not broached.

5.1.1 Mechanical Models

Mechanical systems that we encountered in Chapter 1 include spring–mass systems, the pendulum, and Duffing's equation. The infamous Lorenz system, introduced in Exercise 3a of Chapter 4, also has a basis in mechanics. Of course, the whole field of ODEs got started with the mechanics of planetary motion.

Usually, the ODE for the motion of a mechanical system can be derived from Newton's second law: mass times acceleration equals the sum of the forces. Let's illustrate this by deriving the equations of motion of a bead sliding on a wire hoop that is rotating about a vertical axis, as illustrated in Figure 5.1(a). Let m be the mass of the bead, ℓ the radius of the hoop, and ω the (constant) speed of rotation. The motion of the bead is purely tangential; if x measures its angular position (in radians) as a function of time, then the tangential acceleration is the radius ℓ times the second derivative[2] $d^2x/d\hat{t}^2$. There are three forces acting on the bead:

[1]Even with this restriction, we are forced to omit many interesting problems; just to cite one example, reference [96] studies ODEs derived from neural models that do not belong to any of our three categories.

[2]Regarding the notation $d^2x/d\hat{t}^2$: At this point, we introduce the convention of putting hats over variables—but not parameters—that have dimensions. Time has dimensions, but since x is

(i) gravity, whose projection onto the tangential direction is $-mg\sin x$; (ii) friction, which we model (with the usual oversimplification) as proportional to $dx/d\hat{t}$; and (iii) centrifugal force from rotation about the vertical axis. Centrifugal force from rotation at speed ω in a circle of radius $\ell\sin x$ equals $m(\ell\sin x)\omega^2$; projection of this force, which is directed outward, onto the tangential direction adds a factor of $\cos x$. Thus, Newton's law gives us

$$m\ell\frac{d^2x}{d\hat{t}^2} = -b\frac{dx}{d\hat{t}} - mg\sin x + m\ell\omega^2\sin x\cos x. \tag{5.1}$$

If $\omega = 0$, then (5.1) reduces to the pendulum equation, (1.26) with damping added.

In Exercise 5 we ask you to scale this equation, which *reduces the number of parameters from five to two.*

5.1.2 Electrical Models

Van der Pol's equation (1.33) arose in modeling an electrical circuit. Although we presented the torqued-pendulum equation (4.24) in Chapter 4 as a mechanical system, this equation also describes an electrical device, the Josephson junction (see Strogatz [81], Section 4.6). Interesting ODEs with an electrical basis also occur in models for nerve cells, as in the FitzHugh–Nagumo equations introduced in Exercise 11 and the Morris–Lecar model in Section 9.8. The Chua circuit [53] provides another equation of electrical origin; although this equation exhibits fascinating behavior, we do not study it in this book.

Equations for circuit models may be derived from Kirchhoff's laws plus information about the current–voltage characteristics of the various devices in the circuit. For biologically based models, to understand current–voltage characteristics, one needs to discuss physiological issues as well. We do not attempt to cover any of this background, which can be found online.

5.1.3 "Bathtub" Models

We borrow the phrase *bathtub model* from Ellner and Guckenheimer [20]. This tongue-in-cheek phrase describes models in which a population (of animals, cells, nutrients, biomolecules etc.) is divided into various categories (bathtubs), and the ODEs track the flow of individuals (water) from one category to another. Two examples that fit into this category are the various versions of the Lotka–Volterra equations in Section 1.6, in which the populations consist of animals, and the activator–inhibitor equations (4.30), in which the populations consist of biomolecules. Strictly

measured in radians, it does not. Although this may seem obscure at the moment, we hope that the discussion in Section 5.2 will clarify the reasons for this convention. In the meantime, we suggest that you simply ignore the hats.

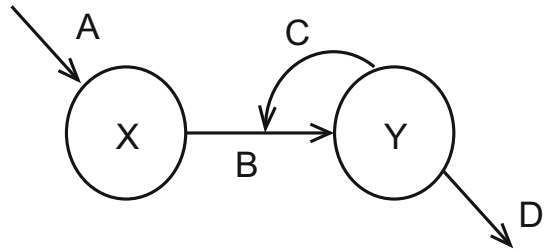

Figure 5.2: *Schematic illustration of the Sel'kov glycolysis model* (5.2).

speaking, the populations in such models are usually integers, but if populations are large, an integer variable may be approximated by a continuous variable.

Biologists often use a schematic diagram to enumerate the terms in bathtub models that arise in their field. We illustrate this in Figure 5.2 for Sel'kov's toy model [71] for the positive feedback in the first step of glycolysis (the metabolic process of converting sugar into energy; see [47], Section 1.5, for more explanation). The governing ODEs are

$$\begin{aligned} d\hat{x}/d\hat{t} &= A - B\hat{x} - C\hat{x}\hat{y}^2, \\ d\hat{y}/d\hat{t} &= B\hat{x} + C\hat{x}\hat{y}^2 - D\hat{y}, \end{aligned} \tag{5.2}$$

where A, B, C, D are positive parameters. (Recall that hats over variables indicate that they have dimensions. For now, you may safely ignore these hats.) Figure 5.2 shows a graph with two vertices, which correspond to the two concentrations $\hat{x}, \hat{y}$ in (5.2). The vertices are the bathtubs[3] in the phrase "bathtub model." The graph has four (directed) edges, with labels taken from coefficients in (5.2). The edges correspond to the following four processes:

A : Continuous addition of X to the system.
B : X decays to Y.
C : The decay of X is enhanced by Y.
D : Y decays to an inert product.

The terms $B\hat{x}$ and $C\hat{x}\hat{y}^2$ appear in both equations because these two processes affect both concentrations. In process D, when we say that Y decays to an inert product, we mean a product that has no influence on the concentrations of X or Y; the product is actually crucial for subsequent processes in glycolysis.

Typically, in bathtub models there are several processes causing movement from one category to another, and it is a wonderful simplification that, as above, equations describing the overall evolution may be obtained simply by adding the rates associated with the various processes. We bore our students by saying frequently,

[3]Sel'kov's model is represented more literally as a bathtub model in the cartoon on the frontispiece.

"This simplification makes it easy to *formulate* differential equations describing physical processes. The challenge lies in *solving* them, which is good news for keeping mathematicians employed."

Concluding remarks: (i) Since the dependent variables in a bathtub model usually represent populations of some kind, they must be nonnegative. (ii) The edges in a schematic graph such as Figure 5.2 enumerate the terms in the ODE but do not specify their exact functional form. Note the different meanings of the edges in Figure 5.2: in Processes A, B, and D, the arrows represent flow into or out of a tub, while in Process C, the arrow represents changes to a reaction rate rather than flow per se. A reaction can be either promoted (as in Process C here) or inhibited (see Section 5.6).

5.2 Scaling Example 0: Two Models from Ecology

As a gentle introduction to scaling, in this section we apply the technique to two population models from Chapter 1.

(a) The logistic equation: The most general scalar logistic equation is

$$\frac{d\hat{x}}{d\hat{t}} = A\hat{x} - B\hat{x}^2. \tag{5.3}$$

Let's ask to what extent this general equation can be simplified by introducing scaled variables, say

$$x = \frac{\hat{x}}{X}, \qquad t = \frac{\hat{t}}{T},$$

where X, T are positive constants. We substitute these scaled variables into (5.3) and manipulate the result to obtain

$$\frac{dx}{dt} = AT\,x - BTX\,x^2.$$

If we choose $T = 1/A$, $X = A/B$, we can reduce (5.3) to

$$x' = x(1 - x),$$

in which the constants A, B have disappeared!

Why is $T = 1/A$ a convenient unit of time? As we know, if the population is small, then solutions of (5.3) grow exponentially, like $e^{A\hat{t}}$. If we measure time in human-centered units—like days, months, etc.—then we have to live with whatever value of the constant A that this choice of units forces on us. However, by measuring time in units of the characteristic growth time of the exponential $e^{A\hat{t}}$, we have effectively scaled the (linear) growth rate to unity.

Similarly, using $X = A/B$ as a unit of population eliminates the coefficient of the nonlinear term in (5.3). This choice of unit also has a natural interpretation for the ODE: $\hat{x} = A/B$ is the population at which the linear growth term $A\hat{x}$ balances the nonlinear decay term $B\hat{x}^2$, i.e., the equilibrium population.

To conclude, there really is only one logistic equation, provided you are willing to measure time and population in equation-appropriate units.

It might seem more natural to use unadorned notation (i.e., without hats) for the variables in the dimensional equation (5.3) and to use hats in the scaled equation. After all, we first encounter the equations in dimensional form, and the scaled variables are then derived from the dimensional variables. However, since we quickly pass from the dimensional equations to the nondimensional equations and then spend most of our time analyzing the latter, we reserve the simpler notation for the dimensionless equations.

(b) The Lotka–Volterra equations: Let's try for the same simplification of the most general Lotka–Volterra equations,

$$\begin{array}{lrcl} (a) & d\hat{x}/d\hat{t} & = & A\hat{x} - B\hat{x}\hat{y}, \\ (b) & d\hat{y}/d\hat{t} & = & C\hat{x}\hat{y} - D\hat{y}. \end{array} \tag{5.4}$$

Thus, we want to introduce scaled variables

$$x = \frac{\hat{x}}{X}, \quad y = \frac{\hat{y}}{Y}, \quad t = \frac{\hat{t}}{T}. \tag{5.5}$$

Immediately, a key difference from the preceding example grabs our attention: there are only three parameters in (5.5), but there are four coefficients in (5.4). Thus, it seems unlikely that we can eliminate all four coefficients, but let's plunge ahead regardless. Equation (5.4) has the equilibrium solution $\hat{x} = D/C$, $\hat{y} = A/B$. Let's use these values as our scales for the population; i.e., in (5.5) we choose

$$X = D/C, \qquad Y = A/B. \tag{5.6}$$

Letting T be undetermined for the time being, we substitute (5.5) into the Lotka–Volterra equations to obtain

$$\begin{array}{lrcl} & dx/dt & = & AT\,(x - xy), \\ & dy/dt & = & DT\,(xy - y). \end{array} \tag{5.7}$$

We can choose T to simplify a coefficient in either equation, but no choice of T will work for the coefficients in both equations.

This difficulty represents a real phenomenon. The parameters A and D are the growth rate for the prey and death rate for the predators, respectively. As suggested by Table 5.1, for different systems, D/A can assume vastly different values. Without

Predator	Prey	Size of D/A
Whales	Plankton	Tiny
Wolves	Deer	Order unity
Germs	Humans	Huge

Table 5.1: *Various predator–prey systems displaying extremes of predator death rate and prey growth rate.*

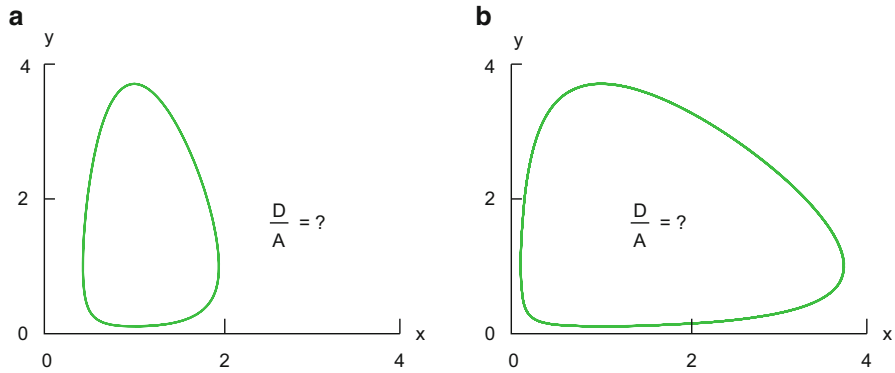

Figure 5.3: *Orbits of the Lotka–Volterra equation* (5.7) *through the point* $(1, 1/10)$ *for two different choices of* D/A. *Can you determine which panel is for* $D/A = \frac{1}{5}$ *and which is for* $D/A = 5$*?*

going to such extremes, suppose for example that $D/A = 1/5$; then in the half-life of the predators, the prey (without predation) would multiply by a factor of $2^5 = 32$, whether you measure time in microseconds, in centuries, in units derived from the ODE, or whatever. On the other hand, if $D/A = 5$, then in a predator half-life, the prey would increase only by a factor of $2^{1/5} \approx 1.149$. In other words, two different systems, say (5.4) with one set of parameter values in which $D/A = 1/5$ and with another set in which $D/A = 5$, will behave differently; it is not reasonable to expect scaling to remove differences in their behavior. Indeed, Figure 5.3 illustrates a typical trajectory for each of these cases; can you puzzle out for which case $D/A = 1/5$ and for which $D/A = 5$? (*No computing. That's cheating!*)

To dig deeper, let's consider dimensions. Every variable and parameter in (5.4) has physical units. For example, the variables $\hat{x}$, $\hat{y}$, and $\hat{t}$ have units of number of prey, number of predators, and time, respectively. We summarize this information in the self-explanatory notation

$$\mathfrak{U}(\hat{x}) = \boxed{\#\text{prey}}, \qquad \mathfrak{U}(\hat{y}) = \boxed{\#\text{predator}}, \qquad \mathfrak{U}(\hat{t}) = \boxed{\text{time}}.$$

The units of the parameters in (5.4) may be determined from the principle that all terms in an equation must have the same units. Applying this principle to the first term in (5.4a), we obtain

$$\mathfrak{U}\left(\frac{d\hat{x}}{d\hat{t}}\right) = \mathfrak{U}(A\hat{x}) = \mathfrak{U}(A)\,\mathfrak{U}(\hat{x}).$$

Making these units explicit, we deduce that

$$\frac{\boxed{\#\text{prey}}}{\boxed{\text{time}}} = \mathfrak{U}(A)\boxed{\#\text{prey}}, \qquad \text{so} \quad \mathfrak{U}(A) = \frac{1}{\boxed{\text{time}}}$$

where we have canceled $\boxed{\#\text{prey}}$. Similar analysis of (5.4b) yields $\mathfrak{U}(D) = 1/\boxed{\text{time}}$. Thus, the ratio D/A is dimensionless; i.e., the ratio will remain unchanged no matter what units are employed.

To summarize: while all parameters in (5.3) could be eliminated through scaling to obtain a unique simplified equation, the best we can hope for with (5.4) is a *one-parameter family* of simplified equations. Indeed, if we choose $T = 1/A$, then (5.7) may be rewritten

$$\begin{aligned} dx/dt &= x - xy, \\ dy/dt &= \rho(xy - y), \end{aligned} \tag{5.8}$$

where $\rho = D/A$ is the (dimensionless) parameter that cannot be scaled away.

Let's push a little further. For the nonlinear terms in (5.4) to have the same units as the other terms, we need

$$\frac{\boxed{\#\text{prey}}}{\boxed{\text{time}}} = \mathfrak{U}(B)\boxed{\#\text{prey}}\boxed{\#\text{predator}} \quad \text{and} \quad \frac{\boxed{\#\text{predator}}}{\boxed{\text{time}}} = \mathfrak{U}(C)\boxed{\#\text{prey}}\boxed{\#\text{predator}},$$

which implies that

$$\mathfrak{U}(B) = \frac{1}{\boxed{\text{time}}\boxed{\#\text{predator}}} \quad \text{and} \quad \mathfrak{U}(C) = \frac{1}{\boxed{\text{time}}\boxed{\#\text{prey}}}.$$

Note that for the x-scale (5.6),

$$\mathfrak{U}(X) = \frac{\mathfrak{U}(D)}{\mathfrak{U}(C)} = \boxed{\#\text{prey}},$$

so that the scaled variable $x = \hat{x}/X$ is dimensionless. Similarly, $\mathfrak{U}(Y) = \boxed{\#\text{predator}}$, so y is also dimensionless. This behavior is typical—*a scale for a variable that simplifies the equation will normally make the scaled variable dimensionless*, and conversely.

Incidentally, it is meaningless to speak of a quantity with nontrivial units being either large or small. Let's support this statement with an apparently outrageous assertion: one normally thinks of the speed of light as a very large quantity, but

in fact, it is only 2×10^{-6} in our choice of units—we use astronomical units per millisecond.[4] The speed of light is fast compared with velocities encountered in ordinary (for human beings) circumstances, like the speed of sound or the speed of a runner. Mathematically, this means that the *dimensionless ratio* of the speed of light divided by a typical reference speed is large.

5.3 Scaling Example 1: A Nonlinear Oscillator

Consider the equation

$$m\frac{d^2\hat{x}}{d\hat{t}^2} + b\frac{d\hat{x}}{d\hat{t}} + k_1\hat{x} + k_2\hat{x}^3 = 0, \tag{5.9}$$

where all constants are positive.[5] Regarding units, for the variables we have

$$\mathfrak{U}(\hat{t}) = \boxed{\text{time}}, \qquad \mathfrak{U}(\hat{x}) = \boxed{\text{length}},$$

and of course $\mathfrak{U}(m) = \boxed{\text{mass}}$, so

$$\mathfrak{U}\left(m\,\frac{d^2\hat{x}}{d\hat{t}^2}\right) = \frac{\boxed{\text{mass}}\,\boxed{\text{length}}}{\boxed{\text{time}}^{\,2}}.$$

From this information we may deduce the units of the other parameters in (5.9). For example, requiring consistent units between the first and second terms gives us

$$\frac{\boxed{\text{mass}}\,\boxed{\text{length}}}{\boxed{\text{time}}^{\,2}} = \mathfrak{U}\left(b\,\frac{d\hat{x}}{d\hat{t}}\right) = \mathfrak{U}(b)\,\frac{\boxed{\text{length}}}{\boxed{\text{time}}},$$

from which it follows that $\mathfrak{U}(b) = \boxed{\text{mass}}/\boxed{\text{time}}$, as given in Table 5.2. The other entries in the table may be verified similarly.

We begin nondimensionalization and scaling by asking two basic questions:

- What dimensionless quantities can be constructed from the parameters in the problem by forming products of powers, i.e., quantities of the form

  $$m^\alpha\, b^\beta\, k_1^\gamma\, k_2^\delta,$$

  where the exponents $\alpha, \beta, \gamma, \delta$ may be chosen arbitrarily?

[4] Would you prefer light-years per year? Unity is *such* a nice velocity.

[5] Thus, this equation differs from Duffing's equation (1.28) in that the linear part of the force law, $k_1\hat{x}$, is attracting. This sign difference has no effect on scaling the equation; it just makes interpretation of the time scale $T = \sqrt{m/k_1}$, which we introduce below, slightly more transparent.

Variables	Description	Units
t	Time	$\boxed{\text{time}}$
x	Length	$\boxed{\text{length}}$
Parameters		
m	Mass	$\boxed{\text{mass}}$
b	Friction coefficient	$\boxed{\text{mass}}\,/\,\boxed{\text{time}}$
k_1	Linear spring constant	$\boxed{\text{mass}}\,/\,\boxed{\text{time}}^2$
k_2	Nonlinear spring constant	$\boxed{\text{mass}}\,/\,\boxed{\text{length}}^2\,\boxed{\text{time}}^2$

Table 5.2: *Units of quantities in equation* (5.9).

- How can the equations be simplified by introducing scaled variables,

$$x = \hat{x}/L, \qquad t = \hat{t}/T? \tag{5.10}$$

Regarding the first question, we find from Table 5.2 that

$$\mathfrak{U}(m^\alpha\, b^\beta\, k_1^\gamma\, k_2^\delta) = \boxed{\text{mass}}^{\alpha+\beta+\gamma+\delta}\,\boxed{\text{time}}^{-\beta-2\gamma-2\delta}\,\boxed{\text{length}}^{-2\delta}.$$

Requiring this to be dimensionless means that we must have

$$\begin{aligned} \alpha+\beta+\gamma+\delta &= 0,\\ \beta+2\gamma+2\delta &= 0,\\ \delta &= 0. \end{aligned}$$

This is a system of three homogeneous linear equations in four unknowns, and the coefficient matrix has rank 3, so there is one linearly independent solution. For this we may take $\beta = 1, \alpha = \gamma = -1/2$, which gives the dimensionless quantity $b/\sqrt{mk_1}$.

To address the second question, let's substitute (5.10) into (5.9); after dividing the equation by mL/T^2, we have

$$\frac{d^2x}{dt^2} + \frac{bT}{m}\frac{dx}{dt} + \frac{k_1T^2}{m}x + \frac{k_2L^2T^2}{m}x^3 = 0.$$

We want to choose L and T to make the three coefficients

$$\frac{bT}{m}, \qquad \frac{k_1T^2}{m}, \qquad \frac{k_2L^2T^2}{m} \tag{5.11}$$

as simple as possible. Since the effect of L and T is mixed together in the third coefficient in (5.11), it makes sense to choose T first: we let $T = \sqrt{m/k_1}$ to make

the middle coefficient equal to unity,[6] so the ODE becomes

$$\frac{d^2x}{dt^2} + \frac{b}{\sqrt{mk_1}}\frac{dx}{dt} + x + \frac{k_2L^2}{k_1}x^3 = 0.$$

Note that the coefficient of the first-order derivative is our dimensionless quantity, $b/\sqrt{mk_1}$; the choice of length scale will not affect this coefficient. However, we may simplify the nonlinear term by choosing $L = \sqrt{k_1/k_2}$, which reduces the equation to

$$\frac{d^2x}{dt^2} + \frac{b}{\sqrt{mk_1}}\frac{dx}{dt} + x + x^3 = 0. \tag{5.12}$$

We have now achieved maximum simplicity: all coefficients in the equation are either simple numbers—in this case, unity—or the dimensionless quantity constructed in answering Question 1.

The task of scaling is not complete until you have interpreted the dimensionless quantities in Question 1 and the scale parameters T and L in Question 2. To this end, let's compare (5.9) with the linear equation

$$m\frac{d^2\hat{x}}{d\hat{t}^2} + b\frac{d\hat{x}}{d\hat{t}} + k_1\hat{x} = 0, \tag{5.13}$$

extracting three points.

- If $b = 0$ (no friction), then (5.13) has trigonometric solutions $\cos(\sqrt{k_1/m}\,t)$ and $\sin(\sqrt{k_1/m}\,t)$. *The time scale T derives from the period $2\pi\sqrt{m/k_1}$ of these linear oscillations.*

- If b is small, the oscillations have approximately the same period but decay as $e^{-bt/2m}$. Over the duration of the time scale T, the oscillations decay by a factor of $e^{-bT/2m} = e^{-b/2\sqrt{mk_1}}$. *The dimensionless parameter $b/\sqrt{mk_1}$ is a measure of the strength of friction compared to other relevant effects in the equation.*

- In the force law of the full equation (5.9), the linear term $k_1\hat{x}$ dominates for small $\hat{x}$, while the cubic term $k_2\hat{x}^3$ dominates for large $\hat{x}$. *The length-scale parameter $L = \sqrt{k_1/k_2}$ separates the two regimes*; it is the displacement where the nonlinear and linear forces are exactly equal.

In some ways, scaling as in (5.10) is analogous to scientific notation, in which very large or small quantities are written as a number on the order of unity times a power of 10. For example, the speed of light in meters/second (more reasonable units than

[6] What about choosing $T = m/b$ to simplify the first coefficient? In Section 5.9, we argue that provided friction is not too large, the spring constant k_1 provides a more useful basis for scaling than b. But such choices are rarely cut-and-dried.

above) is 2.9979×10^8. Similarly, in scaling we write the dimensional variable $\hat{x}$ as a product $\hat{x} = xL$. The scale factor L, which is analogous to the power of 10, might represent the separation between two atoms in a large organic molecule, a typical extension length of a spring in a laboratory, the distance from the sun to a planet, or whatever; for the mathematics, it does not matter which. In studying the equation, it releases brain cells for more creative tasks if you focus on the nondimensional formulation of the problem in terms of x and push L into the background.

Here is another perspective on the importance of dimensionless quantities: no one parameter in (5.13) by itself determines the behavior of solutions of this equation. For example, one may change two parameters in such a way that each change compensates for the other and the behavior of the system is unchanged (apart from scale factors). By contrast, problems for which the dimensionless combination $b/\sqrt{mk_1}$ is large and for which it is small are genuinely different systems. Thus, for instance, the difference between underdamped, overdamped, and critically damped, defined in Section 1.4.1, depends on the size of $b/\sqrt{mk_1}$.

5.4 Scaling Example 2: Sel'kov's Model for Glycolysis

The dimensional formulation (5.2) of Sel'kov's model,

$$\begin{aligned} d\hat{x}/d\hat{t} &= A - B\hat{x} - C\hat{x}\hat{y}^2, \\ d\hat{y}/d\hat{t} &= B\hat{x} + C\hat{x}\hat{y}^2 - D\hat{y}, \end{aligned} \tag{5.14}$$

was introduced as part of the discussion of bathtub models in Section 5.1.3. In this section we show how to reduce (5.14) to the simpler system (4.32) analyzed in Chapter 4.

In Exercise 1 we ask you to perform the first two steps of this reduction, beginning with the following:

- Verify the units of the coefficients A, B, C, D given in Table 5.3.

Note that both the concentrations $\hat{x}$ and $\hat{y}$ have units of molarity. (Since the two equations contain identical terms, for (5.14) to be dimensionally consistent, $\hat{x}$ and $\hat{y}$ must have the *same* units.) It is not necessary to understand what molarity means to do this exercise;[7] you may simply treat molarity as a fundamental unit to be

[7]However, to support scientific literacy we explain molarity in the Pearls. One point is relevant here: concentration in molarity is based on counting molecules per unit volume, not mass per unit volume. This choice allows one to readily compare the concentrations of the different substances X and Y, and it is responsible for the pleasant feature that the coefficients of the terms $B\hat{x}$ and $C\hat{x}\hat{y}^2$ have equal magnitudes in both equations (5.14).

Variables	Description	Units
$\hat{t}$	Time	$\boxed{\text{time}}$
$\hat{x}, \hat{y}$	Concentrations	$\boxed{\text{molarity}}$
Parameters		
A	Addition rate	$\boxed{\text{molarity}} \,/\, \boxed{\text{time}}$
B, D	Decay rates	$1/\boxed{\text{time}}$
C	Nonlinear enhancement of decay	$1/\boxed{\text{molarity}}^{\,2}\,\boxed{\text{time}}$

Table 5.3: *Units of quantities in Sel'kov's model, (5.2). As explained in the Pearls, molarity means moles per liter of solution.*

manipulated along with other units. Thus for example, matching units of terms in the first equation in (5.14), we see that

$$\mathfrak{U}(A) = \frac{\boxed{\text{molarity}}}{\boxed{\text{time}}}.$$

As a second task, we ask you to

- Determine what dimensionless parameters may be constructed from the coefficients A, B, C, D in (5.14).

Specifically, show that there are exactly two independent dimensionless combinations $A^\alpha B^\beta C^\gamma D^\delta$ of the parameters, which we may take to be

$$\rho = A\sqrt{\frac{C}{D^3}}, \qquad \sigma = \frac{B}{D}. \tag{5.15}$$

Usually, in scaling a 2×2 system we would introduce separate scale parameters for $\hat{x}$ and $\hat{y}$. However, in the present context we want to preserve the fact that the terms $B\hat{x}$ and $C\hat{x}\hat{y}^2$ appear with equal coefficients in both equations, so we use the same scale parameter Y for both variables. Thus, we substitute

$$x = \hat{x}/Y, \quad y = \hat{y}/Y, \quad t = \hat{t}/T$$

into (5.2), which yields

$$\begin{aligned} dx/dt &= AT/Y - BT\,x - CTY^2\,xy^2, \\ dy/dt &= BT\,x + CTY^2\,xy^2 - DT\,y. \end{aligned} \tag{5.16}$$

We choose $T = 1/D$ to simplify the decay term proportional to y in the second equation, and we choose $Y = \sqrt{D/C}$ to simplify the terms proportional to xy^2 in both equations. These choices result in the equations

$$\begin{aligned} dx/dt &= \rho - \sigma x - xy^2, \\ dy/dt &= -y + \sigma x + xy^2, \end{aligned} \tag{5.17}$$

where ρ, σ are defined by (5.15), which is the form of Sel'kov's model we analyzed in Chapter 4. We ask you to check that x, y, t are all dimensionless.

It remains to interpret the scaling. The parameter ρ may be viewed as the rate at which X is being added to the system, relative to other scales in the problem, and σ is the ratio of the uncatalyzed reaction rate $\mathsf{X} \to \mathsf{Y}$ to the decay rate of Y. The time scale T is based on the decay rate of Y to an inert product. As occasionally happens, the concentration scale Y is rather unilluminating: if both $\hat{x}$ and $\hat{y}$ equal Y, then $C\hat{x}\hat{y}^2$, the catalyzed contribution to the reaction $\mathsf{X} \to \mathsf{Y}$, equals $D\hat{y}$, which specifies the decay of Y. However, even if its interpretation is uninspiring, this choice for Y gives simple dimensionless equations.

5.5 Scaling Example 3: The Chemostat

The chemostat involves flow through a reactor. Elementary ODE books often pose flow problems, but they seem to cause students an undue amount of trouble. Therefore, before introducing the chemostat, we review such problems.

5.5.1 ODEs Modeling Flow Through a Reactor

Consider the configuration indicated in Figure 5.4. A chemical X is dissolved in water in a tank; the tank is well-stirred, which means that the concentration of X is uniform throughout the tank. Water containing the chemical in constant concentration C is added to the tank at the rate r_1. The well-stirred mixture is drawn off at a rate r_2. The state of the system at every time is specified by $\hat{V}$, the volume of mixture in the tank, and $\hat{x}$, the concentration of the chemical in the mixture. (Let's measure $\hat{x}$ in *mass* per unit volume rather than molarity; see Table 5.4.)

The first variable evolves according to the obvious equation

$$\frac{d\hat{V}}{d\hat{t}} = r_1 - r_2,$$

at least as long as $\hat{V} > 0$.

The equation for $\hat{x}$ is best accessed by considering the total mass of X in the tank, i.e., $\hat{V}\hat{x}$. We claim that this amount evolves according to the equation

$$\frac{d}{d\hat{t}}(\hat{V}\hat{x}) = r_1 C - r_2 \hat{x}.$$

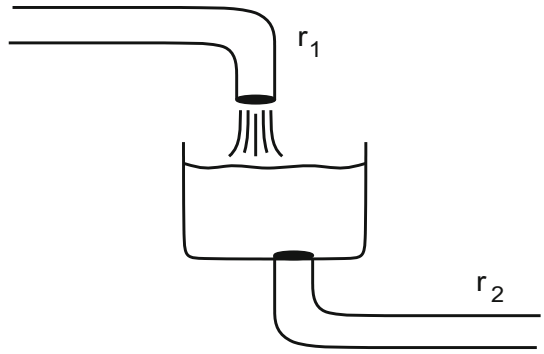

Figure 5.4: *Schematic of the flow problem considered in Section 5.5.1: An aqueous solution containing a chemical of constant concentration C flows at a rate r_1 into a continuously stirred tank, and the mixture drains from the tank at a rate r_2.*

Variables	Description	Units
$\hat{t}$	Time	time
$\hat{V}$	Volume	volume
$\hat{x}$	Concentration	mass / volume
Parameters		
r_1	Addition rate	volume / time
r_2	Withdrawal rate	volume / time
C	Concentration in feed	mass / volume

Table 5.4: *Units of quantities in the reactor model model in Section 5.5.1.*

To derive this equation, consider how the mass changes in a short time interval $(\hat{t}, \hat{t} + \Delta t)$. In time Δt, a mass of X equal to $(r_1 \Delta t)C$ is added to the mixture. Similarly, in the same time interval, a mass of dissolved substance approximately equal to $(r_2 \Delta t)\hat{x}(\hat{t})$ is withdrawn from the tank; this estimate is only approximate, since $\hat{x}$ varies over time, but it will be nearly constant over a short time interval. Thus,

$$(\hat{V}\hat{x})(\hat{t} + \Delta t) - (\hat{V}\hat{x})(\hat{t}) \approx \Delta t\, r_1 C - \Delta t\, r_2 \hat{x}(\hat{t});$$

dividing by Δt and passing to the limit $\Delta t \to 0$, we get the above equation.

While these equations are perhaps too trivial to nondimensionalize, we think you may find it instructive to verify explicitly that

$$\mathfrak{U}(\hat{V}\hat{x}) = \mathfrak{U}((r_1 \Delta t)C) = \boxed{\text{mass}}.$$

As an important special case, suppose $r_1 = r_2$, and write r for the common value. In this case, the volume $\hat{V}$ is a constant, say V, and the concentration equation simplifies to

$$\frac{d\hat{x}}{d\hat{t}} = (r/V)(C - \hat{x}). \tag{5.18}$$

The general solution of this linear equation, particular solution plus homogeneous solution, is

$$\hat{x}(\hat{t}) = C + \text{const } e^{-(r/V)\hat{t}}.$$

Thus, the concentration $\hat{x}$ converges exponentially to the concentration C of the feed at the rate r/V.

5.5.2 The Chemostat

A chemostat is a primarily a humdrum device in the laboratory for maintaining a population of bacteria in a state always ready for experiments. The bacteria live inside a reactor. A solution containing dissolved nutrients is added continuously to the reactor, and the contents of the reactor—a mixture containing bacteria plus whatever nutrients the bacteria have not consumed—are drained off at the same rate. The tank is well stirred, so both bacteria and nutrients are uniformly distributed throughout the reactor.

The state of the system is specified by two variables, the bacteria population per unit volume in the chemostat, $\hat{x}$, and the concentration of nutrients, $\hat{y}$. These variables evolve according to the ODEs

$$\begin{aligned} \frac{d\hat{x}}{d\hat{t}} &= A\frac{\hat{y}}{K+\hat{y}}\hat{x} - \frac{r}{V}\hat{x}, \\ \frac{d\hat{y}}{d\hat{t}} &= -B\frac{\hat{y}}{K+\hat{y}}\hat{x} + \frac{r}{V}(C - \hat{y}), \end{aligned} \tag{5.19}$$

where A, B, C, K, V, r are positive constants. The terms in (5.19) are indicated schematically in Figure 5.5. In the linear terms of either equation, C, V, and r have the same meanings as above: C is the concentration of nutrients in the feed, V is the volume of the reactor, and r is the rate at which fluid is added to and drained from the tank. In the nonlinear terms, say in the second equation, B and K describe the consumption of nutrients by the bacteria. At low concentrations of nutrients ($\hat{y} \ll K$), the consumption rate is proportional to the density of nutrients $\hat{y}$; but at high concentration, the rate saturates at a level independent of $\hat{y}$. (A bacterium can eat only so fast.) The nonlinear term in the first equation expresses the assumption that the growth of the bacteria is proportional to the consumption of nutrients.

As listed in Table 5.5, the units of $\hat{x}$ and $\hat{y}$ are number of cells per unit volume and mass per unit volume, respectively. In Exercise 1 we ask you to (i) verify all the

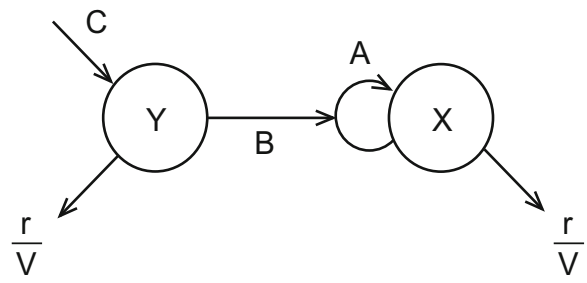

Figure 5.5: *Schematic diagram of the chemostat model* (5.19). *The arrow labeled B does double duty: it represents a reaction that depletes the concentration of* Y, *and it indicates that* Y *promotes the reaction through which* X *grows.*

Variables	Description	Units
$\hat{t}$	Time	time
$\hat{x}$	Cell population	#cells / volume
$\hat{y}$	Nutrient concentration	mass / volume
Parameters		
A	Bacterial growth rate	1/time
B	Consumption rate	mass / #cells time
K	Saturation concentration	mass / volume
C	Concentration in feed	mass / volume
V	Volume of reactor	volume
r	Feed rate	volume / time

Table 5.5: *Units of quantities in the chemostat equations,* (5.19).

dimensions listed in the table and (ii) apply the usual technique to show that exactly two independent dimensionless parameters $A^\alpha B^\beta C^\gamma K^\delta V^\epsilon r^\zeta$ may be constructed from these parameters, which we may take to be

$$\rho = \frac{r}{VA} \quad \text{and} \quad \sigma = \frac{C}{K}. \tag{5.20}$$

Let's introduce scaled variables

$$x = \frac{\hat{x}}{X}, \quad y = \frac{\hat{y}}{K}, \quad t = \frac{\hat{t}}{T}. \tag{5.21}$$

We will choose X and T by substituting (5.21) into (5.19) and making the equations as simple as possible. By contrast, we specify the $\hat{y}$-scale a priori as the saturation

level K in the consumption terms in (5.19); this choice simplifies the factors $\hat{y}/(K+\hat{y})$. If we take $T = 1/A$ and $X = KA/B$, we obtain the scaled equations

$$\begin{aligned} \frac{dx}{dt} &= \frac{y}{1+y}x - \rho x, \\ \frac{dy}{dt} &= -\frac{y}{1+y}x - \rho(y-\sigma), \end{aligned} \tag{5.22}$$

where ρ, σ are given by (5.20), which is the form (4.21) studied in Chapter 4. Check that x, y, t are dimensionless.

Regarding interpretation of the scaling: regrouping the factors in (5.20) as $\rho = (r/V)\,/\,A$, we see that ρ is the ratio of two rates: r/V is the "turnover" rate at which the feed changes the concentration of nutrients in the tank, and A is the bacterial growth rate under conditions of ample food. The other dimensionless parameter, $\sigma = C/K$, is the ratio of two concentrations. The time scale T is set by the bacterial growth rate. As noted above, the nutrient scale is the saturation concentration K. The scaling $x = \hat{x}/X$ relates bacterial concentration to the nutrients that this population would consume. Specifically, $(B/K)\hat{x}$ is the nondimensionalized consumption rate of nutrients under conditions of ample food, and $x = (B/KA)\hat{x}$ is the nondimensionalized quantity of nutrients that would be consumed at this rate in the characteristic growth time $1/A$ of the bacteria.

The choices in scaling are far from unique. Illustrating this point, Exercise 2 proposes an alternative scaling for the chemostat that some authors prefer.

5.6 Scaling Example 4: An Activator–Inhibitor System

In this section and the next we illustrate an additional benefit of scaling: it helps you assess the relative importance of various effects in an equation and suggests appropriate approximations to simplify them.

The activator–inhibitor system is described by the equations[8]

$$\begin{aligned} (a)\quad \frac{d\hat{x}}{d\hat{t}} &= A\frac{L}{L+\hat{y}}\frac{\hat{x}^2}{K^2+\hat{x}^2} - B\hat{x}, \\ (b)\quad \frac{d\hat{y}}{d\hat{t}} &= C\frac{\hat{x}^2}{K^2+\hat{x}^2} - D\hat{y}, \end{aligned} \tag{5.23}$$

[8] A reaction rate of the form $\hat{x}^n/(K^n+\hat{x}^n)$, as appears in (5.23), is called a *Hill* function. If $n = 1$, this rate reduces to Michaelis–Menten kinetics (see Section 5.7). These rates arise from enzymatic reactions; when cooperative effects are important, $n > 1$. See Chapter 1 of Keener and Sneyd [47].

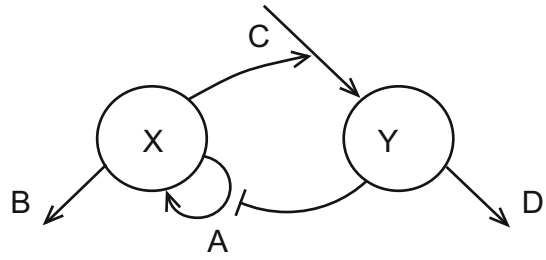

Figure 5.6: *Schematic illustration of the activator–inhibitor model* (5.23).

Variables	Description	Units
$\hat{t}$	Time	[time]
$\hat{x}, \hat{y}$	Concentrations	[molarity]
Parameters		
A, C	Production coefficients	[molarity] / [time]
B, D	Decay rates	1 / [time]
K, L	Saturation concentrations	[molarity]

Table 5.6: *Units of quantities in the activator–inhibitor model,* (5.23).

where A, B, C, D, K, L are positive constants; the terms in (5.23) are indicated schematically[9] in Figure 5.6. The reactant X is called a *self-activator* or *self-promoter*, since its production rate increases with its own concentration, $\hat{x}$. Self-activation is indicated in Figure 5.6 by the closed loop at vertex X . The reactant Y is called an *inhibitor* or *repressor*, because a high value of $\hat{y}$ suppresses the production of X through the factor $L + \hat{y}$ in the denominator of the first term in (5.23a). To indicate inhibition in the figure, the edge originating at Y terminates in a short crossbar at the edge associated to the reaction it inhibits. As usual, nothing in the figure indicates the exact functional dependence of the reaction rates on concentrations; only the presence of an effect and its sign are indicated.

Units of the parameters in (5.23) are given in Table 5.6. In the standard way we find that there are exactly four dimensionless combinations of these parameters, for example

$$\frac{A}{BK}, \quad \frac{C}{DL}, \quad \frac{D}{B}, \quad \frac{L}{K}. \tag{5.24}$$

Interpretation of these parameters is not particularly illuminating. The significant information about these parameters is that in typical applications, A/BK is large

[9]Do you find this figure useful? If not, feel free to ignore it. We find the schematic diagrams less informative as they become more complicated.

and C/DL is huge, the latter on the order of 10^3 or more. A good scaling must accommodate this fact: this means that in choosing the scaling

$$x = \frac{\hat{x}}{X}, \quad y = \frac{\hat{y}}{Y}, \quad t = \frac{\hat{t}}{T},$$

the scaled variables x and y should be on the order of unity near equilibrium. If $\hat{x} = X$ and $\hat{y} = Y$, then equilibrium in the second equation in (5.23) gives

$$\frac{X^2}{K^2 + X^2} = \frac{DY}{C}. \tag{5.25}$$

We propose[10] to take $Y = L$, the saturation concentration in the reaction rate $L/(L + \hat{y})$. With this choice, the RHS of (5.25) equals $DL/C \ll 1$. Therefore, (5.25) implies that $X \ll K$. Neglecting X compared to K in the denominator and solving, we obtain the approximation $X = \sqrt{DL/C}\, K$. With these choices and letting $T = 1/B$, we obtain the scaled equations[11]

$$\begin{aligned} (a) \quad \frac{dx}{dt} &= \sigma \frac{1}{1+y} \frac{x^2}{1 + x^2/\kappa^2} - x, \\ (b) \quad \frac{dy}{dt} &= \rho \left[\frac{x^2}{1 + x^2/\kappa^2} - y \right], \end{aligned} \tag{5.26}$$

where

$$\kappa = \sqrt{\frac{C}{DL}} \gg 1, \qquad \sigma = \frac{A/BK}{\sqrt{C/DL}}, \quad \rho = \frac{D}{B},$$

which is the system (4.30) studied in Section 4.4.1. Incidentally, both the numerator and denominator in σ are large, but they are comparable, and typically σ is of modest size.

[10] Let's admit it—we needed several tries to get this right. After the fact, we can motivate focusing on the second equation, since this equation contains coefficients that appear in the largest parameter.

[11] Why are there only three parameters in (5.26), even though there are four dimensionless combinations of the parameters in (5.23)? The reason is that even though both $\hat{x}$ and $\hat{y}$ have units of molarity, we chose different x- and y-scales, and this gave us an extra degree of freedom in simplifying the equations. By contrast, this option was not attractive to us in scaling the equations (5.14) for Sel'kov's model, because if we had chosen different scales for x and y in that problem, the terms $B\hat{x}$ and $C\hat{x}\hat{y}^2$ would then have had unequal coefficients in the two equations. The difference is that in Sel'kov's model, single reaction terms change *both* concentrations, subtracting from X and adding to Y.

At last we reap the rewards of good scaling. We may suspect the equations (5.23) can be simplified in some way based on the fact that they contain large parameters, but it is far from obvious how to do so. By contrast, it is trivial to take the limit $\kappa \to \infty$ in (5.26):

$$\begin{aligned} (a) \quad x' &= \sigma \frac{x^2}{1+y} - x, \\ (b) \quad y' &= \rho\,[x^2 - y]. \end{aligned} \tag{5.27}$$

This is a much less intimidating system of equations that has only two dimensionless parameters but still exhibits the interesting behavior of the original system.

5.7 Scaling Example 5: Michaelis–Menten Kinetics

This example, like the preceding one, illustrates how scaling can help in assessing the relative importance of various effects in the equations and can suggest simplifying approximations.

Michaelis–Menten kinetics arise as an approximation of certain reaction rates where an enzyme is involved. An enzyme is a catalyst in biochemical reactions; i.e., it facilitates the reaction but is not itself consumed. Consider a chemical reaction of the form[12]

$$\mathsf{R} + \mathsf{E} \rightleftharpoons [\mathsf{RE}] \to \mathsf{P} + \mathsf{E}\ . \tag{5.28}$$

Here R is a reactant (often called a substrate), E is an enzyme, [RE] is a compound in which the reactant and the enzyme are bound to one another, and P is a product. The reaction R+E $\rightleftharpoons$ [RE] is reversible, while the production of the product is considered irreversible. According to the law of mass action (see Chapter 6 of Murray [57] or Chapter 1 of Keener and Sneyd [47]), a binary reaction such as R+E $\to$ [RE] proceeds at a rate proportional to the product of the concentrations of R and E, while the unary reactions [RE] $\to$ R+E and [RE] $\to$ P+E proceed at rates proportional to the concentration of [RE]. Let $\hat{r}$, $\hat{e}$, $\hat{c}$, $\hat{p}$ denote the concentrations of R, E, [RE], P, respectively (mnemonic: "c" for compound). We assume that all concentrations are uniform over some region in space, so that the concentrations are described by ODEs

[12]Reaction (5.28) might seem to violate conservation of mass; otherwise, how could P be different from R? Typically, the product P is what's called an isomer of R, a substance with the same chemical composition but with different spatial configuration. Incidentally, the second reaction [RE] $\to$ P+E is treated as irreversible; this assumption is appropriate if, for example, the product is produced in the cell nucleus but is transported out of the nucleus to the cell body before the reverse reaction can occur.

(rather than PDEs), specifically by the equations[13]

$$\begin{aligned} d\hat{r}/d\hat{t} &= -k_{+1}\hat{r}\hat{e} + k_{-1}\hat{c}, \\ d\hat{e}/d\hat{t} &= -k_{+1}\hat{r}\hat{e} + k_{-1}\hat{c} + k_2\hat{c}, \\ d\hat{c}/d\hat{t} &= k_{+1}\hat{r}\hat{e} - k_{-1}\hat{c} - k_2\hat{c}, \\ d\hat{p}/d\hat{t} &= k_2\hat{c}, \end{aligned} \tag{5.29}$$

where the constants k_j, all positive, specify reaction rates. Although this is a system of four equations, it may be reduced to two equations. In the first place, the equation for $\hat{p}$ decouples from the other three equations, so it may be ignored until the other three have been solved. Less trivially, by adding the $\hat{e}$ and $\hat{c}$ equations, we derive the following claim *(Check this!)*:

Claim 5.7.1. *The sum $\hat{e} + \hat{c}$ is independent of time.*

Acting on the claim, we let E_0 denote the constant value of $\hat{e} + \hat{c}$; for example, if initially none of the compound [RE] is present, then E_0 is the initial value $\hat{e}(0)$. We may use the relation $\hat{e} = E_0 - \hat{c}$ to eliminate $\hat{e}$ from the equations, yielding the two-dimensional system

$$\begin{aligned} d\hat{r}/d\hat{t} &= -k_{+1}\hat{r}(E_0 - \hat{c}) + k_{-1}\hat{c}, \\ d\hat{c}/d\hat{t} &= k_{+1}\hat{r}(E_0 - \hat{c}) - k_{-1}\hat{c} - k_2\hat{c}. \end{aligned} \tag{5.30}$$

In Table 5.7 we have listed plausible values for the parameters[14] in (5.30). We invite you to simulate the IVP for these equations, say with initial conditions $\hat{r}(0) = 10^{-2}$ M and $\hat{c}(0) = 0$. Then compare what you learn from the simulation with the insight gained from the following scaling analysis.

In the Exercises we ask you (i) to verify the units listed in Table 5.7 and (ii) to show that two dimensionless combinations may be constructed from the parameters in (5.30), which may be chosen to be

$$\kappa = \frac{k_2}{k_{-1}} \quad \text{and} \quad \varepsilon = \frac{E_0}{k_{-1}/k_{+1}}. \tag{5.31}$$

For the numbers in Table 5.7, $\varepsilon = 10^{-3}$ is very small, and this is typical.

---

[13] Note that the same reaction constants appear in different equations. As in Sel'kov's model, this simplification arises because we measure all concentrations in molarity, which is based on counting molecules. (Molarity is explained in the Pearls.) In contrast to Sel'kov's model, when we simplify the present problem, we choose different scales for $\hat{r}$ and $\hat{c}$, scales that reflect typical orders of magnitude for the variables.

[14] These are not the parameters for any specific reaction; they are arbitrarily chosen values in the midrange of what is observed experimentally in reactions of this type.

Variables	Description	Units	
t	time	time	
r, c	concentration	molarity	
Parameters			Plausible value
k_{+1}	reaction constant	$1/$ molarity time	10^7 /M sec
k_{-1}	reaction constant	$1/$ time	10^4 /sec
k_2	reaction constant	$1/$ time	10^2 /sec
E_0	enzyme concentration	molarity	10^{-6} M

Table 5.7: *Units of and some plausible values for quantities in (5.30), the reactions leading to Michaelis–Menten kinetics. The molarity unit M stands for moles/liter.*

To nondimensionalize (5.30), we define

$$r = \frac{\hat{r}}{R}, \qquad c = \frac{\hat{c}}{E_0}, \qquad t = \frac{\hat{t}}{T}. \tag{5.32}$$

The scale for $\hat{c}$ seems inevitable: since $\hat{e} + \hat{c} = E_0$, the parameter E_0 is an upper bound for $\hat{c}$, so the scaled variable $c = \hat{c}/E_0$ has the range $0 \le c \le 1$. The scale for $\hat{r}$ is more subtle. We could choose $R = \hat{r}(0)$ as a scale for this variable, but there is a more intrinsic choice that involves only parameters in equation (5.30); this has the advantage of being usable for all initial conditions. To motivate this scale, let us isolate part of the reaction scheme (5.28),

$$\mathsf{R} + \mathsf{E} \rightleftharpoons [\mathsf{RE}]. \tag{5.33}$$

The forward and reverse reactions will be in balance if

$$k_{+1}\hat{r}\hat{e} = k_{-1}\hat{c}.$$

If exactly half of the enzyme is free and half is bound to R—i.e., if $\hat{e} = \hat{c} = E_0/2$—then this equation reduces to $\hat{r} = k_{-1}/k_{+1}$. We propose $R = k_{-1}/k_{+1}$ as the scale[15] to use in (5.32). Note that $E_0/R = \varepsilon$, where ε is defined in (5.31). In physical terms, ε is small because R is abundant, while E and [RE] are in short supply.

Leaving the time scale T undetermined for the moment, we substitute (5.32) into (5.30) and rearrange to obtain

$$\begin{aligned} dr/dt &= \varepsilon k_{-1} T[-r(1-c) + c], \\ dc/dt &= k_{-1} T[r(1-c) - (1+\kappa)c]. \end{aligned} \tag{5.34}$$

[15]Incidentally, if you like jargon: chemists call the quantity k_{-1}/k_{+1} the *dissociation constant* of the reaction (5.33).

Even before we have chosen T, an unexpected consequence has dropped out from systematic scaling: in (5.35), *the reaction rate in the c-equation is effectively* ε^{-1} *times faster than in the r-equation.* This difference in rates may seem surprising, because for every molecule that the term $-k_{+1}\hat{r}(E_0 - \hat{c})$ removes from $\hat{r}$, exactly one molecule is added to $\hat{c}$; and for every molecule that the term $-k_{-1}\hat{c}$ removes from $\hat{c}$, exactly one molecule is added to $\hat{r}$. However, because $\hat{c} \ll \hat{r}$, *on a percentage basis* the same increment or decrement is a much greater change in $\hat{c}$ than in $\hat{r}$, which makes the effective rates so different.

We choose $T = 1/\varepsilon k_{-1}$ to make the overall coefficient in the first equation in (5.34) equal to unity, which yields the fast–slow system

$$\begin{aligned} dr/dt &= -r(1-c) + c, \\ \varepsilon\, dc/dt &= r(1-c) - (1+\kappa)c. \end{aligned} \tag{5.35}$$

As we saw in Section 4.4.5, after a brief transient,[16] evolution in (5.35) is well approximated by the scalar ODE for r derived by setting $\varepsilon = 0$. In other words, scaling has led us to a great simplification of (5.30).

Because of the scientific importance of the Michaelis–Menten approximation, let us return to unscaled variables to express the rate at which the product P is produced after the transient has decayed,

$$\frac{d\hat{p}}{d\hat{t}} = k\frac{\hat{r}}{K + \hat{r}}, \tag{5.36}$$

where $k = k_2 E_0$ and $K = (k_{-1} + k_2)/k_{+1}$. The reaction rate (5.36) is proportional to $\hat{r}$ when this variable is small (compared to K), but the rate saturates at $k_2 E_0$ when $\hat{r}$ is large.

Concluding remarks: (i) The other dimensionless parameter κ specifies what proportion of molecules of the compound [RE] decay back to the separate components R and E and what proportion produce the product P. (ii) If an enzyme *inhibits* a reaction, the reaction rate may have the form $1/(K + \hat{r})$, as with the y-dependence in (5.23a). See Section 1.2.3 of Keener and Sneyd [47] for a derivation of such a rate from a reaction scheme analogous to (5.28).

5.8 Exercises

Beyond the core exercises, there is a subsection of anticipatory exercises that introduce models we will use as examples below.

[16] In Section 4.4.5 our argument was based on nullclines. See Exercise 12 for an analytical treatment of this behavior.

5.8.1 Core Exercises

The core exercises have the following purposes:

To deal with unfinished business	1
To illustrate alternative scalings	2, 3
To practice scaling on new problems	4–6
To complete an idea from the Pearls	7
To keep a geeky tradition alive	8

1. For each of the examples in Sections 5.4–5.7, verify the units given in the corresponding table and check the claims about dimensionless parameters.

2. (a) Regarding the chemostat equations (5.19), determine what scaling for $\hat{x}, \hat{y}, \hat{t}$ produces the simplified equations

$$\begin{aligned} \frac{dx}{dt} &= \kappa \frac{y}{1+y} x - x, \\ \frac{dy}{dt} &= -\frac{y}{1+y} x - y + \mu. \end{aligned}$$

   (b) Express κ, μ in terms of the original dimensional parameters A, B, C, K, V, r, verify that κ, μ are dimensionless, and interpret them in words.

   Discussion: Edelstein-Keshet [19], for example, analyzes this form of the scaled equations. There isn't a compelling case for choosing between the scalings. One minor consideration in favor of the parameters (5.20) is that they are easily controlled by the experimentalist through varying the flow rate r and the feed concentration C.

3. Determine the scaling that reduces the Lotka–Volterra equations (5.4) to

$$\begin{aligned} dx/dt &= x - xy, \\ dy/dt &= xy - \rho y. \end{aligned} \tag{5.37}$$

   Discussion: This scaling makes the coefficients of three of the four terms in the equations equal to unity. In our view, this minor simplification is less important than fixing the location of the equilibrium, which was the basis for (5.8).

4. Find dimensionless parameters and scale the system

$$\begin{aligned} \frac{d\hat{x}}{d\hat{t}} &= \frac{R}{K^4 + \hat{y}^4} - D\hat{x}, \\ \frac{d\hat{y}}{d\hat{t}} &= \frac{R}{K^4 + \hat{x}^4} - D\hat{y}. \end{aligned}$$

   Both $\hat{x}$ and $\hat{y}$ have units of molarity.

Advice: If you are already comfortable with scaling, skip this exercise and proceed to more interesting scaling problems. In this problem we recommend scaling time based on the decay coefficient D and scaling $\hat{x}, \hat{y}$ based on the saturation constant K.

5. (a) Regarding (5.1), which describes a bead sliding on a rotating loop, make a table of variables and parameters, together with their dimensions.

   (b) Show that

$$\beta = \frac{b}{m\sqrt{\ell g}}, \qquad \mu = \frac{\ell\omega^2}{g}$$

   form a basis for the dimensionless parameters that may be constructed from the dimensional parameters in the problem.

   Remark: Of course, β is a dimensionless friction coefficient; the other parameter μ, a dimensionless measure of rotation speed, has a more specific interpretation as the ratio between centrifugal force and gravity.

   (c) Choose a time scale[17] $t = \hat{t}/T$ to transform (5.1) to

$$\frac{d^2x}{dt^2} = -\beta\frac{dx}{dt} - \sin x + \mu \sin x \cos x, \tag{5.38}$$

   where β, μ are given by the above.

   Discussion: You might be tempted to use ω to scale time, i.e., to let $t = \omega\hat{t}$. This choice leads to the equation

$$\frac{d^2x}{dt^2} = -\frac{\beta}{\sqrt{\mu}}\frac{dx}{dt} - \frac{1}{\mu}\sin x + \sin x \cos x. \tag{5.39}$$

   We find this scaling less transparent. For one reason, it seems more natural to scale time based on what is fixed, gravity, than on what can be varied, the rotation speed. This is particularly true in Chapter 8 when we study how the behavior of solutions of this ODE changes as μ varies, shifting the balance between centrifugal force and gravity.

6. *Introduction:* In Exercise 4.3 we considered nondimensionalized equations for the evolution of two interacting species. Those equations may be derived by scaling a dimensional system

$$\begin{aligned} d\hat{x}/dt &= r_1\hat{x}(1 - \hat{x}/K_1 - B\hat{y}), \\ d\hat{y}/dt &= r_2\hat{y}(1 - \hat{y}/K_2 - C\hat{x}), \end{aligned} \tag{5.40}$$

   where r_1, r_2, K_1, K_2, B, C are positive constants.

[17] The angle x is already dimensionless. We prefer not to scale x, because that would mess up the trig functions.

(a) Make a table of units of the variables and parameters in this equation. Interpret the meaning of the parameters in your own words.

(b) Find all dimensionless combinations of the parameters.

(c) Scale (5.40) to obtain the nondimensionalized equations of Exercise 4.3.

7. Verify equations (5.50) and (5.51) in the Pearls.

8. Find the numerical value of the speed of light in furlongs per fortnight.

Discussion: You probably already know this, but just in case: a furlong is one-eighth of a mile, a term used in horse racing, and a fortnight is two weeks, a term more common in Victorian times.

As undergraduates, when punchy from studying during exam week, we distracted ourselves by converting various physical constants into the firkin–furlong–fortnight system of units. (The firkin is a unit of volume in the wine trade, but we made it into a unit of mass whose value is the mass of that quantity of wine.) Care to make the conversion for Planck's constant?

5.8.2 Anticipatory Exercises

9. *Introduction:* This exercise on the continuous stirred-tank reactor (CSTR), which provides the best illustration of some bifurcation phenomena in Chapter 8, is definitely worth doing. The system is described by the equations

$$\begin{aligned} d\hat{c}/d\hat{t} &= \frac{r}{V}(C_f - \hat{c}) - R\,\hat{c}\,e^{\hat{T}/\Delta}, \\ d\hat{T}/d\hat{t} &= \frac{r}{V}(T_f - \hat{T}) + H\,R\,\hat{c}\,e^{\hat{T}/\Delta}, \end{aligned}$$

where $r, V, C_f, R, \Delta, T_f, H$ are constants, all positive except possibly T_f. In Chapter 8 these equations will be a rich source of bifurcation phenomena. The CSTR is a flow problem, as in Section 5.5.1, with the complication that the chemical flowing into and out of the tank undergoes an exothermic reaction. *Exothermic* means that the reaction releases heat as the chemical is consumed.

The state of the system is specified by $\hat{c}$ and $\hat{T}$, the concentration in the tank and its temperature, respectively. As in Section 5.5.1, r is the flow rate into and out of the tank, and V is the volume of the tank; C_f and T_f are the concentration and temperature of the inflow, respectively; R is a rate constant for the reaction at temperature zero, but the rate increases exponentially with temperature; H quantifies the temperature rise due to heat released by the reaction.

Two effects are added together in the concentration equation: the first term represents the change in concentration as the result of flow, as in (5.18), and the second represents the consumption of the chemical in the reaction. The temperature equation has the same two effects, with the difference that the reaction term increases temperature, so this term has a plus sign.

(a) The units of $\hat{c}$ and $\hat{T}$ are mass per unit volume and degrees, respectively. Make a table of the units of all seven parameters in the CSTR equations.

(b) Find a basis for the dimensionless constants that may be constructed from these parameters.

(c) Reduce the equations to the simplified form

$$\begin{aligned} dx/dt &= \mu(1-x) - xe^y, \\ dy/dt &= -\mu y + \sigma x e^y. \end{aligned}$$

Express μ and σ in terms of the dimensional parameters, verify that they are dimensionless, and interpret them in words.

Hint: The scalings for $\hat{c}$ and $\hat{t}$ are straightforward; you will find that $x = \hat{c}/C_f$ is the appropriate scaling for the concentration. However, for temperature, you need to include a translation along with a scale factor; i.e., assume

$$y = (\hat{T} - T_f)/\Delta.$$

Translation is natural in this case, because there is nothing special about the temperature "zero" in the CSTR system.

There are only two dimensionless constants in the reduced equations, which is fewer than the number of dimensionless constants you were able to construct in Part (b). This reduction is a result of the extra flexibility allowed by translating the temperature variable.

10. *Introduction:* The Rosenzweig–MacArthur model, another predator–prey system, is described by the system

$$\begin{aligned} \frac{d\hat{x}}{d\hat{t}} &= A\hat{x} - E\hat{x}^2 - B\frac{\hat{x}\hat{y}}{1+S\hat{x}}, \\ \frac{d\hat{y}}{d\hat{t}} &= C\frac{\hat{x}\hat{y}}{1+S\hat{x}} - D\hat{y}. \end{aligned}$$

If $\hat{x}$ is small (i.e., $\hat{x} \ll S^{-1}$), this system is equivalent to the Lotka–Volterra system with logistic growth of the prey. However, in the present equations the predation rate saturates at a maximum of $(B/S)\hat{y}$ when the prey are plentiful. This Michaelis–Menten-type of saturation also occurred in the chemostat, (5.19).

(a) Make a table of the parameters in these equations and their units.

(b) Find a basis for the dimensionless constants that can be formed from these parameters.

(c) Scale the equations to the form

$$\begin{aligned} \frac{dx}{dt} &= x(1-x) - \frac{xy}{1+\sigma x}, \\ \frac{dy}{dt} &= \kappa\frac{xy}{1+\sigma x} - \rho y. \end{aligned}$$

Express σ, κ, ρ in terms of the dimensional parameters, verify that they are dimensionless, and interpret them in words.

5.8.3 PHD Exercises

11. *Introduction:* The *FitzHugh–Nagumo* model refers to a system of the form

$$\begin{array}{lrcl} \text{(a)} & \eta\, d\hat{x}/d\hat{t} & = & \Gamma(\hat{x}) + A\hat{y}, \\ \text{(b)} & d\hat{y}/d\hat{t} & = & B\hat{x} + C\hat{y} + D, \end{array} \tag{5.41}$$

subject to the following assumptions: (i) $\Gamma(\hat{x}) = \Gamma_3\hat{x}^3 + \Gamma_2\hat{x}^2 + \Gamma_1\hat{x} + \Gamma_0$ is a *nonmonotonic* cubic with $\Gamma_3 < 0$ as sketched in Figure 5.7, (ii) A, B, C, D are constants with $AB < 0$, and (iii) η is a small positive constant. This model was proposed by FitzHugh [27] as a simplification of the Hodgkin–Huxley equations describing the conduction of nerve impulses [41]. (Nagumo [59] later constructed a circuit to simulate a nerve axon.) It might be stretching a point to assign units to the variables in this mathematical model, so we omit this part of scaling. A generalization of the model will figure in Chapters 8 and 9.

(a) Show that scaling may be used to simplify every such system to the form

$$\begin{array}{rcl} \varepsilon\, dx/dt & = & x(1 - x^2) - y + I, \\ dy/dt & = & x - \gamma y, \end{array} \tag{5.42}$$

where ε, I, γ are constants.

Hint: Equations (5.41) contain nine constants, but effectively there are only eight, because only the ratios of coefficients in (5.41a) matter. Let

$$x = \frac{\hat{x} + a}{X}, \qquad y = \frac{\hat{y} + b}{Y}, \qquad t = \frac{\hat{t}}{T},$$

which gives you five constants to play with.

- Choose a to annihilate the quadratic term of Γ.
- Choose $X > 0$ to make the coefficients of the cubic and linear terms in Γ equal in magnitude and opposite in sign. *(Why can't they have the same sign?)*

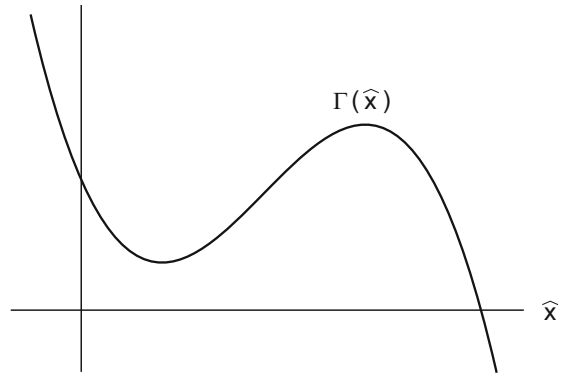

Figure 5.7: *Nonmonotone cubic of the type in the FitzHugh–Nagumo equation* (5.41)*, Exercise 11.*

The first equation now may be written as

$$\frac{\eta}{\overline{\Gamma}_3}\frac{dx}{d\hat{t}} = x(1-x^2) + \overline{A}\,\hat{y} + \overline{\Gamma}_0, \tag{5.43}$$

where the bars indicate some modification of the original constants and $\overline{\Gamma}_3 > 0$. (*Why is* $\overline{\Gamma}_3$ *positive?*)

- Choose $YT = -1/\overline{A}$ to make the coefficient of y in (5.43) equal to -1.
- Preserving YT from the previous step, choose $T > 0$ (and hence Y) to make the coefficient of x in the second equation equal to unity.
- Choose b to annihilate the constant term in the second equation.

The dedicated reader may wish to calculate explicitly the choices for a, b, X, Y, and T, and to identify the constants ε, I, γ.

(b) If $\gamma = 0$, show that a solution of the scaled system satisfies

$$\varepsilon\frac{d^2x}{dt^2} = (1-3x^2)\frac{dx}{dt} - x.$$

Without worrying about the size of ε, scale this equation to obtain the van der Pol equation, $z'' + \beta(z^2-1)z' + z = 0$. Express β in terms of ε.

12. *Introduction:* Consider an IVP for the equations of Michaelis–Menten kinetics (5.35), say

$$\begin{array}{llcll} (a) & dx/dt & = & -x(1-y)+y, & x(0)=b, \\ (b) & \varepsilon dy/dt & = & x(1-y)-(1+\kappa)y, & y(0)=0. \end{array} \tag{5.44}$$

In Section 4.4.5 we argued from nullclines that after a brief transient, the solution of (5.44) was well approximated by the scalar ODE (4.37) that results from letting the fast y-equation proceed to equilibrium; i.e., x solves the IVP

$$\frac{dx}{dt} = -\frac{\kappa x}{x+1+\kappa}, \qquad x(0) = b, \tag{5.45}$$

and y is slaved to x:

$$y = \frac{x}{x+1+\kappa}. \tag{5.46}$$

The present exercise shows how scaling may be used to give an analytical handle on this transient. Consider the scaled time $\tau = t/\varepsilon$, whose greatly expanded scale is appropriate for following the rapid evolution during the transient. With respect to τ, (5.44) becomes

$$\begin{array}{llcll} (a) & dx/d\tau & = & \varepsilon(-x(1-y)+y), & x(0)=b, \\ (b) & dy/d\tau & = & x(1-y)-(1+\kappa)y, & y(0)=0. \end{array} \tag{5.47}$$

Suppose the solution of (5.47) is expanded in a power series in ε, say

$$x(\tau,\varepsilon) = x_0(\tau) + \varepsilon x_1(\tau) + \cdots,$$

and similarly for $y(\tau, \varepsilon)$. Argue that $x_0(\tau) \equiv b$, derive and solve an IVP for $y_0(\tau)$, and show that

$$\lim_{\tau \to \infty} y_0(\tau) = \frac{b}{b + 1 + \kappa}, \tag{5.48}$$

consistent with (5.46).

Discussion: We have two separate approximations for the solution of (5.44), on different time scales: the above transient, called the *inner solution*, and the the fast-equation-to-equilibrium approximation (5.45), called the *outer solution*. Neither approximation is uniformly accurate for all times. Intuitively, they ought to agree in the "overlap region" where $\tau = t/\varepsilon$ may be large, say $\tau \sim \varepsilon^{-1/2}$, even though t is still small. The *method of matched asymptotic expansions* (see Chapter 2 of [43] or Chapters 6 and 7 of [98]) develops these ideas systematically. Both the inner and outer solutions are expanded in power series in ε, and the coefficients of each power of ε in the inner and outer solutions are required to agree in their appropriate limits. This more elaborate procedure has the obvious benefit of giving greater accuracy. It also has a nonobvious benefit: these calculations are often very tricky, and success in matching to higher order builds confidence that one has not gone astray.

5.9 Pearls of Wisdom

5.9.1 Making Scaling Work for You

Students often complain that scaling is more art than science, and we can't really disagree with this complaint. Effective scaling comes only with experience, and getting experience invariably means making errors in some of your trials. We hope that the remarks in this section can at least reduce your grief as you learn to use this collection of techniques. Additional perspective on this topic is provided in Chapter 4 of [98].

We hesitate to call any scaling "wrong," but some are definitely less informative, especially when you are seeking an approximation to exploit a large or small parameter in the equations. Let's illustrate this for a linear spring–mass system

$$m\frac{d^2\hat{x}}{d\hat{t}^2} + b\frac{d\hat{x}}{d\hat{t}} + k\hat{x} = 0. \tag{5.49}$$

Precisely one dimensionless parameter, say $\beta = b/\sqrt{mk}$, can be formed from the parameters in this equation. Since (5.49) is linear, changes in the x-scale have no effect. (*Check this!*) Therefore, we shall simply omit the hat on x and consider scaling only time.

Three scalings that give dimensionless time are listed in Table 5.8, along with the resulting scaled equation. In the scaled equations, the coefficients of two of the three terms in the ODE are equal to one another (and equal to unity). Which scaling is appropriate depends on context.

Scaling	Scaled equation
$\tau_1 = (b/m)\hat{t}$	$d^2x/d\tau_1^2 + dx/d\tau_1 + \beta^{-2}x = 0$
$\tau_2 = (\sqrt{k/m})\hat{t}$	$d^2x/d\tau_2^2 + \beta dx/d\tau_2 + x = 0$
$\tau_3 = (k/b)\hat{t}$	$\beta^{-2}d^2x/d\tau_3^2 + dx/d\tau_3 + x = 0$

Table 5.8: *Scalings of the spring-mass system (5.49), where* $\beta = b/\sqrt{mk}$.

Context 1: small β. Physically, the inequality $\beta \ll 1$ could mean that friction is small. The dominant phenomenon in this context is that solutions of (5.49) oscillate within a decaying envelope set by friction. In this case, the τ_2-scaling is definitely the most useful, because such oscillatory behavior is evident in the scaled equation $d^2x/d\tau_2^2 + \beta dx/d\tau_2 + x = 0$. By contrast, it would be awkward, at best, to extract this behavior from either of the other scalings, in which the term proportional to β^{-2} dominates the other two terms. Moreover, here's another consideration that points to τ_2 as the natural time scale: since (5.49) is linear, it admits exponential solutions $e^{\lambda\hat{t}}$. In Exercise 7 we ask you to show that when $\beta \ll 1$, the roots are

$$\lambda = \pm i\sqrt{\frac{k}{m}} + \mathcal{O}(\beta), \tag{5.50}$$

where the $\mathcal{O}(\beta)$ term contains both real and imaginary parts. In other words, to leading order, (the absolute value of) the exponent $\lambda\hat{t}$ equals τ_2, the preferred time scale.

Context 2: large β. In physical terms, $\beta \gg 1$ could indicate that m is small, that k is small, that b is large, or any combination of these; the concept of large or small is meaningful only when applied to dimensionless quantities. When β is large, the τ_2 time scale is sterile. To explore alternatives, let's look for exponential solutions $e^{\lambda\hat{t}}$ of (5.49). In Exercise 7 we ask you to show that when $\beta \gg 1$, the roots λ are given by

$$-\frac{b}{m} + \mathcal{O}(\beta^{-1}) \qquad \text{and} \qquad -\frac{k}{b} + \mathcal{O}(\beta^{-1}). \tag{5.51}$$

In other words, to leading order, $|\lambda\hat{t}| = \tau_1$ or τ_3. Both scalings have their uses.

In the τ_3-scaled ODE, the first-order and zeroth-order terms dominate the second-order term, $\beta^{-2}d^2x/d\tau_3^2$. Solutions of the equation with the second-order derivative neglected[18] decay like $e^{-\tau_3}$. This scaling captures the long-term decay of the system that follows the initial transient.

[18] If (5.49) were written as a first-order system, it would be a fast–slow system, and neglecting this second-order derivative would be the approximation of letting the fast equation go to equilibrium.

Note that $\tau_1 = \beta^2\tau_3$; i.e., if $\beta \gg 1$, then τ_1 increases very rapidly. The τ_1-scaling is appropriate for examining in detail the structure of the transient (as is done in Exercise 12 for a different problem). For an equation as simple as (5.49) there is little reward in focusing so intensely on this behavior, but in more complicated problems, much can be learned from such analysis. It would take a semester's course in asymptotics to fully support this apparently innocuous remark. (At a lower level of commitment, see Chapter 2 of [43] or Chapters 6 and 7 of [98].) The discussion in Section 7.6.4 provides a hint of the value in studying rapid transients on an expanded time scale.

The *principle of dominant balance* expresses the lessons to be drawn from this section. It articulates two desiderata in choosing a scaling for an equation:

- Every term in the equation should be on the order of unity or smaller, with at least two terms actually having this maximal size.
- All scaled variables should be on the order of unity.

In particular, the second desideratum played a key role in the scalings in Sections 5.6 and 5.7.

Even when no order-of-magnitude issues are involved, you may have to choose between different scalings. Two simple examples of this are given in Exercises 2 and 3, and here is another: Consider the Lotka–Volterra equations when there is a logistic limit to the growth of the prey; i.e., (5.4) is modified to

$$\begin{aligned} d\hat{x}/d\hat{t} &= A\hat{x} - E\hat{x}^2 - B\hat{x}\hat{y}, \\ d\hat{y}/d\hat{t} &= C\hat{x}\hat{y} - D\hat{y}. \end{aligned} \tag{5.52}$$

If we use the same scales as in deriving (5.8), i.e., $x = C\hat{x}/D$, $y = B\hat{y}/A$, $t = A\hat{t}$, then we get the scaled equations

$$\begin{aligned} dx/dt &= x(1 - x/K) - xy, \\ dy/dt &= \rho(xy - y), \end{aligned} \tag{5.53}$$

where $K = AC/DE$ is the nondimensional carrying capacity[19] of the environment and $\rho = D/A$. On the other hand, note that the *dimensional* carrying capacity in (5.52) is $\hat{x} = A/E$. This quantity also provides a natural scale for $\hat{x}$, and if we use it

[19]For the most part, we use Greek letters for dimensionless constants, but K deviates from this convention. However, invoking the pseudo-justification that too much consistency can be oppressive, we stick with K.

to define $x = E\hat{x}/A$ (while keeping the same scalings for y, t), we get the alternative scaled equations

$$\begin{aligned} dx/dt &= x(1-x) - xy, \\ dy/dt &= \kappa xy - \rho y, \end{aligned} \tag{5.54}$$

where $\kappa = C/E$. The parameter κ may be interpreted as the "efficiency with which the predators convert the prey into their own biomass."

The first scaling is more appropriate in studying how a logistic limit on growth changes the behavior of the Lotka–Volterra equations (cf. Section 6.5.3), because in (5.53), the limit $K \to \infty$ is nonsingular. By contrast, the second scaling underlies what is proposed in Exercise 10 on the Rosenzweig–MacArthur model, and this scaling will be more convenient in the bifurcation analysis of this model in Chapter 8. In conclusion, the best choice for scaling depends on the intended application (as well as on personal taste).

5.9.2 A Nod to Scientific Literacy

First, let's define molarity. The term *mole* specifies a quantity of a substance, more precisely, a number of grams of it equal to the molecular weight of the molecules of which the substance is composed. Thus, for example, a mole of glucose, $C_6H_{12}O_6$, equals $6 \times 12 + 12 \times 1 + 6 \times 16 = 180$ grams.[20] A mole of a substance contains a precise number of molecules of it; specifically, it contains N_a molecules, where $N_a \approx 6.02 \times 10^{23}$ is Avogadro's number. In Sel'kov's and other models in which a variable has units of molarity, the chemicals are dissolved in an aqueous solution, and molarity specifies the number of moles per liter of solution.

Exercise 9 involves a chemical reaction that proceeds at a temperature-dependent rate. A more realistic form for such temperature dependence is given by the *Arrhenius kinetics*, a rate proportional to $e^{-T_a/T}$, where T is the absolute temperature and the constant T_a is the activation energy of the reaction converted to a temperature (by means of Boltzmann's constant). The rate-dependence in Exercise 9 provides a tolerable approximation to this formula when, as usually is the case, $T \ll T_a$.

[20] Because of heavier isotopes, the molecular weight of carbon is a little greater than 12, and likewise for hydrogen and oxygen. Thus the molecular weight of glucose is also a little greater than 180, but this integer approximation is adequate for our purposes.

Chapter 6

Trajectories Near Equilibria

In this chapter we relate the flow of an ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ near an equilibrium $\mathbf{b}_*$ to the flow of the *linearization*, by which we mean the equation $\mathbf{w}' = A\mathbf{w}$, where $A = \mathbf{DF}(\mathbf{b}_*)$. There are two main theoretical results. (i) In Section 6.1, we assume $\Re\lambda_j(A) < 0$, which guarantees that all solutions of the linearization converge to the equilibrium; Theorem 6.1.1 shows that under this hypothesis, the full equation shares a version of this behavior, which is called *asymptotic stability*. (ii) In Section 6.6, the stable-manifold theorem (Theorem 6.6.1) characterizes the behavior of solutions when the Jacobian $\mathbf{DF}(\mathbf{b}_*)$ has eigenvalues with both positive and negative real parts.

Other sections mostly supplement or apply results of the above two sections. Specifically, Section 6.2 introduces some of the terminology growing out of Theorem 6.1.1; Sections 6.3 and 6.4 analyze two scientifically interesting applications of Theorem 6.1.1; and Section 6.7 shows how Theorems 6.1.1 and 6.6.1 are useful in sketching trajectories of ODEs.

Section 6.5 pushes in a different direction. It introduces Lyapunov functions, which are another technique for proving stability, a technique not based on linearization.

There are two appendices. The first explores how the stable manifold theorem is proved, and the second informally introduces a generalization of this theorem.

From here on in this book, unless otherwise stated, we shall assume in the generic ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ that the function $\mathbf{F}$ is $\mathcal{C}^1$. (Indeed, otherwise linearization wouldn't make sense.)

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8_6

6.1 Stability of Equilibria

We recall that a point $\mathbf{b}_* \in \mathbb{R}^d$ is called an *equilibrium* for an ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ if at this point, $\mathbf{F}(\mathbf{b}_*) = \mathbf{0}$. For a linear homogeneous constant-coefficient equation $\mathbf{x}' = A\mathbf{x}$, the origin $\mathbf{b}_* = \mathbf{0}$ is always an equilibrium, and it is the only equilibrium if A is invertible. For nonlinear equations, it is more typical to have multiple equilibria. For example, already in Section 1.6 we saw that the augmented Lotka–Volterra system

$$\begin{aligned} x' &= x\left(\frac{x-\varepsilon}{x+\varepsilon}\right)\left(1-\frac{x}{K}\right) - xy, \\ y' &= \rho(xy - y), \end{aligned}$$

has four equilibria in the physical domain if $0 < \varepsilon < 1 < K$. Similarly, the 2×2 system derived from Duffing's equation

$$\begin{aligned} x' &= y, \\ y' &= x - x^3 - \beta y, \end{aligned} \tag{6.1}$$

has equilibrium points $(0, 0)$ and $(\pm 1, 0)$.

6.1.1 The Main Theorem

In Theorem 2.4.1 we showed that for the linear system $\mathbf{x}' = A\mathbf{x}$, if the eigenvalues of the coefficient matrix satisfy

$$\Re\lambda_j(A) < 0, \qquad j = 1, 2, \ldots, d,$$

then every solution $\mathbf{x}(t)$ of the equation decays to zero as $t \to \infty$. The following theorem, a major and quite beautiful result, asserts that one may deduce similar stability behavior for solutions of a nonlinear equation $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ *near* an equilibrium point $\mathbf{b}_*$, provided the eigenvalues of $\mathbf{DF}(\mathbf{b}_*)$, the differential of $\mathbf{F}$ at the equilibrium, satisfy this condition. Here and below we shall abbreviate $\mathbf{DF}(\mathbf{b}_*)$ to $\mathbf{DF}_*$.

Theorem 6.1.1. *Suppose $\mathbf{b}_*$ is an equilibrium point for $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, where $\mathbf{F} \in \mathcal{C}^1(\mathcal{U})$ with $\mathcal{U} \subset \mathbb{R}^d$, and assume that*

$$\Re\lambda_j(\mathbf{DF}_*) < 0, \qquad j = 1, 2, \ldots, d. \tag{6.2}$$

Then there is a neighborhood $\mathcal{V}$ of $\mathbf{b}_*$ in $\mathbb{R}^d$ such that for any initial data $\mathbf{b} \in \mathcal{V}$, the IVP

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0) = \mathbf{b} \tag{6.3}$$

has a solution for all $t \geq 0$, and moreover, $\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{b}_*$.

Remarks. (i) The theorem includes the linear case, because if $\mathbf{F}(\mathbf{x}) = A\mathbf{x}$, then at the origin, $\mathbf{DF}(\mathbf{0}) = A$. (Indeed, $\mathbf{DF}(\mathbf{x}) = A$ at every point $\mathbf{x}$.) *(ii)* In one dimension $\mathbf{DF}_*$ is a scalar, and in this case, Theorem 6.1.1 was anticipated by Problem 1.15.

It may be instructive to compare Theorem 6.1.1 with what may be deduced under the same hypotheses from Theorem 4.6.1, i.e., that $(\forall \eta > 0)(\forall T < \infty)$ there is a neighborhood $\mathcal{V}$ such that if $\mathbf{b} \in \mathcal{V}$, then $\varphi(t, \mathbf{b})$ exists for all $t \le T$ and

$$|\boldsymbol{\varphi}(t, \mathbf{b}) - \mathbf{b}_* - e^{At}(\mathbf{b} - \mathbf{b}_*)| \le \eta |\mathbf{b} - \mathbf{b}_*|, \qquad 0 \le t \le T.$$

Of course, since $\Re\lambda_j(A) < 0$, the term $e^{At}(\mathbf{b} - \mathbf{b}_*)$ tends to zero as $t \to \infty$, so for large time, effectively we have an estimate for $|\boldsymbol{\varphi}(t, \mathbf{b}) - \mathbf{b}_*|$. The new information provided by Theorem 6.1.1 is twofold: (i) the IVP (6.3) has a solution all the way to infinite time, and (ii) $|\boldsymbol{\varphi}(t, \mathbf{b}) - \mathbf{b}_*|$ decays to zero as $t \to \infty$.

Proof of Theorem 6.1.1. Making an appropriate translation in $\mathbb{R}^d$, we may assume without loss of generality that the equilibrium $\mathbf{b}_*$ is located at the origin, i.e., $\mathbf{F}(\mathbf{0}) = \mathbf{0}$. Expand $\mathbf{F}$ at the origin: $\mathbf{F}(\mathbf{x}) = A\mathbf{x} + \mathbf{r}(\mathbf{x})$, where (i) the constant term is missing, since $\mathbf{F}(\mathbf{0})$ vanishes, (ii) in the linear term, $A = \mathbf{DF}(\mathbf{0})$, and (iii) in the order notation of Section 4.6.4, the remainder $\mathbf{r}(\mathbf{x})$ is $o(|\mathbf{x}|)$. It follows by Proposition 2.4.2 that there are constants K, ε, where $\varepsilon > 0$, such that

$$\|e^{At}\| \le K e^{-\varepsilon t}, \qquad t \ge 0; \tag{6.4}$$

of course $K \ge 1$. Choose a positive η such that $\eta < \varepsilon/K$. Since $\mathbf{r}(\mathbf{x}) = o(|\mathbf{x}|)$, there is a $\delta > 0$ such that if $|\mathbf{x}| \le \delta$, then $|\mathbf{r}(\mathbf{x})| \le \eta |\mathbf{x}|$.

Let $\mathcal{V} = \{\mathbf{b} \in \mathbb{R}^d : |\mathbf{b}| < \delta/K\} \subset B(\mathbf{0}, \delta)$. If the IVP (6.3) is not solvable on $[0, \infty)$, then by Theorem 4.1.2, the solution must leave the ball $\overline{B(\mathbf{0}, \delta)}$. We seek to derive a contradiction by assuming that there is a time $t_* > 0$ such that $|\mathbf{x}(t)| < \delta$ for $t < t_*$, while $|\mathbf{x}(t_*)| = \delta$.

With the same ε as in (6.4), let $g(t) = e^{\varepsilon t}|\mathbf{x}(t)|$, which we will estimate with Gronwall's inequality. We rewrite the IVP (6.3) as

$$\mathbf{x}' - A\mathbf{x} = \mathbf{r}(\mathbf{x}), \qquad \mathbf{x}(0) = \mathbf{b} \tag{6.5}$$

and interpret (6.5) as a linear equation with constant coefficients with the inhomogeneous term $\mathbf{r}(\mathbf{x}(t))$. As we saw in (2.51), a solution of (6.5) satisfies the integral equation[1]

$$\mathbf{x}(t) = e^{At}\mathbf{b} + \int_0^t e^{(t-s)A}\, \mathbf{r}(\mathbf{x}(s))\, ds. \tag{6.6}$$

Multiplying by $e^{\varepsilon t}$, we obtain

$$g(t) \le e^{\varepsilon t}|e^{At}\mathbf{b}| + e^{\varepsilon t} \int_0^t |e^{(t-s)A}\, \mathbf{r}(\mathbf{x}(s))|\, ds.$$

[1] A similar integral equation, (3.16), arose in proving the existence theorem. The advantage of (6.6) over (3.16) is twofold: (i) the decay in (6.4) assists the convergence of the integral for large t, and (ii) in the integrand, $\mathbf{r}(\mathbf{x})$ is small when $\mathbf{x}$ is close to zero.

Applying (6.4) to bound the exponentials in each term, we conclude that

$$g(t) \le K|\mathbf{b}| + K\int_0^t e^{\varepsilon s}|\mathbf{r}(\mathbf{x}(s))|\,ds. \tag{6.7}$$

Now for $t \le t_*$, the second term of (6.7) satisfies

$$K\int_0^t e^{\varepsilon s}|\mathbf{r}(\mathbf{x}(s))|\,ds \le K\eta\int_0^t e^{\varepsilon s}|\mathbf{x}(s)|\,ds = K\eta\int_0^t g(s)\,ds.$$

Hence by Gronwall's lemma, $g(t) \le K|\mathbf{b}|e^{K\eta t}$ for $0 \le t \le t_*$. Recalling the definition of g, we conclude that

$$|\mathbf{x}(t)| = e^{-\varepsilon t}g(t) \le K|\mathbf{b}|e^{(K\eta-\varepsilon)t}. \tag{6.8}$$

In particular, since the exponential is decaying, we have $|\mathbf{x}(t_*)| \le K|\mathbf{b}| < \delta$, contradicting our assumption above, so the solution never leaves the ball of radius δ. Thus, the solution exists for all $t \ge 0$, and by (6.8), it tends to zero as $t \to \infty$. □

The following corollary of Theorem 6.1.1 makes the convergence to the equilibrium more quantitative. This result may be derived by exercising a little more care in the proof of Theorem 6.1.1, a task we ask you to complete in Exercise 1(a).

Corollary 6.1.2. *If in Theorem 6.1.1 the eigenvalues satisfy*

$$\Re\lambda_j(\mathbf{DF}_*) < -\varepsilon, \qquad j = 1, 2, \ldots, d, \tag{6.9}$$

where $\varepsilon > 0$, then $\mathcal{V}$ may be chosen with the property that there is a constant K_1 such that for all $\mathbf{b} \in \mathcal{V}$, the solution of the IVP satisfies

$$|\mathbf{x}(t) - \mathbf{b}_*| \le K_1 e^{-\varepsilon t}|\mathbf{b} - \mathbf{b}_*|, \qquad t \ge 0. \tag{6.10}$$

6.1.2 An Easy Application

Let's illustrate these ideas by applying them to Duffing's equation (6.1), which describes motion in the double-well potential $V(x) = -x^2/2 + x^4/4$. At the two minima of V we have

$$\mathbf{DF}(\pm 1, 0) = \mathbf{DF}_* = \begin{bmatrix} 0 & 1 \\ -2 & -\beta \end{bmatrix}.$$

To test whether eigenvalues have negative real parts, we compute

$$\det \mathbf{DF}_* = +2 > 0, \qquad \operatorname{tr} \mathbf{DF}_* = -\beta < 0,$$

where we have assumed $\beta > 0$, i.e., normal friction. Hence by Proposition 2.4.5, the eigenvalues of $\mathbf{DF}_*$ both have negative real parts, and so near the equilibria $(\pm 1, 0)$,

solutions of (6.1) behave as described in Theorem 6.1.1. More interesting examples will be studied below.

Incidentally, Theorem 6.1.1 does not readily extend to nonautonomous equations; Problem 13 in Chapter 3, with a bit of friction added to the ODE, illustrates the difficulties.

6.2 Terminology to Classify Equilibria

6.2.1 Terms Related to Theorem 6.1.1

An equilibrium $\mathbf{b}_*$ of a system $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ is called *Lyapunov stable* if for every neighborhood $\mathcal{V}$ of $\mathbf{b}$ in $\mathbb{R}^d$, there is a smaller neighborhood $\mathcal{V}_1$ such that if $\mathbf{b} \in \mathcal{V}_1$, then the IVP (6.3) is solvable for all positive times and moreover, $\mathbf{x}(t) \in \mathcal{V}$ for all $t \geq 0$. The equilibrium $\mathbf{b}_*$ is called *attracting* if there is some neighborhood $\mathcal{V}_*$ of $\mathbf{b}_*$ such that for all initial data in $\mathcal{V}_*$, the solution of (6.3) exists for all $t \geq 0$ and converges to $\mathbf{b}_*$; in symbols,

$$\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{b}_*. \tag{6.11}$$

The equilibrium $\mathbf{b}_*$ is called *asymptotically stable* if it is Lyapunov stable and attracting.

The results of Section 6.1 imply that $\mathbf{b}_*$ is asymptotically stable if condition (6.2) is satisfied. Theorem 6.1.1 as stated is not sufficient to prove this claim. We need the estimate of Corollary 6.1.2 to conclude that $\mathbf{b}_*$ is Lyapunov stable.

You may think that the language used in these definitions is excessively fussy. Here are two examples to show why such careful wording is required:

(i) Regarding the nested neighborhoods in the definition of Lyapunov stable, consider the following linear 2×2 system:

$$\mathbf{x}' = \begin{bmatrix} -\eta & -K^{-1} \\ K & -\eta \end{bmatrix} \mathbf{x} \tag{6.12}$$

where K is a large constant and $\eta > 0$ a small one. For every real C, the function

$$\mathbf{x}(t) = Ce^{-\eta t} \begin{bmatrix} \cos t \\ K \sin t \end{bmatrix}$$

is a solution of this system. Even though they spiral into the origin, these trajectories are very elongated, as shown in Figure 6.1(a). Suppose we are given a circular neighborhood $\mathcal{V} = \{\mathbf{b} \in \mathbb{R}^2 : |\mathbf{b}| < \varepsilon\}$ of the origin. To prove Lyapunov stability, we want to find another neighborhood, say circular $\mathcal{V}_1 = \{\mathbf{b} \in \mathbb{R}^2 : |\mathbf{b}| < \delta\}$, such that if $\mathbf{x}(0) \in \mathcal{V}_1$, then the trajectory remains confined to $\mathcal{V}$. We have to choose a much smaller radius $\delta < \varepsilon/K$ to achieve

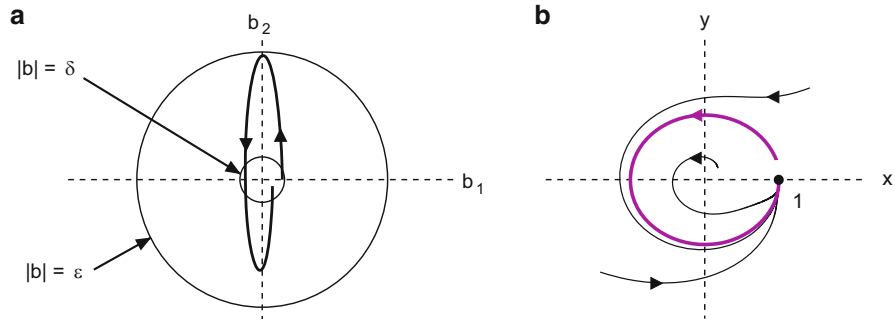

Figure 6.1: *Figures to clarify the definition of asymptotic stability. (a) An elongated spiral converging to the origin need not be contained in a circular neighborhood. (b) An unstable equilibrium may be attracting.*

this, because trajectories are so elongated. If we were smart enough to choose $\mathcal{V}_1$ with a perfect shape adjusted to the orbit, we could arrange that $x(t) \in \mathcal{V}_1$ for all $t \geq 0$. Although this is sometimes possible, it is inconvenient to insist on it in general.

(ii) You might think that if an equilibrium is attracting, then it surely is Lyapunov stable. This is *false*, as shown by the following 2×2 system, which we write in polar coordinates:

$$\begin{aligned} r' &= r - r^3, \\ \theta' &= 1 - \cos\theta. \end{aligned} \tag{6.13}$$

A few trajectories for this system are illustrated in Figure 6.1(b). As suggested by the figure, all nonzero solutions of this system converge to $(r, \theta) = (1, 0)$ as $t \to \infty$. However, a trajectory that starts at a point $(r, \theta) = (1, \varepsilon)$, where ε is positive (and as small as you like), proceeds all the way around the circle before it converges to $(1, 0)$; in particular, it leaves the ball of unit radius around $(1, 0)$.

Another bit of terminology: if $\mathbf{b}_*$ is an attracting equilibrium, the *basin of attraction* of $\mathbf{b}_*$ is the set of initial conditions $\mathbf{b}$ such that the solution of the IVP (6.3) converges to $\mathbf{b}_*$ as $t \to \infty$. For example, the basin of attraction of the equilibrium $x = 1$ of the logistic equation $x' = x - x^2$ is the interval $(0, \infty)$.

Unstable means the negation of Lyapunov stability; i.e., there is some neighborhood $\mathcal{V}$ of $\mathbf{b}_*$ in $\mathbb{R}^d$ such that for every smaller neighborhood $\mathcal{V}_1$, there are initial conditions $\mathbf{b} \in \mathcal{V}_1$ such that the solution to the IVP (6.3) leaves $\mathcal{V}$ at some positive time. The linearization provides a sufficient condition for such behavior:

Proposition 6.2.1. *If $\mathbf{b}_*$ is an equilibrium of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ and if $\Re\lambda_j(\mathbf{DF}_*) > 0$ for some j, then $\mathbf{b}_*$ is unstable.*

The proof of this seemingly obvious result is less straightforward than one might expect. Let's shoot down the natural first try at a proof. Suppose $\mathbf{v}$ is an eigenvector of $\mathbf{DF}_*$ with a real eigenvalue $\lambda > 0$, and let a neighborhood $\mathcal{V}$ of $\mathbf{b}_*$ be given. Consider solutions of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ with $\mathbf{x}(0) = \mathbf{b}_* + \varepsilon\mathbf{v}$, where $\varepsilon \ll 1$. We have the approximation from linearization

$$\mathbf{x}(t) \approx \mathbf{b}_* + \varepsilon e^{\lambda t}\mathbf{v}. \tag{6.14}$$

The growing exponential will eventually push the RHS of (6.14) outside $\mathcal{V}$, as needed to prove instability. *But* the smaller ε is, the longer we have to wait for $\mathbf{x}(t)$ to leave $\mathcal{V}$, and it's not obvious that (6.14) remains valid long enough to guarantee this.

For a direct proof of the proposition, see Exercise 18; for an indirect proof, see Section 6.10.4, where it is derived as a consequence of the stable manifold theorem.

As a simple application of the result, consider the equilibrium at the origin of Duffing's equation (6.1). We have $\det \mathbf{DF}(0,0) < 0$, so the eigenvalues of $\mathbf{DF}$ have opposite signs, one of them being positive, and thus the origin is unstable—no surprise here.

If at an equilibrium, we have only $\Re\lambda_j(\mathbf{DF}_*) \leq 0$, i.e., if the inequality (6.2) is not strict,[2] then no information can be deduced from the linearization. As a trivial example to justify this statement, consider the scalar ODE

$$x' = \pm x^3. \tag{6.15}$$

For either sign, the origin is an equilibrium, and the lone eigenvalue of $DF(0)$ vanishes. However, if the minus sign is chosen, the origin is asymptotically stable, while if the plus sign is chosen, it is unstable, and both behaviors are different from that of the linearization $w' = 0$, whose equilibrium is Lyapunov stable but not asymptotically stable. *(Check these claims!)*

Another borderline case occurs in Duffing's equation (6.1) if the friction coefficient β is equal to zero. The eigenvalues of $\mathbf{DF}$ at the equilibria $(\pm 1, 0)$ are pure imaginary, i.e., $\Re\lambda = 0$. In this case, the equilibria are Lyapunov stable but not asymptotically stable. This can easily be shown using energy as a Lyapunov function, a technique we will introduce in Section 6.5. (Cf. Exercise 11.)

6.2.2 Other Terms Based on Eigenvalues

We call an equilibrium $\mathbf{b}_*$ of an ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ *hyperbolic* if

$$\Re\lambda_j(\mathbf{DF}_*) \neq 0, \qquad j = 1, \ldots, d. \tag{6.16}$$

[2]For example, the equilibrium $(r, \theta) = (1, 0)$ of (6.13) suffers from this degeneracy.

This definition generalizes the usage in Section 2.8.3 for linear systems. We also extend the terminology of Section 2.5 to classify hyperbolic equilibria in two dimensions. Thus, an equilibrium $\mathbf{x}_* \in \mathbb{R}^2$ is called a *sink, source, saddle, node, or focus* if the eigenvalues of $\mathbf{DF}_*$ fit the corresponding description in Table 2.1. (The term "center" is not used, because if the eigenvalues of $\mathbf{DF}_*$ are pure imaginary, then $\mathbf{x}_*$ is not hyperbolic.)

Near a hyperbolic equilibrium $\mathbf{x}_*$, locally the flow of the full system resembles the flow of the linearized system $\mathbf{w}' = A\mathbf{w}$, where $A = \mathbf{DF}_*$. The two main results of this chapter, Theorems 6.1.1 and 6.6.1, support this claim. A precise correspondence between a nonlinear system and its linearization is articulated by the Hartman–Grobman Theorem in Section 6.9.2 of the Pearls. In particular, this result shows that the appropriate phase portrait of a linear system in Section 2.5 provides a reliable qualitative representation of the flow *in some neighborhood* of a hyperbolic equilibrium. (You might find reviewing Section 2.5 a useful investment at this point.)

6.2.3 Section 1.6 Revisited, Part I

Let's apply this classification of equilibria to the augmented Lotka–Volterra system in Section 1.6. This calculation explains how the ε, K-parameter set in Figure 1.9 gets divided into three regions.

We rewrite (1.41) as

$$\begin{array}{lrcl} (a) & x' & = & x\phi(x) - xy, \\ (b) & y' & = & \rho(xy - y), \end{array} \tag{6.17}$$

where

$$\phi(x) = \left(\frac{x-\varepsilon}{x+\varepsilon}\right)\left(1 - \frac{x}{K}\right) \tag{6.18}$$

and $0 < \varepsilon < \min\{K, 1\}$. Because we process pictures more readily than formulas, we have graphed $\phi(x)$ in Figure 6.2. As we showed in Section 1.6, this system has the four equilibria listed in Table 6.1. To classify these equilibria, we calculate that

$$\mathbf{DF} = \begin{bmatrix} x\phi'(x) + \phi(x) - y & -x \\ \rho y & \rho(x-1) \end{bmatrix}. \tag{6.19}$$

(i) *The extinction equilibrium:* Substituting $(x, y) = (0, 0)$ into (6.19), we obtain

$$\mathbf{DF}(0,0) = \begin{bmatrix} \phi(0) & 0 \\ 0 & -\rho \end{bmatrix}.$$

By Figure 6.2, $\phi(0) < 0$. Thus, the eigenvalues of $\mathbf{DF}$ are real and negative, as claimed in the table: this equilibrium is a stable node.

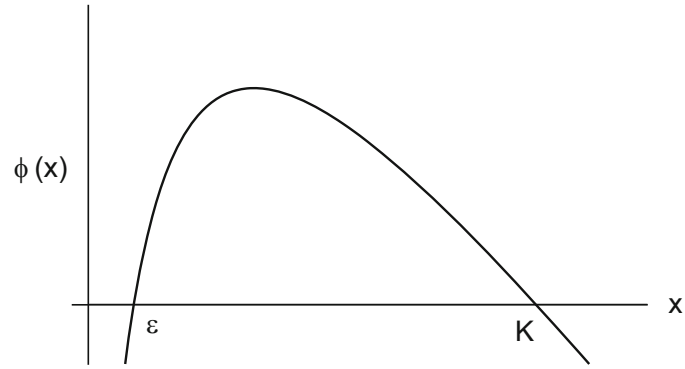

Figure 6.2: *Graph of the function $\phi(x)$ defined in* (6.18).

Equilibrium	Name	Type of equilibrium
$(0,0)$	Extinction equilibrium	Always a sink (and a node)
$(\varepsilon,0)$	Extinction threshold	Always a saddle
$(K,0)$	Prey-only equilibrium	If $K<1$: a sink (and a node)
		If $K>1$: a saddle
$(1,\phi(1))$	Coexistence equilibrium	If $K<1$: an unphysical saddle
		If $1<K<(1+2\varepsilon-\varepsilon^2)/2\varepsilon$: a sink
		If $(1+2\varepsilon-\varepsilon^2)/2\varepsilon<K$: a source

Table 6.1: *Classification of equilibria of the augmented Lotka–Volterra equations* (6.17).

(ii) *The extinction threshold:* Substituting $(\varepsilon, 0)$ into (6.19), we obtain

$$\mathbf{DF}(\varepsilon, 0) = \begin{bmatrix} \varepsilon\phi'(\varepsilon) & -\varepsilon \\ 0 & -\rho(1-\varepsilon) \end{bmatrix}.$$

By Figure 6.2, $\phi'(\varepsilon) > 0$. Thus, the eigenvalues of $\mathbf{DF}$ have opposite signs, as claimed: this equilibrium is a saddle.

(iii) *The prey-only equilibrium:* Substituting $(K, 0)$ into (6.19), we obtain

$$\mathbf{DF}(K, 0) = \begin{bmatrix} K\phi'(K) & -K \\ 0 & -\rho(1-K) \end{bmatrix}.$$

Thus, the eigenvalues of $\mathbf{DF}$ are real. By Figure 6.2, $\phi'(K) < 0$. Both eigenvalues of $\mathbf{DF}$ are negative if $K < 1$ (a stable node) but have opposite signs if $K > 1$ (a saddle). This calculation explains why, in Figure 1.9, the line $\{K = 1\}$ arises as a boundary between regions in parameter space.

(iv) *The coexistence equilibrium:* Substituting $(1, \phi(1))$ into (6.19), we obtain

$$\mathbf{DF}(1, \phi(1)) = \begin{bmatrix} \phi'(1) & -1 \\ \rho\phi(1) & 0 \end{bmatrix}.$$

Now

$$\det \mathbf{DF} = \rho\phi(1) = \rho\left(\frac{1-\varepsilon}{1+\varepsilon}\right)\left(1 - \frac{1}{K}\right).$$

If $K < 1$, then $\det \mathbf{DF} < 0$, so the equilibrium is a saddle; it is unphysical, because $y = \phi(1) < 0$. If $K > 1$, then $\det \mathbf{DF} > 0$, so the equilibrium is either a sink or a source. To determine which, we calculate

$$\operatorname{tr} \mathbf{DF} = \phi'(1) = \frac{2\varepsilon - (1 + 2\varepsilon - \varepsilon^2)/K}{(1+\varepsilon)^2}.$$

If $1 < K < (1 + 2\varepsilon - \varepsilon^2)/2\varepsilon$, then $\operatorname{tr} \mathbf{DF} < 0$, and the equilibrium is a sink; if $(1 + 2\varepsilon - \varepsilon^2)/2\varepsilon < K$, a source. This calculation explains why, in Figure 1.9, the curve $\{K = (1 + 2\varepsilon - \varepsilon^2)/2\varepsilon\}$ arises as a boundary between regions in parameter space. As you can calculate (and we will show in later chapters), these equilibria can be either nodes or foci; which case occurs depends on ρ as well as ε, K.

Some typical phase portraits for (6.17) will be plotted in Section 6.7.3.

	Slope x-nullcline larger	Slope y-nullcline larger
Signs of $\partial F_1/\partial y, \partial F_2/\partial y$ same	$\det \mathbf{DF}_* < 0$	$\det \mathbf{DF}_* > 0$
Signs of $\partial F_1/\partial y, \partial F_2/\partial y$ opposite	$\det \mathbf{DF}_* > 0$	$\det \mathbf{DF}_* < 0$

Table 6.2: *Types of equilibria and slopes of nullclines, assuming* $\partial F_k/\partial y \neq 0$.

6.2.4 Two-Dimensional Equilibria and Slopes of Nullclines

Here is a geometric fact that can sometimes shorten stability calculations for two-dimensional systems: At an equilibrium (x_*, y_*) of

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} F_1(x,y) \\ F_2(x,y) \end{bmatrix}, \tag{6.20}$$

if you know the slopes of the nullclines, you can distinguish between a saddle point (where $\det \mathbf{DF}_* < 0$) and a sink or source (where $\det \mathbf{DF}_* > 0$), as indicated in Table 6.2.

To prove this assertion, suppose that $\det \mathbf{DF}_* \neq 0$. Because the determinant is nonvanishing, (i) both gradients $\nabla F_j(x_*, y_*)$, $j = 1, 2$, are nonzero, so both nullclines are nonsingular curves, and (ii) the nullclines are not tangent. Note that the slope of the x-nullcline is the quotient $-\frac{\partial F_1}{\partial x}/\frac{\partial F_1}{\partial y}$, interpreted as ∞ if the denominator vanishes, and similarly for the slope of the y-nullcline. If either $\partial F_1/\partial y$ or $\partial F_2/\partial y$ vanishes, then one of the two terms in $\det \mathbf{DF}_*$ is zero, so the sign of $\det \mathbf{DF}_*$ may be determined from the signs of elements of $\mathbf{DF}_*$; thus, in the table we assume that $\partial F_1/\partial y \neq 0$, $\partial F_2/\partial y \neq 0$. Multiply and divide $\det \mathbf{DF}$ by the product $\partial F_1/\partial y \;\; \partial F_2/\partial y$ to deduce

$$\det \mathbf{DF} = (\partial F_1/\partial y)(\partial F_2/\partial y) \left\{ \frac{\partial F_1/\partial x}{\partial F_1/\partial y} - \frac{\partial F_2/\partial x}{\partial F_2/\partial y} \right\}. \tag{6.21}$$

Thus (6.21) represents $\det \mathbf{DF}_*$ as a factor times the difference in the slopes of the nullclines. With these observations you may verify the information given in Table 6.2. (Don't forget the minus sign in the formula for the slopes.)

6.3 Activator–Inhibitor Systems and the Turing Instability

As we saw in Section 5.6, in the limit $\kappa \to \infty$, the activator–inhibitor equations[3] reduce to

$$\begin{aligned} (a) \quad x' &= \sigma x^2/(1+y) - x, \\ (b) \quad y' &= \rho\,[x^2 - y], \end{aligned} \tag{6.22}$$

where σ and ρ are positive parameters. In the first subsection below we determine the equilibria of (6.22) and their stabilities; in the second, making a lovely application of Theorem 6.1.1, we present the Turing instability, in which two copies of (6.22) are coupled.

6.3.1 Equilibria of the Activator–Inhibitor System

Let's establish the following facts about the equilibria of (6.22):

- The origin $(0,0)$ is an asymptotically stable equilibrium of (6.22) for all parameter values.
- If $\sigma > 2$, then (6.22) has two nontrivial equilibria, say $\mathbf{P}_\pm = (x_\pm, y_\pm)$, where $x_\pm$ satisfies
$$x^2 - \sigma x + 1 = 0 \tag{6.23}$$
and $y_\pm = x_\pm^2$. (See also Figure 6.3.)
- When $\sigma > 2$, the equilibrium $\mathbf{P}_-$ is a saddle point.
- When $\sigma > 2$, the equilibrium $\mathbf{P}_+$ is a sink if $\rho > 1$ and a source if $\rho < 1$.

Proof of Point 1: It is obvious that the origin $(0,0)$ is an equilibrium of (6.22). To determine its stability, we calculate the Jacobian of the system (at an arbitrary point):

$$\mathbf{DF} = \begin{bmatrix} \dfrac{2\sigma x}{1+y} - 1 & -\dfrac{\sigma x^2}{(1+y)^2} \\[2ex] 2\rho x & -\rho \end{bmatrix}. \tag{6.24}$$

At the origin,

$$\mathbf{DF}_* = \begin{bmatrix} -1 & 0 \\ 0 & -\rho \end{bmatrix}, \tag{6.25}$$

so by Theorem 6.1.1 this equilibrium is asymptotically stable.

[3] In Section 4.4.2 we proved global existence for the activator–inhibitor equations assuming $\kappa < \infty$, and global existence for (6.22) was posed as Exercise 4.3(c).

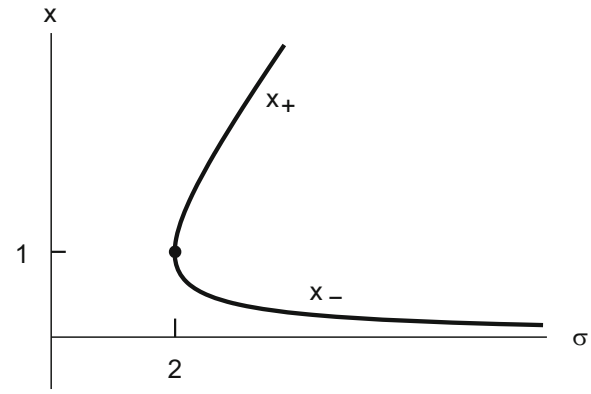

Figure 6.3: *Nonzero equilibria of the activator–inhibitor system* (6.22) *for* $\sigma \geq 2$.

Proof of Point 2: To investigate other equilibria of (6.22), we calculate nullclines. The RHS of (6.22a) vanishes if

$$\text{(a) } x = 0 \quad \text{or} \quad \text{(b) } y = \sigma x - 1. \tag{6.26}$$

The RHS of (6.22b) vanishes if $y = x^2$; substituting into (6.26b), we obtain (6.23). If $\sigma > 2$, then the roots of (6.23) are real, and (6.22) has two nontrivial equilibria, as claimed. Incidentally, the Jacobian (6.24) at these equilibria simplifies to

$$\mathbf{DF}_{\pm} = \begin{bmatrix} 1 & -1/\sigma \\ 2\rho x_{\pm} & -\rho \end{bmatrix}. \tag{6.27}$$

Proof of Point 3: As shown in Section 6.2.4, we may determine the sign of $\det \mathbf{DF}_*$ at an equilibrium by comparing the slopes of the two nullclines. Note from (6.27) that $\partial F_1/\partial y$ and $\partial F_2/\partial y$ always have the same sign, both negative. As we see in Figure 6.4, the x-nullcline has larger slope at $\mathbf{P}_-$ than the y-nullcline. Thus, by Table 6.2, this equilibrium is a saddle point.[4]

Proof of Point 4: Similarly, we see from Table 6.2 that $\mathbf{P}_+$, which we call the "top" equilibrium, is either a sink or a source. By (6.27),

$$\operatorname{tr} \mathbf{DF}_* = 1 - \rho.$$

If $\rho > 1$, then $\operatorname{tr} \mathbf{DF}_* < 0$, so $\mathbf{P}_+$ is a sink, and if $\rho < 1$, then $\mathbf{P}_+$ is a source.

[4]If you don't like the geometric argument we use in deriving Points 3 and 4, you may just calculate $\det \mathbf{DF}_*$ instead.

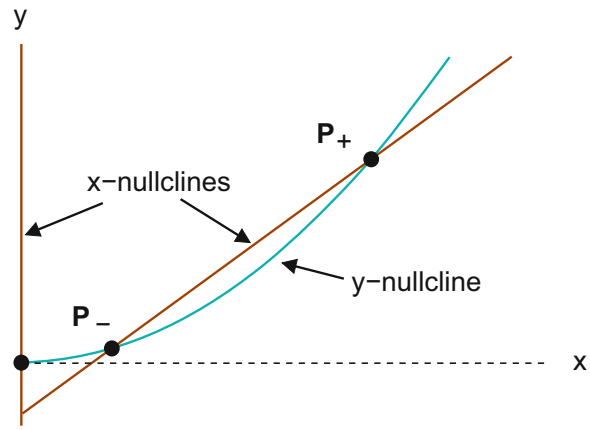

Figure 6.4: *Nullclines and equilibria of the activator–inhibitor system* (6.22) *assuming* $\sigma = 5/2$. *This system has three equilibria, provided* $\sigma > 2$. *Note that this figure differs from the nullclines in Figure 4.6, because* (6.22) *was derived from* (4.30) *by taking the limit* $\kappa \to \infty$.

6.3.2 The Turing Instability: Destabilization by Diffusion

The Turing instability may arise if an activator and an inhibitor react *in a spatially extended environment* in which the chemicals may diffuse. It is believed (see Section 2.2, Volume 2, of Murray [58]) that this mechanism may underlie the formation of periodic structures in the embryo, especially patterns on the coats of animals. The full description of the Turing instability requires both spatial and temporal variables, i.e., a PDE, which is beyond the scope of this book. However, the essential phenomenon occurs in a toy model that we study in this section.

To develop intuition about the effect of diffusion, let's consider a hypothetical scalar reaction, say modeled by $y' = 1 - y$, that takes place in two reaction vessels coupled by diffusion. This situation is described by the equations

$$\begin{aligned} y_1' &= 1 - y_1 + D(y_2 - y_1), \\ y_2' &= 1 - y_2 + D(y_1 - y_2), \end{aligned}$$

where $D > 0$ is a diffusion constant. The diffusive terms cause reactant to move from the cell with higher concentration to the cell with lower concentration, at a rate proportional to the difference in concentration. The original scalar ODE has a unique equilibrium at $y = 1$, and it is asymptotically stable. The system with diffusion has the unique equilibrium $(1, 1)$; moreover, the eigenvalues of the 2×2 coefficient matrix are -1 and $-1 - 2D$, so this equilibrium is also asymptotically stable. Indeed, the extra eigenvalue $-1 - 2D$ is even more negative than the original one. Thus, diffusion is usually regarded as a stabilizing effect.

However, in some circumstances diffusion can be destabilizing. Consider two reaction vessels, each containing an activator–inhibitor system modeled by (6.22). Suppose the inhibitor is allowed to diffuse[5] between the two cells, which leads to the four-dimensional system

$$\begin{array}{lrcl} (a) & x_1' & = & \sigma x_1^2/(1+y_1) - x_1, \\ (b) & y_1' & = & \rho\,[x_1^2 - y_1] + D(y_2 - y_1), \\ (c) & x_2' & = & \sigma x_2^2/(1+y_2) - x_2, \\ (d) & y_2' & = & \rho\,[x_2^2 - y_2] + D(y_1 - y_2). \end{array} \tag{6.28}$$

We assume that $\sigma > 2$ and $\rho > 1$, so that the top equilibrium (x_+, y_+) of (6.22) is asymptotically stable. Let's apply Theorem 6.1.1 to determine the stability of the equal-concentration equilibrium (x_+, y_+, x_+, y_+) of (6.28). The Jacobian $\mathbf{DF}_*$ of (6.28) at the equilibrium may be decomposed into block form

$$\mathbf{DF}_* = \begin{bmatrix} A - B & B \\ B & A - B \end{bmatrix}, \tag{6.29}$$

where A is the 2×2 Jacobian of (6.22) at the equilibrium, and

$$B = \begin{bmatrix} 0 & 0 \\ 0 & D \end{bmatrix}$$

covers the diffusion terms. By applying a similarity transformation with block structure

$$S = \begin{bmatrix} I & -I \\ I & I \end{bmatrix},$$

where I is the 2×2 identity matrix, we may reduce $\mathbf{DF}_*$ to the block diagonal form

$$S^{-1}\mathbf{DF}_* S = \begin{bmatrix} A & 0 \\ 0 & A - 2B \end{bmatrix}. \tag{6.30}$$

The four eigenvalues of this matrix are the two eigenvalues of A and the two eigenvalues of $A - 2B$. Since (x_+, y_+) is asymptotically stable, the eigenvalues of A have negative real parts, so stability hinges on the eigenvalues of $A - 2B$.

Taking A from (6.27), we calculate that

$$\det(A - 2B) = \det \begin{bmatrix} 1 & -1/\sigma \\ 2\rho x_+ & -\rho - 2D \end{bmatrix} = -\rho + 2\rho x_+/\sigma - 2D.$$

[5]Realistically, both chemicals should be allowed to diffuse. For the model problem (6.28), such a perturbation, if not too large, has little consequence. We ask you to verify this in Exercise 20. However, in a PDE formulation of the full problem, diffusion of the activator is an essential effect. In particular, it influences decisively the length scale of periodic structures that may arise from the Turing instability.

If D is large enough, the term $-2D$ makes the determinant of this matrix negative. For such large D, one of the eigenvalues of $A - 2B$, and hence an eigenvalue of $\mathbf{DF}_*$, must be positive, meaning that the equilibrium of the 4×4 system is unstable. *This is the Turing instability: an otherwise stable equilibrium has been destabilized by diffusion!*

What is the long-term behavior of solutions of (6.28) when the equal-concentration equilibrium is unstable? We wouldn't want to stop the determined reader from firing up his/her computer to answer this question right now, but let us mention that in Chapter 8 we will develop analytical methods to attack this question.

For reference when we return to this problem, we record the following information: One eigenvalue of the Jacobian of (6.28) at the equal-concentration equilibrium (x_+, y_+, x_+, y_+) is positive if $D > D_*$, where, after some calculation, we find that the critical value of diffusion is

$$D_* = \frac{\rho\sqrt{1 - 4/\sigma^2}}{2}.$$

If $D = D_*$, the equal-concentration equilibrium of (6.28) is nonhyperbolic, and the null eigenvector $\mathbf{v}$ of $\mathbf{DF}_*$ in this case is

$$\mathbf{v} = \begin{bmatrix} \mathbf{w} \\ -\mathbf{w} \end{bmatrix}, \tag{6.31}$$

where $\mathbf{w} \in \mathbb{R}^2$ spans the kernel of $A - 2B$.

6.4 Feedback Stabilization of an Inverted Pendulum[6]

Consider an inverted pendulum mounted on a cart, as shown in Figure 6.5. The cart may slide along a track, and let $\hat{x}$ measure its displacement along the track. The pendulum may rotate about its pivot in the vertical plane aligned with the track, and let θ be the angle with the vertical.

We don't need any theory to know that the straight-up equilibria of the system

$$\hat{x} = \text{ const}, \qquad \theta = 0, \tag{6.32}$$

are unstable. In this section we ask whether by applying force to the cart, we can reliably drive the system to one of these otherwise unstable equilibria, specifically, a force proportional to the amount the system is out of equilibrium. To put this more mathematically, consider a force on the cart given by

[6]This section is a standalone, moderately difficult application of Theorem 6.1.1. If you feel you need to press ahead, the section may be skipped or read later without loss of continuity.

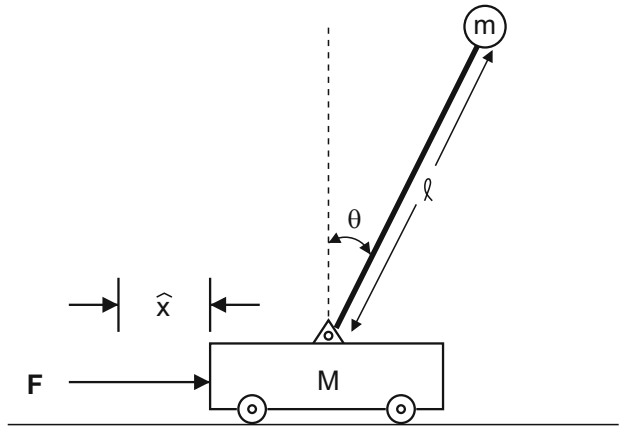

Figure 6.5: *An inverted pendulum mounted on a cart. Can it be stabilized by judicious application of a force* **F***?*

$$\hat{F} = \hat{A}\theta + \hat{B}\frac{d\theta}{d\hat{t}}, \tag{6.33}$$

where $\hat{A}, \hat{B}$ are constants, what is called *feedback control*. Note that the states (6.32) are still equilibria of this system when $\hat{F}$ is given by (6.33). We shall apply Theorem 6.1.1 to determine whether there are any values of the coefficients $\hat{A}, \hat{B}$ in (6.33) that make the equilibria (6.32) of the force–cart–pendulum system stable. Perhaps you, like us, will be surprised by the answer: *no choice of* $\hat{A}, \hat{B}$ *will make the equilibria stable!* However, the surprises don't stop here: with a different feedback control involving the speed of the cart,

$$\hat{F} = \hat{A}\theta + \hat{C}\frac{d\hat{x}}{d\hat{t}}, \tag{6.34}$$

it *is* possible to stabilize the pendulum in the inverted position!

The following ODEs describe the motion of the cart–pendulum system:

$$M\frac{d^2\hat{x}}{d\hat{t}^2} + m\left(\frac{d}{d\hat{t}}\right)^2[\hat{x} + \ell\sin\theta] = \hat{F}(\hat{t}) - c\frac{d\hat{x}}{d\hat{t}}, \tag{6.35}$$

$$m\ell^2\frac{d^2\theta}{d\hat{t}^2} = mg\ell\sin\theta - b\frac{d\theta}{d\hat{t}} - m\ell\cos\theta\frac{d^2\hat{x}}{d\hat{t}^2}. \tag{6.36}$$

As indicated in the figure, M is the mass of the cart, m is the point mass at the end of a (massless) rod of length ℓ, and g is the acceleration of gravity. The term $-c\,d\hat{x}/d\hat{t}$ models a frictional force resisting motion of the cart, and $-b\,d\theta/d\hat{t}$ models a frictional torque resisting rotation of the pendulum.

The simplest derivation of these equations is based on the Lagrangian approach to mechanics[7] (see, for example, Chapter 7 of [88]). Derivations may be found online, or alternatively it might be a nice reward for learning this formalism to be able to perform the derivation yourself. In any case, *you do not need to understand the derivation to proceed with reading this section.*

First, some preliminary processing of the equations: We combine the two terms in (6.35) involving $d^2\hat{x}/d\hat{t}^2$, and then we divide the equation by $(M+m)\ell$ to obtain

$$\frac{1}{\ell}\frac{d^2\hat{x}}{d\hat{t}^2} + \alpha\left(\frac{d}{d\hat{t}}\right)^2[\sin\theta] = \frac{\hat{F}(\hat{t})}{(M+m)\ell} - \frac{c}{(M+m)\ell}\frac{d\hat{x}}{d\hat{t}}, \tag{6.37}$$

where $\alpha = m/(M+m)$. Dividing (6.36) by $m\ell^2$ yields

$$\frac{d^2\theta}{d\hat{t}^2} = \frac{g}{\ell}\sin\theta - \frac{b}{m\ell^2}\frac{d\theta}{d\hat{t}} - \frac{\cos\theta}{\ell}\frac{d^2\hat{x}}{d\hat{t}^2}. \tag{6.38}$$

To nondimensionalize the equations, we define $t = \omega\hat{t}$, where $\omega^2 = g/\ell$ and $x = \hat{x}/\ell$. The scaled equations are

$$\begin{aligned} x'' + \alpha[\sin\theta]'' &= F(t) - \gamma x', \\ \cos\theta\, x'' + \theta'' &= \sin\theta - \beta\theta', \end{aligned}$$

where

$$F(t) = \frac{\hat{F}(\hat{t})}{(M+m)\ell\omega^2}, \qquad \gamma = \frac{c}{(M+m)\omega}, \qquad \beta = \frac{b}{m\ell^2\omega}$$

are all dimensionless, and prime indicates d/dt.

From here, the most straightforward way to process the equations would be to evaluate the derivative

$$[\sin\theta]'' = (\cos\theta)\,\theta'' - (\sin\theta)(\theta')^2,$$

invert the coefficient matrix

$$\begin{bmatrix} 1 & \alpha\cos\theta \\ \cos\theta & 1 \end{bmatrix}$$

[7] A direct derivation of (6.35), (6.36) from Newton's laws is tricky, but after the fact, let us interpret the equations in these terms. Equation (6.35) is the x-component of Newton's second law for the motion of the center of mass of the total system. Equation (6.36) comes from computing moments around the pivot, but with one nonstandard ingredient: the term $m\ell\cos\theta\, d^2\hat{x}/d\hat{t}^2$ is a fictitious torque reflecting the fact that moments are calculated about a point that may be accelerating. Note that the cart exerts an unknown force on the pendulum at the pivot. The above equations sidestep this issue; this force does not appear in (6.35), because an internal force does not contribute to the motion of the center of mass, and in (6.36) the moment of this force vanishes, because the length of the lever arm is zero.

to express (x'', θ'') in terms of lower-order derivatives, derive a first-order four-dimensional system for (x, x', θ, θ'), and invoke Theorem 6.1.1 to assess the stability of the equilibrium. However, this program involves some fairly messy calculations that we may avoid by proceeding as follows.

In the first place, undifferentiated x does not appear in the above equations, only derivatives of x. (Physically, this just expresses the fact that behavior is not affected by exactly where the cart is.) Therefore, we may reduce the order of the system by defining $v = dx/dt$ and rewriting the equations as

$$\begin{array}{rcl} v' + \alpha[\sin\theta]'' & = & F(t) - \gamma v, \\ (\cos\theta)\, v' + \theta'' & = & \sin\theta - \beta\theta'. \end{array} \tag{6.39}$$

Although we could now derive a first-order three-dimensional system for (v, θ, θ'), we continue our nonstandard approach by sticking with the mixed-order two-dimensional system (6.39). Let the applied force be determined by the feedback formula

$$F = A\theta + B\theta' + Cv.$$

Note that $v = \theta = 0$ is still an equilibrium of (6.39) with this force. To investigate whether any coefficients A, B, C make this equilibrium stable, we linearize the equations around the equilibrium:

$$\begin{array}{rcl} v' + \alpha\theta'' & = & A\theta + B\theta' + Cv - \gamma v, \\ v' + \theta'' & = & \theta - \beta\theta'. \end{array}$$

An exponential $(v(t), \theta(t)) = e^{\lambda t}(v_0, \theta_0)$ is a solution of this system iff

$$\begin{bmatrix} \lambda + \gamma - C & \alpha\lambda^2 - B\lambda - A \\ \lambda & \lambda^2 + \beta\lambda - 1 \end{bmatrix} \begin{bmatrix} v_0 \\ \theta_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

We get a nontrivial solution only if the determinant of the matrix is zero, which implies that

$$(1-\alpha)\lambda^3 + (\beta + \gamma - C + B)\lambda^2 + (-1 + \beta(\gamma - C) + A)\lambda - (\gamma - C) = 0. \tag{6.40}$$

This is exactly the same characteristic polynomial that would have resulted from the standard approach based on writing (6.39) as a first-order three-dimensional system. We leave it to the dedicated reader to verify this claim. Therefore, by Theorem 6.1.1, the equilibrium $v = \theta = 0$ of (6.39) will be asymptotically stable if the three roots of (6.40) all lie in the left half-plane.

Claim 1: *If $C = 0$, the equilibrium $v = \theta = 0$ of (6.39) is unstable.*

Proof. If $C = 0$, the product of the roots of (6.40) is $\gamma/(1-\alpha)$. Since $\alpha < 1$, this product is positive, which means that at least one root must lie in the right half-plane. □

Here is a more intuitive take on the above proof. If no force is applied, the system has two eigenvalues in the left half-plane and one in the right. (Heuristically, one eigenvalue may be associated with the speed of the cart, which is negative because of friction, and two eigenvalues with the pendulum, which have opposite signs because the decoupled θ-equation has a saddle-point equilibrium.) Thus if $F = 0$, the product of the eigenvalues is positive. The proof of the claim shows that if feedback does not involve v, it cannot change the sign of the product of the eigenvalues.

Claim 2: *If $B = 0$, the system (6.39) is asymptotically stable if*

$$\text{(a) } C = \gamma + \beta/2 \quad \textit{and} \quad \text{(b) } A > (2-\alpha) + \beta^2/2. \tag{6.41}$$

Proof. As follows from Theorem C.4.1 in Appendix C, the roots of a cubic polynomial $\gamma_0\lambda^3 + \gamma_1\lambda^2 + \gamma_2\lambda + \gamma_3$ all lie in the left half-plane iff

$$\begin{array}{lccc} (a) & \gamma_1/\gamma_0 & > & 0, \\ (b) & \gamma_1\gamma_2/\gamma_0^2 & > & \gamma_3/\gamma_0, \\ (c) & \gamma_3/\gamma_0 & > & 0. \end{array}$$

Condition (c) is satisfied if $C > \gamma$, and Condition (a) is satisfied if $C < \gamma + \beta$; the choice (6.41a) meets both these requirements. Given this value of C, Condition (b) becomes the inequality (6.41b). □

Here is an interpretation of the successful control strategy. First choose C to reverse the effective sign of friction on the block,[8] but keep it small enough that overall, the system is still dissipative. Now choose A to push the cart hard when θ is out of equilibrium, hard enough to overcome gravity and then some. This makes the equilibrium stable.

You might enjoy testing the predictions of this section numerically.[9]

[8] We find this choice surprising, because it seems to make the system *more* unstable. Indeed, if C is given by (6.41) and $A = B = 0$, then two of the roots of (6.40) have positive real parts.

[9] Are you tempted to build a physical model and use it to test the predictions? Contemplating this, you will quickly see the wisdom in the quip due to V.I. Arnol'd, a distinguished Russian mathematician: "Mathematics is the part of physics where experiments are cheap."

6.5 Lyapunov Functions

6.5.1 The Main Results

Lyapunov functions provide another approach to analyzing the stability of an equilibrium. When it can be used, this approach has two noteworthy advantages over Theorem 6.1.1: (i) It may be used to prove asymptotic stability of an equilibrium even in some cases in which one or more eigenvalues of the Jacobian have zero real part. (ii) It is more amenable to obtaining global results (cf. Exercises 12 and 13).

Suppose $\mathbf{b}_*$ is an equilibrium of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, where $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$. Let $L(\mathbf{x})$ be a real-valued function defined on an open neighborhood of $\mathbf{b}_*$, say on $\mathcal{U}_1 \subset \mathcal{U}$, that is continuous on $\mathcal{U}_1$ and $\mathcal{C}^1$ on $\mathcal{U}_1 \sim \{\mathbf{b}_*\}$. We shall call L a *Lyapunov function* for this system near $\mathbf{b}_*$ if it satisfies the following two hypotheses:

$$\begin{array}{ll} (a) & \text{For all } \mathbf{x} \in \mathcal{U}_1 \sim \{\mathbf{b}_*\}, \quad \langle \nabla L(\mathbf{x}), \mathbf{F}(\mathbf{x}) \rangle \le 0 \text{ and} \\ (b) & \text{For all } \mathbf{x} \in \mathcal{U}_1, \quad L(\mathbf{x}) \ge L(\mathbf{b}_*), \text{ with equality only if } \mathbf{x} = \mathbf{b}_*. \end{array} \tag{6.42}$$

Condition (b) requires that $\mathbf{b}_*$ be a strict minimum of L over $\mathcal{U}_1$. Regarding Condition (a): by the chain rule, the derivative of $L(\mathbf{x})$ along a trajectory $\mathbf{x}(t)$ of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ is given by

$$\frac{d}{dt} L(\mathbf{x}(t)) = \langle \nabla L(\mathbf{x}(t)), \mathbf{x}'(t) \rangle = \langle \nabla L(\mathbf{x}(t)), \mathbf{F}(\mathbf{x}(t)) \rangle;$$

thus, Condition (a) implies that $L(\mathbf{x})$ is *nonincreasing along any trajectory* of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$.

The simplest example of a Lyapunov function is provided by the energy $L(x, y) = y^2/2 + x^2/2$ of a spring–mass system, written as a (scaled) first-order system:

$$\begin{aligned} x' &= y, \\ y' &= -x - \beta y. \end{aligned}$$

(Check that (6.42) is satisfied!) In this and other examples, sublevel sets of a Lyapunov function

$$\mathcal{K}_\alpha = \{(x, y) : L(x, y) \le \alpha\} \tag{6.43}$$

provide a one-parameter family of trapping regions.

Theorem 6.5.1. *If $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ admits a Lyapunov function $L(\mathbf{x})$ near $\mathbf{b}_*$, then the equilibrium is Lyapunov stable.*

Proof. Let $\mathcal{V}$, a neighborhood of $\mathbf{b}_*$, be given. Choose a radius δ so small that $\overline{B(\mathbf{b}_*, \delta)} \subset \mathcal{V} \cap \mathcal{U}_1$. Let

$$\alpha = \min_{|\mathbf{x} - \mathbf{b}_*| = \delta} L(\mathbf{x}); \tag{6.44}$$

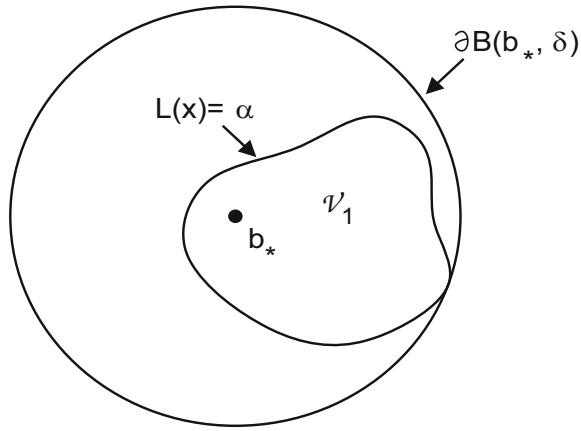

Figure 6.6: *Hypothetical neighborhoods in the proof of Theorem 6.5.1.*

by (6.42b) and compactness, $\alpha > L(\mathbf{b}_*)$. Let

$$\mathcal{V}_1 = \{\mathbf{x} \in B(\mathbf{b}_*, \delta) : L(\mathbf{x}) < \alpha\}.$$

(This data is indicated schematically in a hypothetical example in Figure 6.6.)

If $\mathbf{b} \in \mathcal{V}_1$, let $\mathbf{x}(t) = \boldsymbol{\varphi}(t, \mathbf{b})$ be the solution to the IVP. Since $L(\mathbf{x})$ decreases along trajectories, we know that for as long as the solution is in $\mathcal{U}_1$,

$$L(\mathbf{x}(t)) \le L(\mathbf{x}(0)) < \alpha.$$

In light of (6.44), $\mathbf{x}(t)$ cannot cross $\partial B(\mathbf{b}_*, \delta)$; i.e., the trajectory is confined to the compact set $\overline{B(\mathbf{b}_*, \delta)}$. By Theorem 4.1.2, the solution exists for all $t \ge 0$. Moreover, $\mathbf{x}(t) \in \overline{B(\mathbf{b}_*, \delta)} \subset \mathcal{V}$, and this shows that $\mathbf{b}_*$ is Lyapunov stable. □

Regarding Condition (b) in (6.42), the following example shows that $\mathbf{b}_*$ must be a strict minimum of $L(\mathbf{x})$ to get a useful concept: the origin is an equilibrium of

$$\begin{aligned} x' &= x, \\ y' &= -y, \end{aligned}$$

the function $L(x, y) = y^2$ satisfies Condition (a) and has a minimum at $\mathbf{0}$, but the equilibrium is unstable.

If in Condition (a) of (6.42), less-than-or-equal-to is replaced by strict inequality, $\langle \nabla L(\mathbf{x}), \mathbf{F}(\mathbf{x}) \rangle < 0$, then L is called a *strict* Lyapunov function. In this case, of course, along every trajectory $L(\mathbf{x})$ is strictly decreasing.

Theorem 6.5.2. *If the equation $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ admits a strict Lyapunov function near $\mathbf{b}_*$, then the equilibrium is asymptotically stable.*

Proof. By Theorem 6.5.1, if a trajectory $\mathbf{x}(t)$ starts near $\mathbf{b}_*$, then it exists for all positive time and stays within a compact neighborhood of $\mathbf{b}_*$. Suppose such a trajectory does not converge to $\mathbf{b}_*$. Then there exists a sequence $\{t_n\}$ tending to infinity such that $\{\mathbf{x}(t_n)\}$ is bounded away from $\mathbf{b}_*$. By invoking compactness and passing to a subsequence if necessary, we may assume without loss of generality that the sequence $\{\mathbf{x}(t_n)\}$ has a limit, say $\mathbf{x}(t_n) \to \mathbf{b}$, for some point $\mathbf{b} \neq \mathbf{b}_*$. Since L is continuous,

$$\lim_{n\to\infty} L(\mathbf{x}(t_n)) = L(\mathbf{b}) = \lim_{t\to\infty} L(\mathbf{x}(t)), \tag{6.45}$$

the latter equality because $L(\mathbf{x}(t))$ is a decreasing function.

Now consider the IVP for $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ with initial condition $\mathbf{b}$; we write this solution as $\boldsymbol{\varphi}(s, \mathbf{b})$. On the one hand, since L is a strict Lyapunov function, for any $s > 0$

$$L(\boldsymbol{\varphi}(s, \mathbf{b})) < L(\boldsymbol{\varphi}(0, \mathbf{b})) = L(\mathbf{b}). \tag{6.46}$$

On the other hand, by the continuity of $\boldsymbol{\varphi}$ (Theorem 4.5.1), we have

$$\boldsymbol{\varphi}(s, \mathbf{b}) = \lim_{n\to\infty} \boldsymbol{\varphi}(s, \mathbf{x}(t_n));$$

by the semigroup property (Proposition 4.5.3), we have $\boldsymbol{\varphi}(s, \mathbf{x}(t_n)) = \mathbf{x}(t_n + s)$, so $\boldsymbol{\varphi}(s, \mathbf{b}) = \lim_{n\to\infty} \mathbf{x}(t_n + s)$; hence,

$$L(\boldsymbol{\varphi}(s, \mathbf{b})) = \lim_{n\to\infty} L(\mathbf{x}(t_n + s)) = L(\mathbf{b}), \tag{6.47}$$

the latter equality by (6.45). But (6.47) contradicts (6.46), which proves the theorem. □

6.5.2 Lasalle's Invariance Principle

Let's attempt to apply Lyapunov functions to show that the equilibria $(\pm 1, 0)$ of Duffing's equation (6.1) are asymptotically stable. We propose the energy $E(x, y) = y^2/2 - x^2/2 + x^4/4$ as our Lyapunov function. The equilibria $(\pm 1, 0)$ are strict minima of E, and

$$\frac{dE}{dt} = -\beta y^2 \leq 0;$$

thus E is indeed a Lyapunov function. Unfortunately, it is *not* a strict Lyapunov function, because dE/dt vanishes along the x-axis. Thus we may conclude from this line of analysis only that $(\pm 1, 0)$ are Lyapunov stable, even though we know from Theorem 6.1.1 that these equilibria are in fact asymptotically stable.

Difficulties of this type, which are fairly common, can be resolved with the following result, known as *Lasalle's invariance principle*. The principle brings in information about the set on which the Lyapunov inequality fails to be strict,

$$S = \{\mathbf{x} \in \mathcal{U}_1 \sim \{\mathbf{b}_*\} : \langle \nabla L(\mathbf{x}), \mathbf{F}(\mathbf{x}) \rangle = 0\}. \tag{6.48}$$

Specifically, suppose that

no trajectory that starts in S remains in S for all positive time. (6.49)

Theorem 6.5.3. *If near an equilibrium $\mathbf{b}_*$, $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ has a Lyapunov function L that satisfies assumption (6.49), then $\mathbf{b}_*$ is asymptotically stable.*

Proof. The proof of this result closely follows that of Theorem 6.5.2. Suppose there is a trajectory $\mathbf{x}(t)$ starting close to $\mathbf{b}_*$ that does not converge to $\mathbf{b}_*$. Then proceeding as in the previous proof, we may choose a sequence $\{t_n\}$ such that $\mathbf{x}(t_n) \to \mathbf{b}$ and such that (6.45) holds. Again we consider the solution $\boldsymbol{\varphi}(s, \mathbf{b})$ of the IVP and show that it satisfies (6.47). The difference appears with (6.46): before, we could guarantee that this inequality held for any $s > 0$, but now, if $\mathbf{b} \in S$, it might happen that $L(\boldsymbol{\varphi}(s, \mathbf{b})) = L(\mathbf{b})$ for a range of s. However, the crucial point is this: by Assumption (6.49), the trajectory $\boldsymbol{\varphi}(s, \mathbf{b})$ cannot remain in S indefinitely, and hence there must be *some* value of s such that (6.46) holds. Thus, we can still obtain the contradiction needed to complete the proof. □

In Exercise 12 we ask you to use Lasalle's invariance principle to complete the proof of asymptotic stability for the equilibria $(\pm 1, 0)$ of Duffing's equation using Lyapunov functions.

6.5.3 Construction of Lyapunov Functions: An Example

The real mystery regarding Lyapunov functions is not how to use them but how to *find* them. There are a few standard tricks: In mechanical problems, such as Duffing's equation above, energy is an obvious candidate. For two classes of equations—Hamiltonian systems (cf. Exercise 10) and gradient systems (cf. Section 6.9.1)—the structure of such equations automatically provides a Lyapunov function. Another technique that sometimes works is introduced in Exercise 13(b). Failing these special cases, you are forced to rely on ingenuity and insight.

This early in your study of ODEs, you may feel that insight is in short supply, but we urge you to take the long view: you will find that in studying an equation over an extended period of time, you develop intuition about it that may astonish those less familiar with the equation, and this intuition will, among other benefits, help in constructing a Lyapunov function.

Let's illustrate how insight can be exploited to construct a Lyapunov function for the Lotka–Volterra equations with logistic limits to growth of the prey[10],

$$\begin{array}{lrcl} (a) & x' & = & x(1-x/K-y), \\ (b) & y' & = & \rho y(x-1). \end{array} \tag{6.50}$$

We assume that $K > 1$, so that the coexistence equilibrium of (6.50), located at $(x, y) = (1, 1-1/K)$, lies in the physical domain $\{x > 0, y > 0\}$. It is readily verified from Theorem 6.1.1 that this equilibrium is asymptotically stable.

Note that setting $K = \infty$ in (6.50) yields the original Lotka–Volterra equation (1.39) from Chapter 1. As we saw there, the function

$$\rho(x - \ln x) + y - \ln y, \tag{6.51}$$

which has a strict minimum at $(1, 1)$, is constant on the (periodic) orbits of (1.39). This is of course a special case of a Lyapunov function. Let's try to modify (6.51) to obtain a Lyapunov function for (6.50) when $K < \infty$, say

$$L(x, y) = Ax - B \ln x + y - D \ln y \tag{6.52}$$

for some constants A, B, D. (By scaling (6.52), we have assumed without loss of generality that the coefficient of y in (6.52) is unity.) To determine appropriate coefficients in (6.52), we first require that this function assume its minimum at the equilibrium $(x, y) = (1, 1-1/K)$ of (6.50); this yields that $A = B$ and $D = 1-1/K$. Thus we may rewrite (6.52) as

$$L(x, y) = A(x - \ln x) + y - (1 - 1/K) \ln y.$$

A calculation shows that

$$\frac{dL}{dt} = A(x-1)(1-y-x/K) + \rho(y-1+1/K)(x-1).$$

If we choose $A = \rho$, then all terms that are *not* $\mathcal{O}(1/K)$ cancel, and this equation simplifies to

$$\frac{dL}{dt} = -\frac{\rho(x-1)^2}{K}.$$

In particular, $dL/dt \leq 0$, so L is a Lyapunov function for (6.50) at the coexistence equilibrium.

It follows from Theorem 6.5.1 that the equilibrium $(1, 1-1/K)$ of (6.50) is Lyapunov stable. Although L is not a strict Lyapunov function, by invoking Lasalle's

[10] In Exercise 13(c) we relate this Lyapunov function to behavior of the augmented Lotka–Volterra equation that we discussed in Section 1.6.

invariance principle, you may prove, independently of Theorem 6.1.1, that this equilibrium is asymptotically stable. In fact, in Exercise 13(c) we outline how L may be used to prove that the equilibrium is *globally* attracting.

6.6 Stable and Unstable Manifolds

In this section we consider equilibria where the Jacobian has eigenvalues with both positive and negative real parts.

6.6.1 A Linear Example

Stable and unstable manifolds[11] can be seen already in the linear system

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{6.53}$$

This matrix has eigenvalues equal to ± 1, and the general solution of (6.53) is

$$\begin{bmatrix} x \\ y \end{bmatrix} = C_1 e^t \begin{bmatrix} 1 \\ 1 \end{bmatrix} + C_2 e^{-t} \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{6.54}$$

It is easily calculated that the trajectory (6.54) is contained in the hyperbola

$$y^2 - x^2 = C, \tag{6.55}$$

where $C = -4C_1C_2$. (Alternatively, C may be interpreted as twice the total energy, kinetic ($y^2/2$) plus potential ($-x^2/2$), which a simple calculation shows is constant along trajectories.) The case $C = 0$ is qualitatively different from C nonzero. Specifically, please verify the following points:

- If $C \neq 0$, then the set (6.55) consists of exactly two distinct orbits[12]—each branch of the hyperbola is an orbit. Both orbits are bounded away from the the equilibrium of (6.53) at the origin. (Cf. Figures 6.7a,b.)

[11] The general term "manifold" is defined in Section B.3.3, but in most examples the stable and unstable manifolds will be just curves in the plane.

[12] Recall that the *orbit* of a solution of an ODE means the curve traced out by the solution, considered merely as a subset of $\mathbb{R}^d$, independent of any parametrization. By contrast, the term *trajectory* includes a specific parametrization that yields a solution of the ODE. Each orbit corresponds to infinitely many different trajectories, time translates of one another. Thus, it is more convenient to enumerate orbits than trajectories.

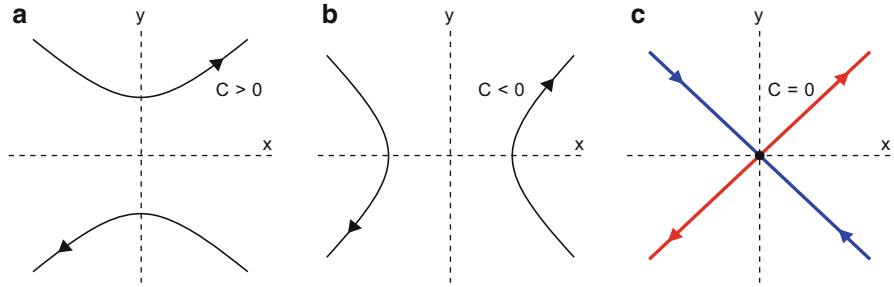

Figure 6.7: *Orbits of the model problem* (6.53). *Stable (blue) and unstable (red) manifolds are identified in Panel (c).*

- If $C = 0$ (illustrated in Figure 6.7c), the singular hyperbola consists of five orbits, the equilibrium $\{\mathbf{0}\}$ and four rays of slope $\pm 45°$ starting at the equilibrium, i.e.,

$$\{y = x > 0\}, \quad \{y = x < 0\}, \quad \{y = -x > 0\}, \quad \{y = -x < 0\}. \tag{6.56}$$

Stable and unstable manifolds relate to the asymptotic behavior of solutions as $t \to \pm\infty$. If the initial conditions $x(0), y(0)$ of a solution of (6.53) lie on one of the hyperbolas (6.55) with $C \neq 0$, then this solution tends to infinity in both limits $t \to \pm\infty$. By contrast, if $x(0), y(0)$ lies on the line $\{y = -x\}$, then this solution tends to the equilibrium as $t \to +\infty$. Conversely, for every solution that tends to the equilibrium as $t \to +\infty$, its initial conditions must lie on this line. Reflecting this behavior, the line $\{y = -x\}$ is called the *stable manifold* of the equilibrium, denoted by $\mathcal{M}_s$. Note that the stable manifold, which is colored blue in Figure 6.7c, is the union of three orbits, the equilibrium and two of the rays in (6.56).

Analogously, the line $\{y = +x\}$, which is colored red in the figure, is called the *unstable manifold* $\mathcal{M}_u$. A solution of (6.53) tends to the equilibrium as $t \to -\infty$ if and only if $(x(0), y(0)) \in \mathcal{M}_u$.

Remark: We shall always observe the above color conventions—blue for stable manifolds, red for unstable manifolds.

It is not difficult to generalize these ideas to a linear system of any dimension whose equilibrium is hyperbolic, say $\mathbf{x}' = A\mathbf{x}$, where $\Re\lambda_j(A) \neq 0$. As discussed in Exercise 6, the stable manifold is given by the linear span of all generalized eigenvectors[13] of A associated with eigenvalues such that $\Re\lambda < 0$; for the unstable manifold, it is the eigenvalues such that $\Re\lambda > 0$.

[13]Recall from the discussion of Jordan normal forms in Section C.1 that $\mathbf{v}$ is a generalized eigenvector of a matrix A with eigenvalue λ if $(A - \lambda I)^p \mathbf{v} = \mathbf{0}$ for some power p.

6.6.2 Statement of the Local Theorem

The following result generalizes the stable/unstable-manifold concept to the IVP

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0) = \mathbf{b} \tag{6.57}$$

for a nonlinear system that has a hyperbolic equilibrium at $\mathbf{x} = \mathbf{b}_*$. For brevity we focus only on the stable manifold. The first change from the linear case is that we restrict attention to initial conditions $\mathbf{b}$ that are close to $\mathbf{b}_*$. Such a restriction is hard to avoid with a nonlinear equation. Indeed, even with it, in a neighborhood of $\mathbf{b}_*$ there may be initial conditions such that the IVP (6.57) does not have a solution for all time $t > 0$ (cf. Exercise 9). The second change is that $\mathcal{M}_s$ is a curved manifold, not a linear subspace.

To set the notation, suppose that d_s eigenvalues of the Jacobian $\mathbf{DF}(\mathbf{b}_*) = \mathbf{DF}_*$ have negative real parts and that the remaining $d - d_s$ eigenvalues of $\mathbf{DF}_*$ have positive real parts. Let $E_s \subset \mathbb{R}^d$ denote the span of all generalized eigenvectors of $\mathbf{DF}_*$ associated with eigenvalues such that $\Re\lambda < 0$, a subspace of dimension d_s.

Theorem 6.6.1. *Given a hyperbolic equilibrium $\mathbf{b}_*$ of an ODE as described above, there exist a (bounded) neighborhood $\mathcal{V}$ of $\mathbf{b}_*$ in $\mathbb{R}^d$ and a differentiable manifold $\mathcal{M}_s \subset \mathcal{V}$ of dimension d_s tangent to E_s at $\mathbf{b}_*$ such that: (i) If $\mathbf{b} \in \mathcal{M}_s$, then the IVP (6.57) has a solution $\boldsymbol{\varphi}(t, \mathbf{b})$ for all positive time and*

$$\lim_{t\to\infty} \boldsymbol{\varphi}(t, \mathbf{b}) = \mathbf{b}_*. \tag{6.58}$$

(ii) If $\mathbf{b} \in \mathcal{V} \sim \mathcal{M}_s$, then $\boldsymbol{\varphi}(t, \mathbf{b})$ leaves $\mathcal{V}$ at some positive time.

Although it is mathematically redundant to do so, it may be pedagogically useful to describe this result in nontechnical language: $\mathcal{M}_s$ characterizes certain initial conditions for (6.57) in terms of the asymptotic behavior of the corresponding solution of the IVP. To develop this thought, consider the IVP

$$\begin{aligned} x' &= -x & x(0) &= b_1, \\ y' &= y + x^2 & y(0) &= b_2. \end{aligned} \tag{6.59}$$

By finding an explicit solution you can show (cf. Exercise 8) that the solution of this IVP converges to the equilibrium $(0, 0)$ iff

$$b_2 = -b_1^2/3; \tag{6.60}$$

i.e., (6.60) defines the stable manifold of the equilibrium (and $\mathcal{V}$ can be chosen arbitrarily).

Remarks: (i) If in the theorem $d_s = d$, then the stable manifold is simply an entire neighborhood of $\mathbf{b}_*$. In this case, Theorem 6.6.1 is effectively just a restatement of

Theorem 6.1.1. *(ii)* Apart from being defined in a bigger or smaller neighborhood of $\mathbf{b}_*$, the manifold $\mathcal{M}_s$ is unique. *(iii)* Under the hypotheses of Theorem 6.6.1, there is also an *unstable* manifold $\mathcal{M}_u$ through $\mathbf{b}_*$. Its dimension equals the number of eigenvalues of $\mathbf{DF}_*$ with positive real parts. It is most easily described as the stable manifold of the time-reversed system $\mathbf{x}' = -\mathbf{F}(\mathbf{x})$. In particular, if $\mathbf{b} \in \mathcal{M}_u$, then $\boldsymbol{\varphi}(t, \mathbf{b})$ is defined for all $t \leq 0$ and tends to $\mathbf{b}_*$ as $t \to -\infty$.

The proof of this result, which is long and rather technical, is outlined in an appendix. This material may be omitted without loss of continuity. However, to continue profitably reading this book, you need to develop intuition about stable and unstable manifolds, *especially in the most common case of a saddle point in a two-dimensional system*, where $\mathcal{M}_s$ and $\mathcal{M}_u$ reduce to curves. This may be achieved through interpreting the conclusions of Theorem 6.6.1 in various examples, including those in the rest of Section 6.6, in Section 6.7, and in the exercises of Section 6.8.3.

6.6.3 A Nonlinear Example

In general it is not possible to find a formula for the stable or unstable manifolds at an equilibrium. In this section we consider a nonlinear equation with a hyperbolic equilibrium for which such formulas can be found, the ODE

$$\begin{aligned} x' &= y, \\ y' &= x + x^3. \end{aligned} \tag{6.61}$$

This Duffing-like equation may be viewed as a nonlinear perturbation of (6.53). It describes a particle under a force $F(x) = x + x^3$ in which both linear and cubic terms are repulsive. Its potential energy, $V(x) = -x^2/2 - x^4/4$, is shown in Figure 6.8. We discuss solutions of this equation in terms of the loose analogy of a marble rolling in the x, z-plane over the hill given by $z = V(x)$. Because there is no friction, each orbit of (6.61) is contained in (but not necessarily equal to) a level set of the energy function,[14]

$$\{(x, y) : H(x, y) = C\}, \tag{6.62}$$

where $H(x, y) = y^2/2 + V(x)$.

For the case $C > 0$, representative level sets of (6.62) are shown in Figure 6.9a. The orbit $y = +\sqrt{2C + x^2 + x^4/2}$ in the upper half-plane derives from a particle that at large negative times is far to the left of the origin and is moving to the right toward the top of the hill at $x = 0$ (the situation shown in Figure 6.8); it slows down as it approaches the top of the hill, but it has enough energy to clear it; after it passes $x = 0$, it sails off to the right, at ever increasing speeds. In focusing on

[14]Equation (6.61) is an example of what's called a *Hamiltonian* system, a class of ODEs that is defined in Exercise 10. From now on, we will use the letter H for the energy of Hamiltonian systems.

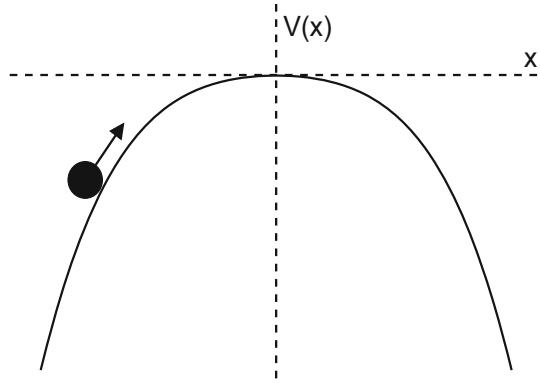

Figure 6.8: *Illustration of the rolling-marble analogy for* (6.61) *described in the text.*

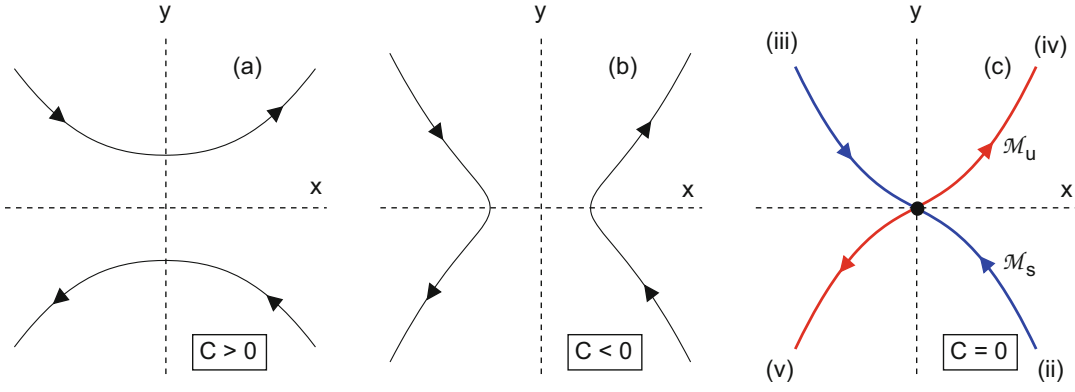

Figure 6.9: *Level sets of* (6.62)*. In Panels (a) and (b), trajectories tend to infinity as* $t \to \pm\infty$*. In Panel (c), trajectories on the stable manifold (colored blue) tend to* $(0,0)$ *as* $t \to +\infty$*; on the unstable manifold (red), as* $t \to -\infty$*.*

the *orbit*, we suppress information regarding exactly when the particle passes over the hill. Similarly, the orbit $y = -\sqrt{2C + x^2 + x^4/2}$ in the lower half-plane may be interpreted in terms of a particle moving from right to left that clears the hill. These orbits are bounded away from the origin $(x, y) = (0, 0)$ in the phase plane, and they tend to infinity as $t \to \pm\infty$.

If $C < 0$, (6.62) again consists of exactly two orbits (see Figure 6.9b), one in the right half-plane $\{x > 0\}$ and one in $\{x < 0\}$. The orbit in the right half-plane derives from a particle that at large negative times is far to the right of the origin and is moving toward the hill but does not have enough energy to clear it; thus the particle is turned around and sails back to the right as time increases. The orbit in the left half-plane is similarly described.

When $C = 0$, the level set (6.62) consists of two crossed curves, $y = \pm x\sqrt{1 + x^2/2}$ (see Figure 6.9c). This level set decomposes into exactly five orbits, as follows:

$$\begin{array}{lll} \text{(i)} & x = y = 0, & \\ \text{(ii)} & y = -x\sqrt{1 + x^2/2}, & 0 < x < \infty, \\ \text{(iii)} & y = -x\sqrt{1 + x^2/2}, & -\infty < x < 0, \\ \text{(iv)} & y = x\sqrt{1 + x^2/2}, & 0 < x < \infty, \\ \text{(v)} & y = x\sqrt{1 + x^2/2}, & -\infty < x < 0. \end{array} \tag{6.63}$$

Orbit (ii) derives from a particle that at large negative times was far to the right of the origin and is moving to the left with just enough energy to converge to the top of the hill as $t \to \infty$; this is a single orbit. Similarly, orbit (iii) derives from a particle moving to the right that converges to the top of the hill. The stable manifold is the union of orbits (i), (ii), and (iii), the curve $y = -x\sqrt{1 + x^2/2}$, where $-\infty < x < \infty$.

Mathematically, orbits (iv) and (v) are quite similar to orbits (ii) and (iii), but the description in words is harder to swallow: the particle "falls off" the hilltop at time minus infinity and is moving away from the equilibrium for all time, starting at infinitesimal speeds but continuously accelerating; it takes an infinite amount of time to fall off an equilibrium at time minus infinity, just as it takes an infinite amount of time to converge to equilibrium as t tends to plus infinity. The motion is to the right or left for orbits (iv) or (v), respectively. These two orbits, together with the origin, constitute the unstable manifold.

As regards Theorem 6.6.1, we ask you to verify that for every $\delta > 0$, the definitions

$$\mathcal{V} = \{(x, y) \in \mathbb{R}^2 : \sqrt{x^2 + y^2} < \delta\}, \quad \mathcal{M}_s = \{(x, y) \in \mathcal{V} : y = -x\sqrt{1 + x^2/2}\}, \tag{6.64}$$

satisfy all the claims of the theorem. The manifold $\mathcal{M}_u$ admits a similar parametrization.

For (6.61), conservation of energy allows us to derive explicit formulas for $\mathcal{M}_s$ and $\mathcal{M}_u$. If we modified (6.61) slightly, for example by including a friction term, we

could no longer find explicit parametrizations, but we could still invoke the theorem to guarantee that stable and unstable manifolds exist.

Incidentally, this example illustrates another difference between orbits and trajectories. Most solutions of (6.61) blow up in finite time, both forward and backward. In such cases, the trajectory is defined only for a finite interval of time, and the parametrization is singular at the endpoints of this interval. By contrast, although the orbits (6.63) extend to infinity, their behavior is nonsingular.

6.6.4 Global Behavior of Stable/Unstable Manifolds

Theorem 6.6.1 concerns what properly should be called a *local* stable submanifold. When it makes for greater clarity, we use the notation $\mathcal{M}_s^{(loc)}$ to indicate this idea explicitly. It is useful to extend local stable and unstable manifolds to global objects. We want a global stable manifold $\mathcal{M}_s^{(glob)}$ to satisfy the following two properties:

(i) $\mathcal{M}_s^{(glob)}$ is invariant under the flow.

(ii) $\mathcal{M}_s^{(glob)}$ contains all initial conditions $\mathbf{b}$ such that $\boldsymbol{\varphi}(t, \mathbf{b})$ converges to the equilibrium as $t \to \infty$.

Regarding Property (i), let us make the intuitive concept "invariant" precise: a set S is called *invariant* if for every initial condition $\mathbf{b} \in S$, we have that $\boldsymbol{\varphi}(t, \mathbf{b}) \in S$ for all t, positive or negative, such that the IVP has a solution.

In the example (6.61), the curve $y = -x\sqrt{1 + x^2/2}$ where $-\infty < x < \infty$ is such a global stable manifold. Although the extension is trivial in this example, let us consider Duffing's equation (without friction)

$$\begin{aligned} x' &= y, \\ y' &= x - x^3, \end{aligned} \tag{6.65}$$

which differs from (6.61) only in the sign of the cubic term in the force. The origin is a hyperbolic equilibrium of (6.65), and the Jacobian there is given by

$$\mathbf{DF}(\mathbf{0}) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

which has eigenvalues ± 1. By Theorem 6.6.1, $\mathcal{M}_s^{(loc)}$ is a curve through the origin tangent to $(1, -1)$, the eigenvector of $\mathbf{DF}(\mathbf{0})$ with eigenvalue -1, and similarly $\mathcal{M}_u^{(loc)}$ is tangent to $(1, 1)$.

We may use energy to identify stable and unstable manifolds for (6.65). Indeed, both local and global versions of stable and unstable manifolds are contained in the zero-energy level set,

$$S = \{(x, y) : y^2/2 - x^2/2 + x^4/4 = 0\}, \tag{6.66}$$

a "figure eight" that is sketched in Figure 6.10(a). Although the zero-energy set for (6.61) decomposed into the five orbits (6.63), the set (6.66) decomposes into only *three* orbits,

$$\begin{array}{lll} \text{(i)} & x = y = 0, & \\ \text{(ii)} & y^2/2 - x^2/2 + x^4/4 = 0, & x > 0, \\ \text{(iii)} & y^2/2 - x^2/2 + x^4/4 = 0, & x < 0. \end{array} \tag{6.67}$$

We may describe these orbits in terms of the "rolling marble" analogy. (You should sketch the double-well potential $-x^2/2 + x^4/4$, to which the following descriptions refer.) In orbit (ii) at $t = -\infty$ the particle falls off the equilibrium $x = 0$, it moves to the right and gets turned around by the hill, and it converges back to the origin as $t \to \infty$. The set (ii) is a called a *homoclinic* orbit: this term refers to an orbit that converges to the same equilibrium as $t \to \infty$ and $t \to -\infty$. Similarly for orbit (iii), but to the left.

Regarding Theorem 6.6.1, let us take $\mathcal{V} = \{(x, y) : |x| < 1, |y| < 1\}$. Solving the equation in (6.66) for y, we obtain the parametrization

$$\mathcal{M}_s^{(loc)} = \{(x, y) : y = -x\sqrt{1 - x^2/2}, \quad -1 < x < 1\} \tag{6.68}$$

and a similar parametrization for $\mathcal{M}_u^{(loc)}$ using the function $y = +x\sqrt{1 - x^2/2}$.

Although the local manifolds $\mathcal{M}_s^{(loc)}$ and $\mathcal{M}_u^{(loc)}$ are distinct, at the global level they merge—both $\mathcal{M}_s^{(glob)}$ and $\mathcal{M}_u^{(glob)}$ equal the whole level set (6.66). This follows from the fact that to be invariant, these sets must contain the entire orbit through each of their points. Reflecting this behavior, we extend our color conventions: *purple* when the stable and unstable manifolds coincide. (In Figure 6.10(a) we show the portions of $\mathcal{M}_s$ and $\mathcal{M}_u$ inside the neighborhood $\mathcal{V}$ as blue and red, respectively, according to our previous convention; in the future we will use purple for the entire manifold.)

Parenthetically, let us use this example to correct a possible misunderstanding about what Conclusion (ii) in Theorem 6.6.1 asserts: although solutions with initial conditions in $\mathcal{V} \sim \mathcal{M}_s^{(loc)}$ eventually leave $\mathcal{V}$, it is possible for them to return to $\mathcal{V}$ at some later time.

Nontrivial intersections[15] of $\mathcal{M}_s^{(glob)}$ and $\mathcal{M}_u^{(glob)}$, as in (6.65), are not robust. The slightest perturbation of the equation is likely to remove them. For example, if we modify (6.65) by including a small amount of friction, then $\mathcal{M}_s^{(glob)}$ and $\mathcal{M}_u^{(glob)}$ intersect only at the origin, as sketched in Figure 6.10(b). (See Exercise 12.)

[15] $\mathcal{M}_s$ and $\mathcal{M}_u$ always intersect at the equilibrium; we use "nontrivial" to refer to the existence of other intersection points. For (6.65), $\mathcal{M}_s^{(glob)}$ and $\mathcal{M}_u^{(glob)}$ coincide, but in higher-dimensional equations $\mathcal{M}_s^{(glob)} \cap \mathcal{M}_u^{(glob)}$ may be a proper, but nontrivial, subset of both manifolds. Of course the intersection consists of one or more complete orbits.

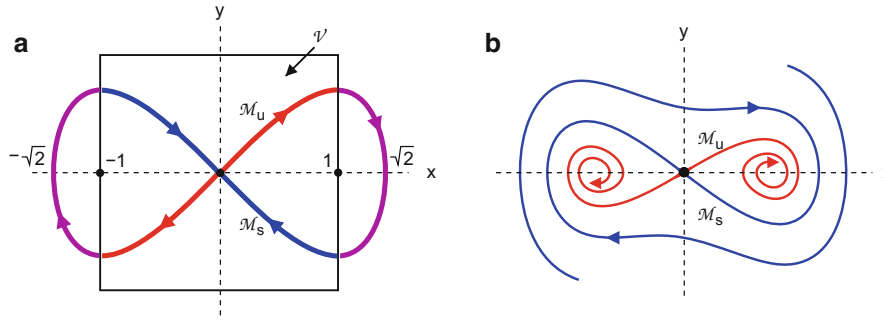

Figure 6.10: *(a) The zero-energy level set (6.66) for Duffing's equation without friction (6.65). The local manifolds $\mathcal{M}_s^{(loc)}$ and $\mathcal{M}_u^{(loc)}$, i.e., the portions inside the neighborhood $\mathcal{V}$, intersect only at the origin, but the global manifolds actually coincide and cannot be distinguished. (b) With friction the global unstable manifold spirals into the equilibria at $(\pm 1, 0)$, while the stable manifold comes in from infinity and converges to the origin. (In the figure, $\beta = 1/4$.)*

Two other examples for which $\mathcal{M}_s^{(glob)}$ and $\mathcal{M}_u^{(glob)}$ intersect one another nontrivially to form a homoclinic orbit are given in Exercises 10(f) and 17.

At a general hyperbolic equilibrium point $\mathbf{b}_*$ of an equation $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, a global stable manifold may be defined as a union of orbits through $\mathcal{M}_s^{(loc)}$; in symbols,[16]

$$\mathcal{M}_s^{(glob)} = \bigcup \{\boldsymbol{\varphi}(t, \mathbf{b}) : t \in \mathbb{R},\ \mathbf{b} \in \mathcal{M}_s^{(loc)}\}, \tag{6.69}$$

with a similar representation for $\mathcal{M}_u^{(glob)}$. Of course only negative times contribute anything new to the union (6.69). In simple examples, $\mathcal{M}_s^{(glob)}$ is a "nice" submanifold of $\mathbb{R}^d$, and mostly we restrict our attention to such cases. More generally, however, especially in higher-dimensional cases where the ODE exhibits chaos, a global stable or unstable manifold may become entangled with itself as $t \to \pm\infty$. We refer you to Section 5.5 of Meiss [54] for a careful discussion of these issues.

6.7 Drawing Phase Portraits

Recall that a *phase portrait* is a sketch of the trajectories of a few key solutions of an ODE that suggests the behavior of an arbitrary solution. Nullclines and local knowledge near equilibria from this chapter greatly help in making such portraits. Indeed, sometimes with these tools no computer solutions are required; and even

[16] Strictly speaking, in this union we may take only those times t for which the initial value problem has a solution. This limitation may be expressed quite precisely in the flow notation of Section 4.5.2, but in our opinion such precision obscures more than it clarifies.

when numerics are required, these tools are invaluable in interpreting the computations. Let us illustrate these techniques with several examples.

6.7.1 Example 1: The Chemostat

We recall the (scaled) chemostat equations

$$\begin{aligned} x' &= \frac{y}{y+1}x - \rho x, \\ y' &= -\frac{y}{y+1}x - \rho(y-\sigma), \end{aligned} \tag{6.70}$$

where the variables belong to the quarter-plane $\{x \geq 0, y \geq 0\}$. There are two equilibria:

$$x = 0,\ y = \sigma \quad \text{and} \quad x = \sigma - \rho/(1-\rho),\ y = \rho/(1-\rho). \tag{6.71}$$

At the first equilibrium, which we call "trivial," the bacteria population x vanishes. Let's assume

$$\sigma > \rho/(1-\rho) > 0, \tag{6.72}$$

so that at the second equilibrium both variables are positive; i.e., the nontrivial equilibrium is in the physical domain.

We ask you to verify the following information:

- At the nontrivial equilibrium, the entries of the Jacobian $\mathbf{DF}_*$ satisfy

  $$a_{11} = 0, \quad a_{12} > 0, \quad a_{21} < 0, \quad a_{22} < 0.$$

  Thus $\det \mathbf{DF}_* = -a_{12}a_{21} > 0$ and $\operatorname{tr} \mathbf{DF}_* = a_{22} < 0$, so this equilibrium is a sink.

- The trivial equilibrium, at which the Jacobian is lower triangular, is a saddle point. The stable eigenvector is $(0, 1)$, and in fact the stable manifold of this equilibrium is the y-axis. The unstable eigenvector is $(1, -1)$.

Thus, the local unstable manifold emerges from the saddle point $(0, \sigma)$ with slope $-45°$. In Exercise 4.9 you showed that every trajectory in the open first quadrant converges to the sink.[17] In particular, the global unstable manifold converges to the sink,[18] as sketched in the phase portrait of Figure 6.11.

Surprisingly, $\mathcal{M}_u$ can be located exactly: by adding the equations, we conclude that the line $\{x+y = \sigma\}$ is invariant under (6.70) and therefore contains the unstable manifold through the saddle point.

[17] In fact, using Theorem 6.1.1, you could now derive this behavior with less effort.

[18] *"What else could it do?"* we naively ask. Although phase portraits are often made using this kind of "logic," be careful. Not seeing other alternatives might just reflect limitations of our imagination.

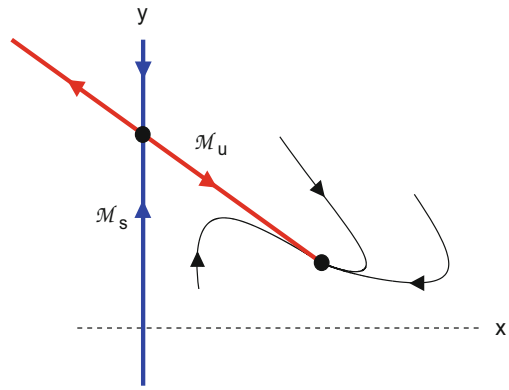

Figure 6.11: *Phase portrait for the chemostat* (6.70) *in the first quadrant, assuming* $\sigma > \rho/(1-\rho) > 0$. *In this figure,* $\sigma = 1$ *and* $\rho = 1/4$.

Incidentally, the eigenvalues of $\mathbf{DF}_*$ at the sink are real, so convergence to it will be monotone, as shown in the figure. Most trajectories will converge to the sink tangent to the less rapidly decaying eigenvector, which, for the parameters in the figure, is $(1, -2/3)$.

6.7.2 Example 2: The Activator–Inhibitor

In Section 6.3.1 we saw that if $\sigma > 2$, which we assume here, the activator–inhibitor system (6.22) has three equilibria: a sink at the origin, a saddle $\mathbf{P}_-$ at (x_-, x_-^2), where x_- is the smaller root of (6.23), and a third equilibrium $\mathbf{P}_+$ at (x_+, x_+^2), where x_+ is the larger root of (6.23). The equilibrium $\mathbf{P}_+$ is a sink or a source according as $\rho > 1$ or $\rho < 1$, respectively. The eigenvalues of the Jacobian at $\mathbf{P}_+$ are complex if

$$1 + \sqrt{1 - 4/\sigma^2} \;>\; \frac{(1+\rho)^2}{4\rho}. \tag{6.73}$$

For $\sigma = 2.1$ as in Figure 6.12, inequality (6.73) means $0.348 < \rho < 2.871$.

Two computed phase portraits are sketched in the figure. If $\rho > 1$, then one half of the unstable manifold $\mathcal{M}_u$ from the saddle point $\mathbf{P}_-$ connects to the origin, while the other half connects to $\mathbf{P}_+$; if $\rho < 1$, both halves of $\mathcal{M}_u$ connect to the origin. The phase portrait for the borderline case $\rho = 1$ is described in Exercise 10(f).

In contrast to the previous example, here the *stable* manifold $\mathcal{M}_s$ of $\mathbf{P}_-$ is also interesting: if $\rho > 1$, this curve separates the basins of attractions for the two sinks.[19] Specifically, for all initial conditions (in the physical domain $\{x \geq 0,\ y \geq 0\}$) lying "outside" $\mathcal{M}_s$, the solution of (6.22) converges to the origin as $t \to \infty$, and for all

[19]Reflecting this behavior, in the case of a saddle point in the plane ($d = 2$, $d_s = 1$), the stable and unstable manifolds (i.e., curves) were traditionally called *separatrices*.

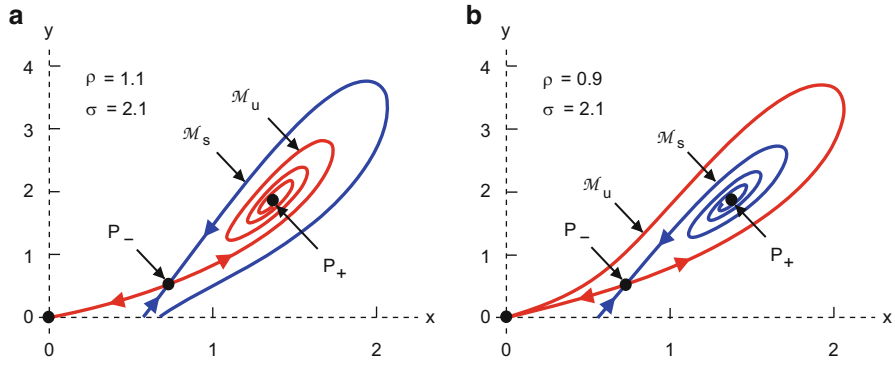

Figure 6.12: *Stable and unstable manifolds of* $\mathbf{P}_-$ *for the activator–inhibitor equations* (6.22). *Note that* $\rho > 1$ *in the left panel and* $\rho < 1$ *in the right panel, and in both cases inequality* (6.73) *is satisfied. (The phase portrait for* $\rho = 1$ *is plotted in Figure 7.2 below.)*

initial conditions lying "inside" $\mathcal{M}_s$, the solution converges to $\mathbf{P}_+$. This is an example of what is commonly referred to as *bistability*; i.e., there are two asymptotically stable equilibria, and the choice of initial conditions determines which of the two equilibria is approached as $t \to \infty$.

In Exercise 14 you will see how we used the theory to get a more faithful representation of the stable and unstable manifolds near the equilibria in Figure 6.12.

6.7.3 Example 3: Section 1.6 Revisited, Part II

The augmented Lotka–Volterra equations (6.17) have three or four equilibria in the physical domain $\{x \geq 0,\ y \geq 0\}$. The number and stabilities of these equilibria depend on ε, K as itemized in Table 6.1; the inequalities in the table are also displayed graphically in Figure 6.13(d). In Panels (a)–(c) of the figure we show one possible phase plane portrait for each of the three regions in Panel (d). (This figure augments Figure 1.9 by including stable and unstable manifolds.) As it happens, all phase portraits for parameters in Regions I and III are qualitatively equivalent, but behavior in Region II is more complicated. We shall piece together the full story of phase portraits in Region II in Chapters 8 and 9.

The flows shown for Regions I and II are bistable. The boundary between the basins of attraction of the two sinks is the stable manifold through the threshold equilibrium $(\varepsilon, 0)$.

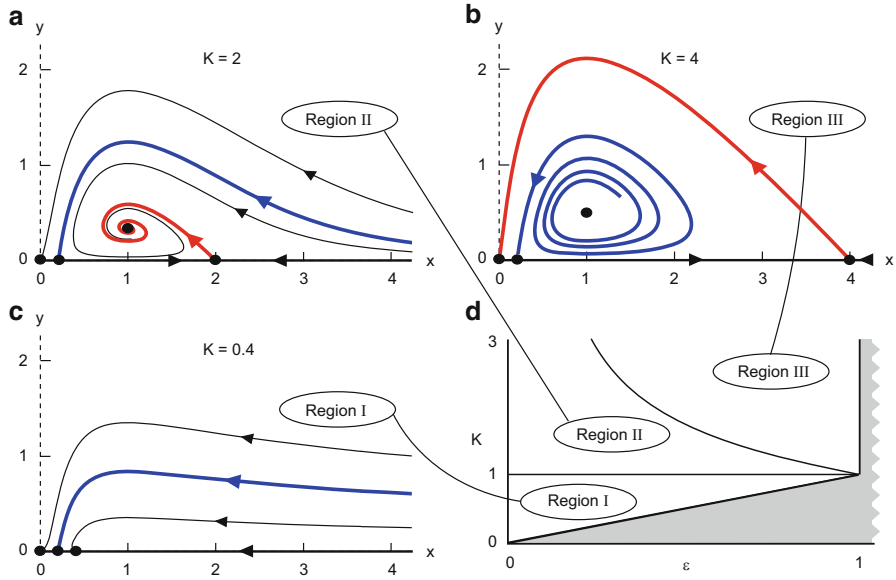

Figure 6.13: *Some representative phase portraits (Panels a, b, c) for the augmented Lotka–Volterra equations, (6.17) with $\varepsilon = 0.2$. In each case, the stable manifold of $(\varepsilon, 0)$ is shown; the unstable manifold of $(K, 0)$ is shown when this equilibrium is a saddle point (Panels a and b). For stable/unstable manifolds that are contained in the x-axis we have suppressed our usual color conventions. Panel (d) identifies regions in parameter space where the different phase portraits occur.*

6.8 Exercises

After the core exercises, there are subsections of exercises on Lyapunov functions and on sketching phase portraits.

6.8.1 Core Exercises

The primary purposes of the core exercises are:

To complete a proof	1
To apply Theorem 6.1.1 in various examples	2
To reinforce the proof techniques of Theorem 6.1.1	3
To explore cases of borderline stability	4
To relate an ecological principle to this chapter	5
To work with stable manifolds	6–8
To practice related skills	9
To introduce Hamiltonian systems	10

Don't skip Exercise 10. The Hamiltonian framework helps elucidate numerous ODEs below. In particular, in Part (f) of this exercise you find another example of a homoclinic orbit.

1. Prove Corollary 6.1.2.

2. (a) Show that for all $\mu > 0$, the equations

$$x' = \frac{\mu}{1+y^4} - x, \qquad y' = \frac{\mu}{1+x^4} - y,$$

have an equilibrium with $x = y$. Determine for what values of μ the equilibrium is asymptotically stable. More generally, for all $\mu > 0$, specify which terms in Table 2.1 apply to the equilibrium.

Food for thought: Are there equilibria with $x \neq y$? Chapter 8 will provide new tools to address such questions (cf. Exercise 8.15).

(b) For Sel'kov's model for glycolysis (from Chapter 4)

$$\begin{aligned} x' &= \rho - \sigma x - xy^2, \\ y' &= -y + \sigma x + xy^2, \end{aligned}$$

show that there is a unique equilibrium in the first quadrant. Assuming $\sigma = 0.1$, determine the ranges of ρ for which this equilibrium is asymptotically stable and for which it is unstable. For the range of ρ where it is unstable, specify which terms in Table 2.1 apply to this equilibrium.

(c) *Introduction:* Recall the (scaled) equations for a bead sliding on a rotating loop (Exercise 5.5),

$$\begin{aligned} x' &= y, \\ y' &= -\beta y - \sin x + \mu \sin x \cos x. \end{aligned} \tag{6.74}$$

For $\mu < 1$, this system (considered as an ODE on $S^1 \times \mathbb{R}$) has two equilibria, while for $\mu > 1$ it has four.

Find all these equilibria, determine their stabilities, and classify them according to Table 2.1.

Puzzle: As you can easily show, the function

$$L(x, y) = y^2/2 - \mu(\sin^2 x)/2 - \cos x \tag{6.75}$$

decreases along orbits. On the other hand, $L(x, y)$ cannot always be a Lyapunov function for the equilibrium $(0, 0)$ of (6.74)—there are values of μ such that this equilibrium is unstable. How is this possible?

Incidentally, $L(x, y)$ resembles energy but differs subtly from it: the term $\mu(\sin^2 x)/2$ represents the contribution to kinetic energy from the mass rotating about the axis of the hoop, but it is *subtracted* in (6.75), not added.

(d) *Introduction:* The Lorenz equations

$$\begin{aligned} x' &= \sigma(y - x), \\ y' &= \rho x - y - xz, \\ z' &= -\beta z + xy, \end{aligned} \tag{6.76}$$

where σ, ρ, β are positive parameters, were introduced in Exercise 3(a) in Chapter 4. For all values of the parameters, the origin is an equilibrium of this system.

Determine the parameter values for which $(0, 0, 0)$ is asymptotically stable and for which it is unstable.

3. Consider an ODE

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f}(\mathbf{x}, t), \tag{6.77}$$

where the eigenvalues of A satisfy (6.2) and $\mathbf{f} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ is Lipschitz continuous with respect to both variables. Suppose $\mathbf{f}(\mathbf{x}, t)$ is estimated by

$$|\mathbf{f}(\mathbf{x}, t)| \le \phi(t)|\mathbf{x}|,$$

where ϕ is continuous and $\lim_{t\to\infty} \phi(t) = 0$. Show that every solution of (6.77) tends to zero exponentially fast as $t \to \infty$.

Hint: Theorem 4.7.1 guarantees that solutions exist for all time. Prove the decay result by imitating the proof of Theorem 6.1.1. Specifically, the ODE in the theorem may be written

$$\mathbf{x}' = A\mathbf{x} + \mathbf{r}(\mathbf{x}),$$

where $\mathbf{r}(\mathbf{x}) = o(|\mathbf{x}|)$; i.e., $\mathbf{r}$ is small when $\mathbf{x}$ is close to the origin. Similarly, in (6.77), the remainder is small when t is large. Use this information to estimate the solution on $[t_0, \infty)$ for appropriately large t_0.

Remark: Under the slightly weaker hypothesis that

$$|\mathbf{f}(\mathbf{x},t)| \le \phi(t)(|\mathbf{x}|+1),$$

where ϕ tends to zero, solutions of (6.77) still converge to zero, but not necessarily exponentially fast. Moreover, the proof requires different ideas; a fixed-point argument (as in the proof of Theorem 6.6.1) is one possibility.

4. (a) Is the equilibrium of $x' = \sin^2 x$ at $x = 0$ Lyapunov stable, possibly even asymptotically stable, or unstable?

   (b) Is the equilibrium of $x' = -x^3 + e^x \sin^2(x^2)$ at $x = 0$ Lyapunov stable, possibly even asymptotically stable, or unstable?

   (c) When $\sigma = 2$, the activator–inhibitor system (6.22) has an equilibrium $(x, y) = (1, 1)$. Use nullclines to predict whether this equilibrium is Lyapunov stable, possibly even asymptotically stable, or unstable.

   Discussion: Recall that for $\sigma > 2$, this system has three equilibria, but for $\sigma < 2$ only one. The equilibrium at $(1, 1)$ in this borderline case is the limit of two different equilibria that are separate for $\sigma > 2$, merge as $\sigma \to 2$, and become complex for $\sigma < 2$.

   (d) For equation (4.83),

$$\begin{aligned} x' &= x^2 - y^2, \\ y' &= 2xy, \end{aligned}$$

   is the equilibrium at the origin Lyapunov stable, possibly even asymptotically stable, or unstable?

   Food for thought: After reading Section 6.11, ask yourself, what is the center manifold for this nonhyperbolic equilibrium?

5. *Introduction:* The three-dimensional system

$$\begin{aligned} x' &= rx(K - x) - p_1 xy - p_2 xz, \\ y' &= \rho_1 xy - d_1 y, \\ z' &= \rho_2 xz - d_2 z, \end{aligned} \tag{6.78}$$

   where r, K, p_k, ρ_k, d_k are positive constants, is analogous to the Lotka–Volterra system with logistic growth of the prey, except in (6.78) there are two species (y and z) that attack the third.

   (a) Show that unless $d_1/\rho_1 = d_2/\rho_2$, (6.78) has no equilibria at which all three populations are nonzero.

   Discussion: In the language of Section 2.8.3, the relation $d_1/\rho_1 = d_2/\rho_2$ is "nongeneric." In nontechnical language, such an equality could be satisfied only by accident, and even if it were satisfied, the slightest perturbation of the system would undo it. In ecology, this behavior is known as the *law*

of competitive exclusion; i.e., it is ecologically unstable for two species to compete for exactly the same resources. (Cf. Section 3.5 of [57].)

(b) Restricting attention to generic cases only, determine all equilibria in the (closed) first quadrant and their stabilities as functions of the parameters.

Advice: There are annoyingly many cases to consider, and it's far from clear how to organize them systematically. You might want to skip Part (b) for now; we will return to this question in Exercise 8.16, where bifurcation theory will guide us in organizing the calculations.

6. If $\mathbf{0}$ is a hyperbolic equilibrium of a linear system $\mathbf{x}' = A\mathbf{x}$, let E_s be the span of all (generalized) eigenvectors with $\Re\lambda < 0$. Show that the solution of the IVP, $e^{At}\mathbf{b}$, converges to $\mathbf{0}$ as $t \to \infty$ iff $\mathbf{b} \in E_s$.

Remark: No doubt it has already occurred to you that there is an analogous result for E_u and $t \to -\infty$.

7. For the Lorenz equations (Exercise 2(d) above) with $\rho > 1$, find the stable and unstable subspaces of the saddle point at the origin.

Discussion: In itself, this exercise poses little difficulty. We include it primarily to start you thinking about the Lorenz system. You will find that E_u has dimension one. According to Theorem 6.6.1, the origin has a one-dimensional unstable manifold $\mathcal{M}_u$. In Chapter 9 we will study, analytically and numerically, how $\mathcal{M}_u^{(glob)}$ changes as ρ increases, and we will see that the behavior becomes very complicated indeed. If you wish to get a jump on this work, you could perform such numerical computations now: if $\sigma = 10$ and $b = 8/3$, the real complexity occurs when $24 < \rho < 25$.

8. By finding the explicit solution of (6.59), show that (6.60) defines the stable manifold (where $\mathcal{V}$ may be chosen arbitrarily).

9. For (6.61), a particle in a repulsive cubic force, show that if the initial conditions $\mathbf{b}$ are not in $\mathcal{M}_s^{(glob)}$, then the solution to the IVP blows up in finite (positive) time.

10. *Introduction:* Let $H(x_1, \ldots, x_d, y_1, \ldots, y_d)$ be a smooth function of $2d$ real variables. A system of the form

$$x_j' = \frac{\partial H}{\partial y_j}, \quad y_j' = -\frac{\partial H}{\partial x_j} \qquad (j = 1, \ldots, d), \tag{6.79}$$

is called *Hamiltonian*. A particle moving in an n-dimensional potential $V(\mathbf{x})$ without friction,

$$x_j' = y_j, \quad y_j' = -\partial V/\partial x_j(\mathbf{x}) \qquad (j = 1, \ldots, d),$$

is Hamiltonian with $H(\mathbf{x}, \mathbf{y}) = |\mathbf{y}|^2/2 + V(\mathbf{x})$ equal to the total energy. A classical example is the two-body problem for gravitational attraction, which can be formulated as a Hamiltonian system on $(\mathbb{R}^3 \sim \{\mathbf{0}\}) \times \mathbb{R}^3$ with $V(\mathbf{x}) = -1/|\mathbf{x}|$. On more familiar ground, the frictionless Duffing equation (6.65) and the frictionless pendulum are Hamiltonian.

(a) Show that the frictionless version of (6.74) (i.e., with $\beta = 0$) is Hamiltonian by finding a Hamiltonian.

Hint: If you have already done Exercise 2(c), it may occur to you to consult that problem for ideas. (Even if you haven't done the problem, we've alerted you.)

(b) In the general case, show that $H(\mathbf{x}, \mathbf{y})$ is constant along trajectories of a Hamiltonian system.

Remark: Thus the Hamiltonian is a natural candidate for a Lyapunov function for such equations. If friction is added to a Hamiltonian system of mechanical origin, then typically the Hamiltonian decreases along trajectories, so $H(\mathbf{x}, \mathbf{y})$ is still is a natural candidate for a Lyapunov function.

(c) *Introduction:* A point $(\mathbf{x}_*, \mathbf{y}_*)$ is an equilibrium of a Hamiltonian system iff it is a critical point of $H(\mathbf{x}, \mathbf{y})$, i.e., $\partial H/\partial x_j = \partial H/\partial y_j = 0$. The following part of the problem generalizes Exercise 11, which you might want to do first.

Show that an equilibrium $(\mathbf{x}_*, \mathbf{y}_*)$ of a Hamiltonian system is Lyapunov stable if $H(\mathbf{x}, \mathbf{y})$ has a strict local minimum there. (In particular, if $H(\mathbf{x}, \mathbf{y}) = |\mathbf{y}|^2/2 + V(\mathbf{x})$, this condition means that $\mathbf{y}_* = \mathbf{0}$ and $V(\mathbf{x})$ has a local minimum at $\mathbf{x}_*$.)

Remark: An equilibrium $(\mathbf{x}_*, \mathbf{y}_*)$ of a Hamiltonian system is never asymptotically stable, because $\operatorname{tr} \mathbf{DF}_* = 0$. This follows from equality of mixed second-order partials. So the sum of the eigenvalues of $\mathbf{DF}_*$ is zero.

(d) *Introduction:* Two-dimensional Hamiltonian systems often have many periodic orbits. In light of Part (b), this is easily understood: every orbit is contained in a level set $\{H(x, y) = C\}$, and if these level sets form nested closed curves, one may expect a family of periodic orbits (Cf. Exercise 17).

This behavior carries over to certain non-Hamiltonian systems. A system is called *reparametrized Hamiltonian* if it can be written

$$x_j' = \phi(\mathbf{x}, \mathbf{y}) \frac{\partial H}{\partial y_j}, \quad y_j' = -\phi(\mathbf{x}, \mathbf{y}) \frac{\partial H}{\partial x_j} \qquad (j = 1, \ldots, d), \tag{6.80}$$

where $\phi(\mathbf{x}, \mathbf{y})$ is a nonzero function. The next part of the exercise explains the name.

Suppose $\mathbf{x}(t), \mathbf{y}(t)$ is a solution of a Hamiltonian system (6.79). Show that if $\tau(t)$ satisfies the ODE $d\tau/dt = \phi(\mathbf{x}(t), \mathbf{y}(t))$, then $\mathbf{x}(\tau(t)), \mathbf{y}(\tau(t))$ solves the reparametrized Hamiltonian system.

Remark: Hence although (6.79) and (6.80) have different solutions, these solutions trace out the same orbits.

(e) Show that the Lotka–Volterra equations

$$x' = x - xy, \qquad y' = \rho(xy - y),$$

are reparametrized Hamiltonian with

$$\phi(x, y) = xy, \qquad H(x, y) = -\rho(x - \ln x) - y + \ln y.$$

Remark: This result "explains" why all orbits of the Lotka–Volterra equations are periodic.

(f) Show that the activator–inhibitor equations (6.22) with $\rho = 1$ are reparametrized Hamiltonian with

$$\phi(x, y) = x^2, \qquad H(x, y) = \sigma \ln(1 + y) - \frac{y}{x} - x.$$

Remark: With this information we can describe the phase portrait of (6.22) when $\rho = 1$, the portrait that if $\sigma = 2.1$, interpolates between Panels (a) and (b) in Figure 6.12. Specifically, the unstable manifold $\mathcal{M}_u$ through the saddle point $(x_-, y_-) \approx (0.730, 0.533)$ is contained in the level set

$$\left\{(x, y) : 2.1 \ln(1 + y) - \frac{y}{x} - x = C\right\}, \tag{6.81}$$

where $C = 2.1 \ln(1 + y_-) - y_-/x_- - x_-$. Thus, one half of $\mathcal{M}_u$ connects to the origin, but the other half is a homoclinic orbit that connects back to the saddle point. The region inside (6.81) is filled by a one-parameter family of periodic orbits, which are also level sets of H. (Cf. Figure 7.2.)

6.8.2 Uses of a Lyapunov Function

11. Use the energy $H(x, y) = y^2/2 - x^2/2 + x^4/4$ as a Lyapunov function to show that the equilibria $(\pm 1, 0)$ of Duffing's equation (without friction), (6.65), are Lyapunov stable but not asymptotically stable.

Remark: While you are thinking about (6.65), it would be good practice to use the energy to enumerate the orbits of this equation, as we did for the repulsive cubic (6.61). Interpret the orbits physically in terms of the rolling-marble analogy.

12. *Introduction:* In this exercise you verify some of the features of the phase portrait in Figure 6.10(b) for Duffing's equation (with positive friction), (6.1).

(a) Use the energy $H(x, y) = y^2/2 - x^2/2 + x^4/4$ as a Lyapunov function, together with Lasalle's principle, to show that for every initial condition $\mathbf{b}$ in the set

$$\{(x, y) : x > 0,\ H(x, y) < 0\}, \tag{6.82}$$

the solution $\boldsymbol{\varphi}(t, \mathbf{b})$ tends to $(1, 0)$ as $t \to \infty$. *(Draw this set!)*

Remark: Thus, we have a second proof that the equilibrium $(1,0)$ is asymptotically stable, supplementing the proof based on linearization at the equilibrium. In Part (b) of this problem you use the Lyapunov function to obtain additional information not immediately available from the linearization.

Incidentally, Condition (6.49) is so easy to verify in this example that you might wonder how could it fail: this happens, for example, if $\mathcal{U}_1$ contains another equilibrium of the ODE.

(b) Show that the two halves of the unstable manifold $\mathcal{M}_u^{(glob)}$ through $(0,0)$ converge to the equilibria at $(\pm 1, 0)$.

Hint: Find the unstable eigenvector of $\mathbf{DF}(0,0)$. Use the fact that $\mathcal{M}_u$ is tangent to this eigenvector to show that points on $\mathcal{M}_u$ with $x > 0$ close to $(0,0)$ lie in the set (6.82).

(c) Describe the stable manifold $\mathcal{M}_s^{(glob)}$ through $(0,0)$ in terms of the rolling-marble analogy.

Remarks: (i) Thus, adding friction to (6.65) restores the generic situation that $\mathcal{M}_s$ and $\mathcal{M}_u$ intersect only at the saddle point. *(ii)* Incidentally, $\mathcal{M}_s^{(glob)}$ defines the boundary between the intertwined basins of attraction of $(1,0)$ and $(-1,0)$.

13. *Introduction:* As you will see in this exercise, sometimes a Lyapunov function may be used to prove that an equilibrium is globally attracting. Say that a function $L : \mathcal{U} \to \mathbb{R}$ tends to infinity at the boundary of $\mathcal{U}$ if for every $M > 0$ there is a compact subset $\mathcal{K} \subset \mathcal{U}$ such that $L(\mathbf{x}) > M$ on $\mathcal{U} \sim \mathcal{K}$.

(a) Show that if $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ has a strict Lyapunov function $L : \mathcal{U} \to [0, \infty)$ that tends to infinity at the boundary of $\mathcal{U}$, then this ODE has a unique equilibrium in $\mathcal{U}$ that moreover is globally attracting for any initial conditions in $\mathcal{U}$.

(b) Show for the Lorenz equations (6.76) that if $\rho < 1$, then

$$L(x, y, z) = x^2/\sigma + y^2 + z^2$$

is a strict Lyapunov function. Use this to deduce that the origin is *globally* asymptotically stable when $\rho < 1$, an improvement over Exercise 2(d).

Remark: This part of the problem suggests another technique for finding Lyapunov functions: play around with an arbitrary positive definite quadratic form to see whether you can make its derivative negative.

(c) *Introduction:* Note that, provided $K > 1$, the Lyapunov function

$$L(x, y) = \rho(x - \ln x) + y - (1 - 1/K) \ln y$$

of Section 6.5.3 tends to infinity at the boundary of the open first quadrant.

Combine Lasalle's invariance principle with your proof in Part (a) to prove that the equilibrium $(1, 1 - 1/K)$ of (6.50) is globally attracting for initial data in the open first quadrant.

Remark: The parameters considered for (6.50), i.e., $\varepsilon = 0$, $K > 1$, lie on the boundary of Region II in Figure 6.13. You might find it amusing to contemplate how as $\varepsilon \to 0$, the phase portrait in Panel (a) of the figure can evolve into one in which all orbits in the first quadrant converge to a single equilibrium.

6.8.3 Phase-Portrait Exercises

See also Exercise 21 below.

14. *Introduction:* This problem illustrates for Figure 6.12 how theory may be used to enhance the accuracy of phase-portrait sketches. We have already remarked that the unstable manifold spirals into $\mathbf{P}_+$ if (6.73) holds, so we focus on $\mathbf{P}_-$.

    (a) Draw the nullclines and flow-quadrant diagram for the activator–inhibitor system (6.22), paying special attention to the region around $\mathbf{P}_-$.

    (b) Argue from the flow-quadrant diagram that the unstable manifold leaves the saddle point $\mathbf{P}_-$ in the wedge between the horizontal direction and the tangent to the y-nullcline $y = x^2$.

    (c) Argue from the flow-quadrant diagram that the stable manifold converges to the saddle point $\mathbf{P}_-$ in the wedge between the tangent to the x-nullcline $y = \sigma x - 1$ and the vertical direction.

    Remark: The dedicated reader can improve on Part (b) by evaluating the Jacobian at $\mathbf{P}_-$ and calculating the unstable eigenvector, which is tangent to the unstable manifold. Similarly for Part (c) and the stable eigenvector.

15. *Introduction:* The following is a purely made-up, easy exercise to practice on.

    Sketch the phase portrait of the system

    $$\begin{aligned} x' &= y - x^2, \\ y' &= x - y. \end{aligned}$$

    That is, locate all equilibria, classify them according to Table 2.1, sketch the stable and unstable manifolds of any saddle points, and indicate a few other trajectories.

16. *Introduction:* Recall from Exercise 4.3(b) the equations for the evolution of two interacting species,

    $$x' = x\,(1 - x - by)\,, \qquad y' = \rho y\,(1 - y - cx)\,.$$

    For all parameter values, this system has equilibria at $(0,0)$, $(0,1)$, and $(1,0)$. Moreover, in each of the three cases (i) $b, c > 1$, (ii) $0 < b, c < 1$, and (iii) $-1 < b, c < 0$, there is also a coexistence equilibrium of this system in the open first quadrant.

(a) For each of the three cases, determine the stability of all equilibria of this system.

(b) For each case, use *what-else-could-it-be?* "logic" to sketch phase portraits, including the stable and unstable manifolds through all saddle points.

Remark: If you're feeling stuck, you could calculate eigenvectors of **DF** at the saddle point.

(c) For each case, identify the basins of attraction of all asymptotically stable equilibria.

17. Use the energy $H(x,y) = y^2/2 - \cos x$ to construct the phase portrait of the frictionless pendulum

$$\begin{aligned} x' &= y, \\ y' &= -\sin x. \end{aligned}$$

Identify the stable and unstable manifolds through saddle points.

Remark: If the equation is considered on the cylinder $\mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R}$, the stable and unstable manifolds through the saddle point $(\pi, 0)$ coincide. In other words, this equation gives another example of a homoclinic orbit.

6.8.4 PHD Exercises

18. Prove Proposition 6.2.1.

Hint: Suppose the equilibrium $\mathbf{b}_*$ is at the origin. Write $\mathbf{F}(\mathbf{x}) = A\mathbf{x} + \mathbf{r}(\mathbf{x})$, where $\mathbf{r}(\mathbf{x}) = o(|\mathbf{x}|)$. Applying an appropriate similarity transformation, without loss of generality you may assume that A is block diagonal,

$$A = \begin{bmatrix} A_s & 0 \\ 0 & A_u \end{bmatrix},$$

where the eigenvalues of these two submatrices satisfy

$$\Re\lambda_k(A_s) \le 0, \qquad \Re\lambda_k(A_u) > 0. \tag{6.83}$$

Decompose vectors $\mathbf{x} = (\mathbf{x}_s, \mathbf{x}_u)$ into stable and unstable components. Given positive constants η, δ, define a truncated cone around the unstable subspace,

$$\Gamma_{\eta,\delta} = \{\mathbf{x} \in \mathbb{R}^d : \ |\mathbf{x}_s|^2 \le \eta |\mathbf{x}_u|^2, \quad |\mathbf{x}| \le \delta\}.$$

(Make appropriate sketches for the case that $\mathbf{x}_s$ and $\mathbf{x}_u$ are one-dimensional.) Your primary task is to show that η and δ may be chosen small enough to obtain the following two estimates:

- There is an $\varepsilon > 0$ such that for all $\mathbf{x} \in \Gamma_{\eta,\delta}$,

$$\langle \mathbf{x}, \mathbf{F}(\mathbf{x}) \rangle \ge \varepsilon |\mathbf{x}|^2. \tag{6.84}$$

- On the conical portion of $\partial\Gamma_{\eta,\delta}$,

$$\{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x}_s|^2 = \eta|\mathbf{x}_u|^2,\ |\mathbf{x}| \le \delta\}, \tag{6.85}$$

the inward normal $\mathbf{N} = (-\mathbf{x}_s, \eta\mathbf{x}_u)$ satisfies

$$\langle \mathbf{N}, \mathbf{F}(\mathbf{x})\rangle \ge 0,$$

with equality only at the origin.

Then, since $(d/dt)|\mathbf{x}(t)|^2 = 2\langle \mathbf{x}, \mathbf{F}(\mathbf{x})\rangle$, it follows from the first estimate that for as long as $\mathbf{x}(t) \in \Gamma_{\eta,\delta}$, a solution satisfies $|\mathbf{x}(t)|^2 \ge |\mathbf{x}(0)|^2 e^{2\varepsilon t}$. On the other hand, it follows from the second estimate, as in the analysis of trapping regions, a solution starting in $\Gamma_{\eta,\delta}$ cannot cross the conical portion of the boundary. Thus, the solution of an IVP with nonzero initial conditions in $\Gamma_{\eta,\delta}$ arbitrarily close to the origin will grow until it eventually leaves the sphere of radius δ; i.e., the equilibrium is not Lyapunov stable.

The derivation of the estimates is based on condition (6.83). Given this condition, it follows from Proposition C.2.1 that, possibly after applying another similarity transformation, there is an $\varepsilon_1 > 0$ such that

$$(a)\ \langle \mathbf{x}_s, A_s\mathbf{x}_s\rangle \le (\varepsilon_1/4)|\mathbf{x}_s|^2, \qquad (b)\ \langle \mathbf{x}_u, A_u\mathbf{x}_u\rangle \ge \varepsilon_1|\mathbf{x}_u|^2. \tag{6.86}$$

For (6.84) we have

$$\langle \mathbf{x}, \mathbf{F}(\mathbf{x})\rangle = \langle \mathbf{x}_s, A_s\mathbf{x}_s\rangle + \langle \mathbf{x}_u, A_u\mathbf{x}_u\rangle + \langle \mathbf{x}, \mathbf{r}(\mathbf{x})\rangle.$$

According to (6.86b), the middle term on the RHS is positive. To obtain (6.84), choose η small to control the size of the first term and invoke the fact that $\mathbf{r}(\mathbf{x}) = o(|\mathbf{x}|)$ to control the third. Similarly, for (6.85) we have

$$\langle \mathbf{N}, \mathbf{F}(\mathbf{x})\rangle = -\langle \mathbf{x}_s, A_s\mathbf{x}_s\rangle + \eta\langle \mathbf{x}_u, A_u\mathbf{x}_u\rangle + \langle \mathbf{N}, \mathbf{r}(\mathbf{x})\rangle.$$

By (6.86), if $\mathbf{x} \in \partial\Gamma_{\eta,\delta}$, the first two terms satisfy

$$-\langle \mathbf{x}_s, A_s\mathbf{x}_s\rangle + \eta\langle \mathbf{x}_u, A_u\mathbf{x}_u\rangle \ge \frac{3}{4}\eta\varepsilon_1|\mathbf{x}_s|^2.$$

To obtain (6.85), invoke the fact that $\mathbf{r}(\mathbf{x}) = o(|\mathbf{x}|)$ to control the third term.

19. Construct an alternative proof of Theorem 6.1.1 using a fixed-point argument, as in the proof of the stable manifold theorem, Theorem 6.6.1.

Remark: Although this exercise requires considerable effort, it could help you penetrate the messy details of the proof of Theorem 6.6.1.

20. Suppose (6.28) is generalized to include diffusion of both species,

$$\begin{aligned}
(a)\quad x_1' &= \sigma x_1^2/(1+y_1) - x_1 + D_1(x_2 - x_1),\\
(b)\quad y_1' &= \rho\,[x_1^2 - y_1] + D_2(y_2 - y_1),\\
(c)\quad x_2' &= \sigma x_2^2/(1+y_2) - x_2 + D_1(x_1 - x_2),\\
(d)\quad y_2' &= \rho\,[x_2^2 - y_2] + D_2(y_1 - y_2).
\end{aligned}$$

Assuming $\sigma > 2$ and $\rho > 1$ as in Section 6.3.2, determine for what values of the diffusion coefficients the equal-concentration equilibrium (x_+, y_+, x_+, y_+) is stable.

21. Sketch the phase portrait, including stable and unstable manifolds, for the torqued pendulum

$$\begin{aligned} x' &= y, \\ y' &= \mu - \sin x - \beta y, \end{aligned}$$

when the torque is just below critical, i.e., $\mu < 1$ but barely so.

Suggestion: If you feel stuck on this problem, the phase portrait appears in Figure 9.6(c). It is instructive to ask what happens to the center manifold in Figure 6.14(a) when μ is perturbed.

22. *Introduction:* This problem gives you another example of a homoclinic orbit in a different context. The *Korteweg–de Vries (KdV) equation*

$$u_t + 6uu_x + u_{xxx} = 0, \qquad x \in \mathbb{R}, t > 0, \tag{6.87}$$

is a partial differential equation that has been used to model surface waves in shallow water. We seek a *soliton* solution of the KdV equation. Such a solution has *traveling-wave* form $u(x,t) = v(x - ct)$, where c, the wave speed, is a parameter to be determined; without loss of generality we may assume $c > 0$. Let $\xi = x - ct$. For a soliton, v, v', and v'' all tend to 0 as $\xi \to \pm\infty$, where a prime indicates differentiation with respect to ξ.

(a) Show that the PDE for $v(x - ct)$ reduces to the ODE $v''' + 6vv' - cv' = 0$.

(b) Integrate the equation once to show that $v'' + 3v^2 - cv = 0$.

Remark: The integration constant must vanish for the solution to decay at infinity.

(c) Let $w = v'$ and derive the first-order system that v, w satisfy. Show that this system is Hamiltonian with $H(v,w) = w^2/2 - cv^2/2 + v^3$.

(d) Deduce from the Hamiltonian structure that for every wave speed, the global stable and unstable manifolds through the saddle point at the origin intersect to form a homoclinic orbit.

Discussion: Since this construction works for every value of c, we get a one-parameter family of soliton solutions of the PDE. Changing the viewpoint, we may say that the speed of a soliton depends on its amplitude.

Incidentally, this exercise illustrates a common technique: looking for a solution of a PDE with a specific form often leads to an ODE.

23. *Introduction:* For certain parameter values, the FitzHugh–Nagumo equations (5.42),

$$\begin{aligned} \varepsilon x' &= x(1 - x^2) - y + I, \\ y' &= x - \gamma y, \end{aligned} \tag{6.88}$$

where I, ε, γ are parameters with ε small and positive, provide an example of what is called an *excitable system*. To explain this term: the equations have a sink (x_*, y_*) (so given initial conditions that are a *sufficiently small* perturbation of (x_*, y_*), the solution simply will decay back to the sink), but for certain initial conditions that differ from (x_*, y_*) only by $\mathcal{O}(\varepsilon)$, the solution of the IVP undergoes a long-lived, large-amplitude excursion before it returns to (x_*, y_*). Such behavior is central to the function of nerve cells.

(a) Supposing that $0 < \gamma < 1$, sketch the nullclines of (6.88) and show that it has a unique equilibrium solution.

(b) Deduce from Table 6.2 that the unique equilibrium is either a sink or a source.

(c) Calculate from the Jacobian $\mathbf{DF}_*$ that if the equilibrium is located at (x_*, y_*), then

$$\operatorname{tr} \mathbf{DF}_* = \frac{1 - 3x_*^2}{\varepsilon} - \gamma. \tag{6.89}$$

(d) *Introduction:* Excitability occurs if $x_* \approx -1/\sqrt{3}$, i.e., if an equilibrium (which by (6.89) is a sink) lies near the local minimum of the x-nullcline. For definiteness, let's choose I, γ such that $x_* = -1/\sqrt{3}$ exactly. Then the second equation requires that $y_* = -1/\gamma\sqrt{3}$, and the first that $I = 2/3\sqrt{3} + y_*$. We can achieve this with simple numbers if we let $\gamma = 3/5$ and $I = -1/\sqrt{3}$.

To demonstrate excitability numerically, solve the IVP for (6.88), say with

$$x(0) = -1\sqrt{3}, \quad y(0) = -5/3\sqrt{3} - \delta,$$

and I, γ as above. Choose your own δ as small as you wish.

Remark: Excitability in (6.88) can be understood through analysis of the nullclines as a fast–slow system. This analysis is similar to what will be developed in Section 7.6.3, and we propose it as Exercise 7.24.

24. *Comment:* How's this for a change of pace?

The Heisenberg uncertainty principle imposes an upper bound $T_{\max}$ on how long you can reliably stand a pencil on its point; estimate this time.

Hint: Model the pencil as a frictionless pendulum, governed by the equation $\theta'' + (g/\ell)\sin\theta = 0$. Behavior near the vertical is well approximated by the linearization[20] of this equation at $\theta = \pi$, i.e., $\theta'' - \omega^2\theta = 0$, where $\omega = \sqrt{g/\ell}$. With initial conditions $\theta(0) = a, \quad \theta'(0) = b$, the linear IVP has the solution

$$\theta(t) = A(a,b)e^{\omega t} + B(a,b)e^{-\omega t}, \tag{6.90}$$

[20] In case you were wondering: linearization is the issue that connects this problem to the rest of the present chapter: for most of the time while the solution is growing, θ is so close to vertical that solutions of the full equation and of the linearization could not be distinguished on a computer screen.

the sum of growing and decaying exponentials. The Heisenberg uncertainty principle,

$$\Delta x\, \Delta p \geq \hbar/2, \tag{6.91}$$

where $\hbar \approx 1.05 \times 10^{-34}$joule-sec is Planck's constant divided by 2π, limits the precision with which initial displacement and velocity may be specified; you are able to only require that

$$l|a| \leq \Delta x, \quad ml|b| \leq \Delta p,$$

with Δx, Δp subject to (6.91). If you choose an accuracy Δx for the position, then in the best case, $\Delta p = \hbar/2\Delta x$. Thus for a given Δx,

$$\varepsilon(\Delta x) = \max_{|a|\leq \Delta x/\ell} \max_{|b|\leq \hbar/2m\ell\Delta x} |A(a,b)|$$

is the smallest value that can be guaranteed for the coefficient of the growing exponential in (6.90). Calculate $\varepsilon_{\min} = \min_{\Delta x} \varepsilon(\Delta x)$ and estimate $T_{\max}$ by solving $\varepsilon_{\min} e^{\omega t} = 1$, the time when the optimized solution becomes of order unity.

25. Verify the claim in the Pearls that the two ODEs in (6.94) are topologically equivalent.

Hint: Rewrite both equations in polar coordinates. Given initial conditions $\mathbf{b} = (r_0, \theta_0)$, then we have

$$\boldsymbol{\varphi}_1(t, \mathbf{b}) = (e^{-2t} r_0, \theta_0) \qquad \text{and} \qquad \boldsymbol{\varphi}_2(t, \mathbf{b}) = (e^{-t} r_0, \theta_0 + t)$$

for the solutions of the $\mathbf{x}$ and $\mathbf{y}$ equations, respectively. Prove that the map (in polar coordinates)

$$\Psi(r, \theta) = (\sqrt{r}, \theta - (\ln r)/2)$$

is continuous and that (6.93) is satisfied for this map.

26. (a) Find the global stable manifold for the equilibrium $(1, 0)$ of (6.13)

    (b) Describe all local center manifolds (defined in Section 6.11) through the equilibrium.

Discussion: The natural choice for the center manifold, which moreover is global, is the homoclinic orbit, $\{r = 1, 0 < \theta < 2\pi\}$. Although you could obtain parametric representations for the other choices by solving the ODE, we say don't bother. It is sufficient to recognize the existence of these other possible choices and to describe their behavior.

6.9 Pearls of Wisdom

6.9.1 Miscellaneous

Theorem 6.1.1 is such an important result that we urge you not only to know how to apply it, but also to develop and retain some understanding about how it is proved.

For example, be prepared to reply cogently if a few months from now a stranger stops you on the street and asks, *"Hey, how do you prove that an equilibrium of an ODE is asymptotically stable if all eigenvalues of the Jacobian are negative there?"*

The term *gradient system* refers to an ODE of the form

$$x_j' = -\frac{\partial V}{\partial x_j}, \qquad j = 1, \ldots, d,$$

where $V : \mathcal{U} \to \mathbb{R}$ is a smooth function; in vector notation, $\mathbf{x}' = -\nabla V(\mathbf{x})$. You can easily show that for such a system, (i) V decreases along orbits and (ii) a point $\mathbf{b}_*$ is an equilibrium iff $\nabla V(\mathbf{b}_*) = \mathbf{0}$. Thus, V provides a Lyapunov function for an equilibrium of a gradient system, *provided* V has a strict local minimum there.

Incidentally, a Hamiltonian system can be written as a matrix times a gradient,

$$\begin{bmatrix} \mathbf{x}' \\ \mathbf{y}' \end{bmatrix} = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \begin{bmatrix} \nabla_x H(\mathbf{x}, \mathbf{y}) \\ \nabla_y H(\mathbf{x}, \mathbf{y}) \end{bmatrix}.$$

6.9.2 The Hartman–Grobman Theorem and Topological Conjugacy

As we have seen in this chapter, the flow of a nonlinear system $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ near a hyperbolic equilibrium resembles the flow of the linearization. This resemblance is made precise in the *Hartman–Grobman theorem*, which is proved in [37]; see also [63] for a readable detailed sketch of the proof.

Theorem 6.9.1. *Suppose $\mathbf{b}_*$ is a hyperbolic equilibrium point for $\mathbf{x}' = \mathbf{F}(\mathbf{x})$. Let $\boldsymbol{\varphi}$ be the flow map for this equation and let $A = \mathbf{DF}_*$. Then there exist a neighborhood $\mathcal{U}$ of $\mathbf{b}_*$ and a continuous map $\boldsymbol{\Psi} : \mathcal{U} \to \mathbb{R}^d$, where $\boldsymbol{\Psi}(\mathbf{b}_*) = \mathbf{0}$, that is a homeomorphism onto its range such that*

$$\boldsymbol{\Psi}(\boldsymbol{\varphi}(t, \mathbf{b})) = e^{tA}\boldsymbol{\Psi}(\mathbf{b}) \tag{6.92}$$

for all $\mathbf{b} \in \mathcal{U}$ and all times such that $\boldsymbol{\varphi}(t, \mathbf{b})$ is defined and belongs to $\mathcal{U}$.

A noteworthy consequence of this result: the stable and unstable manifolds of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ at a hyperbolic equilibrium $\mathbf{b}_*$ are equal to $\boldsymbol{\Psi}^{-1}(E_s)$ and $\boldsymbol{\Psi}^{-1}(E_u)$, respectively, where E_s and E_u are the stable and unstable *subspaces* of A.

Theorem 6.9.1 is definitely false at a nonhyperbolic equilibrium, which may be seen by comparing either of the scalar ODEs $x' = \pm x^3$ to its linearization, $w' = 0$.

The theorem motivates the following definition: Consider two ODEs, say $\mathbf{x}' = \mathbf{F}_1(\mathbf{x})$ and $\mathbf{y}' = \mathbf{F}_2(\mathbf{y})$, on open subsets $\mathcal{U}_1$ and $\mathcal{U}_2$ of $\mathbb{R}^d$, respectively, and let these ODEs have flows $\boldsymbol{\varphi}_1(t, \mathbf{b})$ and $\boldsymbol{\varphi}_2(t, \mathbf{b})$, respectively. We call these two flows *topologically conjugate* if there is a homeomorphism $\boldsymbol{\Psi}$ from $\mathcal{U}_1$ onto $\mathcal{U}_2$ such that for all $(t, \mathbf{b}) \in \mathbb{R} \times \mathcal{U}_1$,

$$\boldsymbol{\Psi}(\boldsymbol{\varphi}_1(t, \mathbf{b})) = \boldsymbol{\varphi}_2(t, \boldsymbol{\Psi}(\mathbf{b})). \tag{6.93}$$

In case solutions are not global, equality in (6.92) means that if either side has meaning, then both sides have meaning and are equal. Thus, we may paraphrase the conclusion of Theorem 6.9.1 to state that on some neighborhood of a hyperbolic equilibrium, $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ is topologically conjugate to (the restriction to an appropriate neighborhood of zero of) its linearization.

Note that topological conjugacy requires only that the homeomorphism $\boldsymbol{\Psi}$ be *continuous*. This lack of smoothness is a serious issue that undercuts much of the initial appeal of the theorem. For example, in Exercise 25 we guide you through a proof that the two systems (both on the unit disk, say)

$$\mathbf{x}' = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix} \mathbf{x} \qquad \text{and} \qquad \mathbf{y}' = \begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix} \mathbf{y} \tag{6.94}$$

are topologically conjugate, even though the x-flow converges along radial lines to the origin, while the y-flow has spirals that encircle the origin infinitely many times as they converge, less rapidly, to the origin. Of course, these two systems are definitely not conjugate via a *differentiable* homeomorphism. In the first place, if (6.93) is satisfied with a $\mathcal{C}^1$ diffeomorphism, then $\mathbf{DF}_1$ and $\mathbf{DF}_2$ have the same eigenvalues at corresponding critical points. But more emphatically, no differentiable homeomorphism could map radial lines onto spirals.

In general, asking for the homeomorphism in the theorem to have one or more derivatives raises technical issues involving "nonresonance conditions" on the eigenvalues of $\mathbf{DF}_*$. The discussion in Section 19.12 of Wiggins [95] may help you understand how these conditions arise. In case the equilibrium $\mathbf{b}_*$ is asymptotically stable, these complications can be sidestepped: Hartman [36] has shown that *if* $\mathbf{F}$ *is* $\mathcal{C}^2$ *and if* $\Re\lambda_j(\mathbf{DF}_*) < 0$, *then there is a* $\mathcal{C}^1$ *diffeomorphism* $\boldsymbol{\Psi}$ *such that* (6.92) *is satisfied.*

6.9.3 Structural Stability

In topological conjugacy, (6.93) requires that trajectories match up point by point, *including the parametrization by time.* Structural stability is based on a weaker notion, topological *equivalence*, in which only orbits need match up. Specifically, two ODEs $\mathbf{x}' = \mathbf{F}_1(\mathbf{x})$ and $\mathbf{y}' = \mathbf{F}_2(\mathbf{y})$ as above are called *topologically equivalent* if there is a homeomorphism $\boldsymbol{\Psi} : \mathcal{U}_1 \to \mathcal{U}_2$ such that for each $\mathbf{b} \in \mathcal{U}_1$, $\boldsymbol{\Psi}$ maps the orbit[21] $\boldsymbol{\varphi}_1(\mathbb{R}, \mathbf{b})$ of $\mathbf{x}' = \mathbf{F}_1(\mathbf{x})$ through $\mathbf{b}$ onto the orbit $\boldsymbol{\varphi}_2(\mathbb{R}, \boldsymbol{\Psi}(\mathbf{b}))$ of $\mathbf{y}' = \mathbf{F}_2(\mathbf{y})$ through $\boldsymbol{\Psi}(\mathbf{b})$, preserving the direction of flow along orbits. For example, the two ODEs

[21] Although we write $\boldsymbol{\varphi}_1(\mathbb{R}, \mathbf{b})$, only times such that $\boldsymbol{\varphi}_1(t, \mathbf{b})$ is defined are to be substituted.

$$(a)\ \mathbf{x}' = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\mathbf{x} - |\mathbf{x}|^2\mathbf{x}, \qquad (b)\ \mathbf{y}' = \begin{bmatrix} 1 & -a \\ a & 1 \end{bmatrix}\mathbf{y} - |\mathbf{y}|^2\mathbf{y}, \tag{6.95}$$

are topologically equivalent if $a > 0$ but not topologically conjugate unless $a = 1$. The problem is that the periodic orbit $|\mathbf{x}| = 1$ of (6.95a) has period 2π, while the corresponding orbit $|\mathbf{y}| = 1$ of (6.95b) has period $2\pi/a$; i.e., the parametrizations of trajectories do not match.

An ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ on an open set $\mathcal{U} \subset \mathbb{R}^d$ is called *structurally stable* if it is topologically equivalent to every other ODE $\mathbf{y}' = \mathbf{G}(\mathbf{y})$ on $\mathcal{U}$ that is $\mathcal{C}^1$-close to it, i.e., for which

$$\sup_{\mathbf{x}\in\mathcal{U}} |\mathbf{G}(\mathbf{x}) - \mathbf{F}(\mathbf{x})| \qquad \text{and} \qquad \sup_{\mathbf{x}\in\mathcal{U}} \max_j |\partial_j \mathbf{G}(\mathbf{x}) - \partial_j \mathbf{F}(\mathbf{x})|$$

are sufficiently small. For example, it follows from Peixoto's theorem (see Section 1.9 of [33]) that (6.95a) is structurally stable on $\mathbb{R}^2$. By contrast, the Lotka–Volterra system (1.39) is not structurally stable.

Historically, there was a period in which structurally stable ODEs were a primary focus of the theory. The thinking was that if a model of a physical system was not structurally stable, the inevitable errors in the model, no matter how small, might pose perturbations that invalidated conclusions drawn from the model. In response to better understanding of chaotic solutions of higher-dimensional ODEs, this dogma is currently regarded as too naive. (Cf. the discussion on pp. 258–259 of [33].)

6.10 Appendix 1: Partial Proof of Theorem 6.6.1

Our proof is nearly complete; we merely refer you to an external reference for a couple of technical points at the end of the argument.

6.10.1 Reformulation of the IVP as an Integral Equation

The construction of the stable manifold of an equilibrium is facilitated by two reductions. First, we translate coordinates so that the equilibrium is at the origin, $\mathbf{b}_* = \mathbf{0}$, and then we change coordinates to separate the eigenvectors of $\mathbf{DF}(\mathbf{0})$ associated with eigenvalues having positive and negative real parts. After performing an appropriate similarity transformation, we may assume without loss of generality that $\mathbf{DF}(\mathbf{0})$ has block-diagonal form, and we may write a Taylor series expansion

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} B & 0 \\ 0 & -C \end{bmatrix}\mathbf{x} + \mathbf{r}(\mathbf{x}), \tag{6.96}$$

where B and C are (square) matrices of dimensions d_s and $d - d_s$ whose eigenvalues have negative real parts and where $\mathbf{r}(\mathbf{x}) = o(|\mathbf{x}|)$ near the origin. By Corollary 2.4.4, possibly performing an additional similarity transformation, we may assume that there is a constant $\varepsilon > 0$ such that

$$\left\|e^{Bt}\right\| \le e^{-\varepsilon t}, \quad \left\|e^{Ct}\right\| \le e^{-\varepsilon t} \qquad (t \ge 0). \tag{6.97}$$

Theorem 6.6.1 concerns the IVP

$$\mathbf{x}' = \begin{bmatrix} B & 0 \\ 0 & -C \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{p}(\mathbf{x}) \\ \mathbf{q}(\mathbf{x}) \end{bmatrix}, \qquad \mathbf{x}(0) = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}, \tag{6.98}$$

where we have decomposed both the remainder $\mathbf{r}(\mathbf{x})$ and the initial condition into a d_s-dimensional "stable" component plus a $(d - d_s)$-dimensional "unstable" component; i.e., we let

$$\mathbf{x}^{(stb)} = \begin{bmatrix} x_1 \\ \dots \\ x_{d_s} \end{bmatrix}, \qquad \mathbf{x}^{(nst)} = \begin{bmatrix} x_{d_s+1} \\ \dots \\ x_d \end{bmatrix}.$$

We suggest that you keep the concrete academic example (6.59) handy as you read the remainder of this section.[22] Rewriting (6.59) in the notation of (6.98), note that $B = C = -1$ and the components of the remainder term are scalar-valued functions $p(x, y) = 0$ and $q(x, y) = x^2$.

The proof of the theorem is based on an integral equation, which we now formulate.

Proposition 6.10.1. *If* $\mathbf{x}(t)$ *satisfies* (6.98) *for* $0 \le t < \infty$ *and if* $\sup_{t\ge 0} |\mathbf{x}(t)| < \infty$, *then*

$$\mathbf{x}(t) = \begin{bmatrix} e^{Bt}\mathbf{c} + \displaystyle\int_0^t e^{(t-s)B}\, \mathbf{p}(\mathbf{x}(s))\, ds \\ \\ -\displaystyle\int_t^\infty e^{(s-t)C}\, \mathbf{q}(\mathbf{x}(s))\, ds \end{bmatrix}. \tag{6.99}$$

Conversely, if $\mathbf{x}(t)$ *is a bounded function that satisfies* (6.99), *then it satisfies the ODE in* (6.98).

Remark. Regarding initial conditions, a solution of (6.99) satisfies $\mathbf{x}^{(stb)}(0) = \mathbf{c}$, but no simple formula for $\mathbf{x}^{(nst)}(0)$ is available.

Proof of Proposition 6.10.1. The first component of (6.99) follows by integrating the first component of the ODE (6.98) over $(0, t)$, as in the derivation of (6.6). For the second component we integrate over (t, ∞). More precisely, we rewrite the second component of (6.98) as

$$\frac{d}{dt}\left(e^{Ct}\mathbf{x}^{(nst)}\right) = e^{Ct}\mathbf{q}(\mathbf{x}(t)),$$

[22] If you continually refer back to example (6.59) throughout the technical results that follow, the main payoff arrives once you reach equation (6.107). Namely, you'll find that $\psi(c) = -c^2/3$ (scalar-valued for this toy example), in agreement with formula (6.60) for the stable manifold.

integrate over (t, T) to deduce

$$e^{CT}\mathbf{x}^{(nst)}(T) - e^{Ct}\mathbf{x}^{(nst)}(t) = \int_t^T e^{Cs}\mathbf{q}(\mathbf{x}(s))\,ds,$$

and use the fact that $\mathbf{x}(t)$ is bounded together with (6.97) to obtain (6.99) in the limit $T \to \infty$.

Regarding the converse, it follows by differentiating (6.99) that the ODE in (6.98) is satisfied. □

6.10.2 Fixed-Point Analysis

When we used the integral equation (6.6) to prove Theorem 6.1.1, we already knew that this equation had a solution for at least some range of t and this solution was unique; we needed only to estimate the decay of the solution. By contrast, for (6.99), existence and uniqueness must be derived before other issues are addressed. A fixed-point argument is the appropriate tool for this job.

Let $\mathcal{CB}([0,\infty), \mathbb{R}^d)$ be the space of bounded continuous vector-valued functions on the half-line, which is a Banach space with respect to the sup norm,

$$\|\mathbf{x}\|_{\sup} = \sup_{0 \le t < \infty} |\mathbf{x}(t)|.$$

For $\mathbf{c} \in \mathbb{R}^{d_s}$ with $|\mathbf{c}|$ sufficiently small, we will define an integral operator on a subset S of $\mathcal{CB}([0,\infty), \mathbb{R}^d)$ by the formula

$$\mathfrak{T}_{\mathbf{c}}[\mathbf{x}](t) = \begin{bmatrix} e^{Bt}\mathbf{c} + \displaystyle\int_0^t e^{(t-s)B}\,\mathbf{p}(\mathbf{x}(s))\,ds \\ \\ \displaystyle -\int_t^\infty e^{(s-t)C}\,\mathbf{q}(\mathbf{x}(s))\,ds \end{bmatrix}. \tag{6.100}$$

The domain S will be chosen with the help of the following lemma.

Lemma 6.10.2. *If $\eta > 0$, there is a positive constant δ such that for every two points in $\mathbb{R}^d$ with $|\mathbf{x}|$, $|\mathbf{y}| \le \delta$, we have*

$$|\mathbf{r}(\mathbf{x}) - \mathbf{r}(\mathbf{y})| \le \eta|\mathbf{x} - \mathbf{y}|.$$

Remark. From the definition of little-o, we know that we can satisfy $|\mathbf{r}(\mathbf{x})| \le \eta|\mathbf{x}|$; the lemma asserts a little more.

Proof of Lemma 6.10.2. Since $\mathbf{r}(\mathbf{x}) = o(|\mathbf{x}|)$ and is $\mathcal{C}^1$, its Jacobian $\mathbf{Dr}(\mathbf{x})$ is $o(1)$ near $\mathbf{x} = \mathbf{0}$. Thus, there is a $\delta > 0$ such that

$$\|\mathbf{Dr}(\mathbf{x})\| \leq \eta \quad \text{provided } |\mathbf{x}| < \delta.$$

The lemma follows by integrating along the line between $\mathbf{x}$ and $\mathbf{y}$, as in the proof of Lemma 3.2.3. □

To define the domain of $\mathfrak{T}_\mathbf{c}$, choose $\eta > 0$ such that

$$\eta < \frac{\varepsilon}{4}, \tag{6.101}$$

where ε is defined by (6.97). Then choose δ as in the lemma, and let

$$\mathsf{S} = \{\mathbf{x} \in \mathcal{CB}([0,\infty), \mathbb{R}^d) : \|\mathbf{x}\|_{\sup} \leq \delta\}. \tag{6.102}$$

Please verify that if $\mathbf{x} \in \mathsf{S}$, then $\mathfrak{T}_\mathbf{c}[\mathbf{x}] \in \mathcal{CB}([0,\infty), \mathbb{R}^d)$.

Proposition 6.10.3. *If $|\mathbf{c}| < \delta/2$, then $\mathfrak{T}_\mathbf{c}$ is a contraction that maps S into itself.*

Proof. If $\mathbf{x}, \mathbf{y} \in \mathsf{S}$, then by the lemma,

$$|\mathbf{p}(\mathbf{x}(s)) - \mathbf{p}(\mathbf{y}(s))| \leq \eta |\mathbf{x}(s) - \mathbf{y}(s)| \leq \eta \|\mathbf{x} - \mathbf{y}\|_{\sup}, \tag{6.103}$$

and similarly for $\mathbf{q}(\mathbf{x})$. Substituting this estimate into (6.100) and using (6.97) to estimate the exponentials, we calculate that

$$|\mathfrak{T}_\mathbf{c}[\mathbf{x}] - \mathfrak{T}_\mathbf{c}[\mathbf{y}]|(t) \leq \int_0^t e^{-\varepsilon(t-s)} \left(\eta \|\mathbf{x} - \mathbf{y}\|_{\sup}\right) ds + \int_t^\infty e^{-\varepsilon(s-t)} \left(\eta \|\mathbf{x} - \mathbf{y}\|_{\sup}\right) ds. \tag{6.104}$$

Regarding the first integral,

$$\int_0^t e^{-\varepsilon(t-s)} ds = \frac{1 - e^{-\varepsilon t}}{\varepsilon} \leq \frac{1}{\varepsilon},$$

and the second integral equals $1/\varepsilon$ exactly. Thus,

$$|\mathfrak{T}_\mathbf{c}[\mathbf{x}] - \mathfrak{T}_\mathbf{c}[\mathbf{y}]|(t) \leq \frac{2\eta}{\varepsilon} \|\mathbf{x} - \mathbf{y}\|_{\sup},$$

both terms in (6.104) contributing equally. Taking the supremum over t and inserting our choice (6.101) of η, we calculate that

$$\|\mathfrak{T}_\mathbf{c}[\mathbf{x}] - \mathfrak{T}_\mathbf{c}[\mathbf{y}]\|_{\sup} \leq \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_{\sup}. \tag{6.105}$$

Now suppose $|\mathbf{c}| < \delta/2$ and $\mathbf{x} \in \mathsf{S}$. By the triangle inequality,

$$\|\mathfrak{T}_{\mathbf{c}}[\mathbf{x}]\|_{\sup} \leq \|\mathfrak{T}_{\mathbf{c}}[\mathbf{0}]\|_{\sup} + \|\mathfrak{T}_{\mathbf{c}}[\mathbf{x}] - \mathfrak{T}_{\mathbf{c}}[\mathbf{0}]\|_{\sup}\,.$$

For the first term we have

$$\|\mathfrak{T}_{\mathbf{c}}[\mathbf{0}]\|_{\sup} = \left\| \begin{bmatrix} e^{Bt}\mathbf{c} \\ \mathbf{0} \end{bmatrix} \right\|_{\sup} = \sup_{0 \leq t < \infty} |e^{Bt}\mathbf{c}| = |\mathbf{c}| < \frac{\delta}{2}.$$

For the second, we know that $\|\mathbf{x}\|_{\sup} \leq \delta$, so it follows from (6.105) with $\mathbf{y} = \mathbf{0}$ that $\|\mathfrak{T}_{\mathbf{c}}[\mathbf{x}] - \mathfrak{T}_{\mathbf{c}}[\mathbf{0}]\|_{\sup}$ is less than $\delta/2$. Thus $\|\mathfrak{T}_{\mathbf{c}}[\mathbf{x}]\|_{\sup} \leq \delta$, so $\mathfrak{T}_{\mathbf{c}}[\mathbf{x}] \in \mathsf{S}$, as claimed. □

To conclude: it follows from the fixed-point theorem that *for every* $\mathbf{c} \in \mathbb{R}^{d_s}$ *with* $|\mathbf{c}| < \delta/2$, *the operator* $\mathfrak{T}_{\mathbf{c}}$ *has a unique fixed point in* S, which we will denote by $\mathbf{X}_{\mathbf{c}}(t)$.

6.10.3 The Stable Manifold

Let the neighborhood $\mathcal{V}$ in Theorem 6.6.1 be given by

$$\mathcal{V} = \{(\mathbf{c}, \mathbf{d}) \in \mathbb{R}^{d_s} \times \mathbb{R}^{d-d_s} : |\mathbf{c}| < \delta/2,\ |\mathbf{d}| < \delta/2\}$$

with δ defined as above. If $\mathbf{x}(t)$ satisfies (6.98) and if $\mathbf{x}(t)$ never leaves $\mathcal{V}$, then $\mathbf{x}(t)$ is a bounded solution of the integral equation (6.99) and must in fact be the unique fixed point of $\mathfrak{T}_{\mathbf{c}}$, i.e., $\mathbf{x}(t) \equiv \mathbf{X}_{\mathbf{c}}(t)$. In particular, regarding initial conditions, $\mathbf{d} = \mathbf{X}_{\mathbf{c}}^{(nst)}(0)$. We use this fact to define the stable manifold as the graph of a mapping $\boldsymbol{\psi}$ from a ball in $\mathbb{R}^{d_s}$ into $\mathbb{R}^{d-d_s}$, i.e.,

$$\mathcal{M}_s = \{(\mathbf{c}, \mathbf{d}) \in \mathbb{R}^{d_s} \times \mathbb{R}^{d-d_s} : |\mathbf{c}| < \delta/2,\ \mathbf{d} = \boldsymbol{\psi}(\mathbf{c})\}, \tag{6.106}$$

where if $\mathbf{c} \in \mathbb{R}^{d_s}$ with $|\mathbf{c}| < \delta/2$, we define

$$\boldsymbol{\psi}(\mathbf{c}) = \mathbf{X}_{\mathbf{c}}^{(nst)}(0) = -\int_0^\infty e^{Cs}\,\mathbf{q}(\mathbf{X}_{\mathbf{c}}(s))\,ds. \tag{6.107}$$

Reflecting over the construction so far, we see that two claims in Theorem 6.6.1 remain to be proved:

- If $(\mathbf{c}, \mathbf{d}) \in \mathcal{M}_s$, the fixed point $\mathbf{X}_{\mathbf{c}}(t)$ has the decay indicated in (6.58).
- The set (6.106) is a differentiable manifold tangent to E_s.

For the proof of the first claim, a task similar to proving Theorem 6.1.1, we refer you to Section 5.4 of Meiss [54], especially Lemma 5.4. The following proposition establishes most of the second claim, and we also refer you to Meiss for the missing piece, i.e., that $\boldsymbol{\psi}$ is $\mathcal{C}^1$.

Proposition 6.10.4. *The map $\boldsymbol{\psi}$ is Lipschitz continuous, and $\boldsymbol{\psi}(\mathbf{c}) = o(|\mathbf{c}|)$ near the origin.*

Proof. If $\mathbf{c}_1, \mathbf{c}_2 \in B(\mathbf{0}, \delta/2) \subset \mathbb{R}^{d_s}$, then by the fixed-point property,

$$\mathbf{X}_{\mathbf{c}_1} - \mathbf{X}_{\mathbf{c}_2} = \boldsymbol{\mathfrak{T}}_{\mathbf{c}_1}[\mathbf{X}_{\mathbf{c}_1}] - \boldsymbol{\mathfrak{T}}_{\mathbf{c}_2}[\mathbf{X}_{\mathbf{c}_2}].$$

Adding and subtracting $\boldsymbol{\mathfrak{T}}_{\mathbf{c}_2}[\mathbf{X}_{\mathbf{c}_1}]$ and invoking the triangle inequality, we deduce

$$\|\mathbf{X}_{\mathbf{c}_1} - \mathbf{X}_{\mathbf{c}_2}\|_{\sup} \leq \|\boldsymbol{\mathfrak{T}}_{\mathbf{c}_1}[\mathbf{X}_{\mathbf{c}_1}] - \boldsymbol{\mathfrak{T}}_{\mathbf{c}_2}[\mathbf{X}_{\mathbf{c}_1}]\|_{\sup} + \|\boldsymbol{\mathfrak{T}}_{\mathbf{c}_2}[\mathbf{X}_{\mathbf{c}_1}] - \boldsymbol{\mathfrak{T}}_{\mathbf{c}_2}[\mathbf{X}_{\mathbf{c}_2}]\|_{\sup}. \quad (6.108)$$

But for the first term,

$$\|\boldsymbol{\mathfrak{T}}_{\mathbf{c}_1}[\mathbf{X}_{\mathbf{c}_1}] - \boldsymbol{\mathfrak{T}}_{\mathbf{c}_2}[\mathbf{X}_{\mathbf{c}_1}]\|_{\sup} = \left\| \begin{bmatrix} e^{Bt}(\mathbf{c}_1 - \mathbf{c}_2) \\ \mathbf{0} \end{bmatrix} \right\|_{\sup} = |\mathbf{c}_1 - \mathbf{c}_2|.$$

We estimate the second term by (6.105) and bring this term to the LHS of (6.108), thereby verifying Lipschitz continuity:

$$\|\mathbf{X}_{\mathbf{c}_1} - \mathbf{X}_{\mathbf{c}_2}\|_{\sup} \leq 2|\mathbf{c}_1 - \mathbf{c}_2|.$$

In particular, letting one of $\mathbf{c}$'s equal zero, we have

$$\|\mathbf{X}_{\mathbf{c}}\|_{\sup} \leq 2|\mathbf{c}|. \quad (6.109)$$

To show that $\boldsymbol{\psi}(\mathbf{c}) = o(|\mathbf{c}|)$, let $\alpha > 0$ be given. Choose a positive constant η_1 such that $\eta_1 < \varepsilon\alpha/2$. By Lemma 6.10.2, we may choose $\delta_1 > 0$ such that $|\mathbf{q}(\mathbf{x})| < \eta_1|\mathbf{x}|$ for $|\mathbf{x}| < \delta_1$. If $|\mathbf{c}| < \delta_1/2$, then by (6.109), $\|\mathbf{X}_{\mathbf{c}}\|_{\sup} < \delta_1$, so $|\mathbf{q}(\mathbf{X}_{\mathbf{c}}(s))| < \eta_1 \|\mathbf{X}_{\mathbf{c}}\|_{\sup}$. Thus,

$$|\boldsymbol{\psi}(\mathbf{c})| < \int_0^\infty e^{-\varepsilon s} \left(\eta_1 \|\mathbf{X}_{\mathbf{c}}\|_{\sup}\right) ds \leq \frac{2\eta_1}{\varepsilon}|\mathbf{c}| < \alpha|\mathbf{c}|,$$

as desired. □

6.10.4 Stable Manifolds at Nonhyperbolic Equilibria

Even if an equilibrium $\mathbf{b}_*$ fails to be hyperbolic, stable/unstable manifolds still exist. This behavior is illustrated by the example

$$\begin{aligned} x' &= x^2, \\ y' &= -y. \end{aligned} \quad (6.110)$$

Note that the Jacobian $\mathbf{DF}_*$ has the eigenvalues -1 and 0, with -1 having the eigenvector $\mathbf{e}_2$. The stable manifold $\mathcal{M}_s$ of the equilibrium is one-dimensional—the y-axis, tangent to $\mathbf{e}_2$—and solutions lying along $\mathcal{M}_s$ converge exponentially fast to the origin as $t \to \infty$. The behavior of other solutions is discussed in the next section.

We formulate a partial generalization of Theorem 6.6.1 to such cases. Let $\mathbf{b}_*$ be a nonhyperbolic equilibrium of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, and suppose E_s, the subspace spanned by eigenvectors whose eigenvalues have negative real parts, has dimension d_s.

Proposition 6.10.5. *Under the above circumstances, there is a differentiable manifold $\mathcal{M}_s$ of dimension d_s defined near $\mathbf{b}_*$, tangent to E_s at $\mathbf{b}_*$, such that if $\mathbf{b} \in \mathcal{M}_s$, then the IVP (6.57) has a solution $\boldsymbol{\varphi}(t, \mathbf{b})$ for all positive time and*

$$\lim_{t\to\infty} \boldsymbol{\varphi}(t, \mathbf{b}) = \mathbf{b}_*. \tag{6.111}$$

Regarding proofs, near $\mathbf{b}_*$, we may still split $\mathbf{F}(\mathbf{x})$ into stable and unstable components as in (6.96), where on E_s the inequality $\Re\lambda_k(B) < 0$ remains valid, but on E_u we have only $\Re\lambda_k(C) \le 0$. Thus, Proposition 6.10.1 may fail in the present context, because $\int_t^\infty e^{(s-t)C}\,\mathbf{q}(\mathbf{x}(s))\,ds$ need not converge. This issue is addressed as part of the theory of center manifolds, as for example in [12].

Changing directions, let us use the unstable-manifold analogue of Proposition 6.10.5 to prove[23] Proposition 6.2.1.

Proof of Proposition 6.2.1. Since $\mathbf{DF}_*$ has at least one unstable eigenvalue, the unstable manifold $\mathcal{M}_u$ through $\mathbf{b}_*$ is nontrivial; i.e., there is at least one point $\mathbf{b} \in \mathcal{M}_u$ different from $\mathbf{b}_*$. By Proposition 6.10.5,

$$\lim_{t\to-\infty} \boldsymbol{\varphi}(t, \mathbf{b}) = \mathbf{b}_*.$$

Choose a neighborhood $\mathcal{V}$ of $\mathbf{b}_*$ that does not contain $\mathbf{b}$.

Now, given any neighborhood $\mathcal{V}_1$ of $\mathbf{b}_*$, no matter how small, there is a $t_0 > 0$ such that $\boldsymbol{\varphi}(-t_0, \mathbf{b}) \in \mathcal{V}_1$. By translational invariance, the IVP

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0) = \boldsymbol{\varphi}(-t_0, \mathbf{b})$$

has solution $\mathbf{x}(t) = \boldsymbol{\varphi}(t - t_0, \mathbf{b})$, so $\mathbf{x}(t_0) = \mathbf{b} \notin \mathcal{V}$. That is, at least one solution with initial conditions in the arbitrarily small neighborhood $\mathcal{V}_1$ leaves the fixed neighborhood $\mathcal{V}$. □

6.11 Appendix 2: Center Manifolds and Nonhyperbolicity

As we just saw in Section 6.10.4, stable and unstable manifolds still exist at a nonhyperbolic equilibrium, but the sum of the dimensions of $\mathcal{M}_s$ and $\mathcal{M}_u$ is less than the overall dimension of the ODE. The *center manifold*, which we describe without

[23] If you object to using such heavy theory to prove such a basic result, we remind you that Exercise 18 outlines a direct proof of Proposition 6.2.1.

proofs[24] in this appendix, fills this gap in dimensions. The center manifold at a nonhyperbolic equilibrium $\mathbf{b}_*$ is tangent to the subspace spanned by all eigenvectors of $\mathbf{DF}_*$ that have zero real part. While the stable and unstable manifolds are robust objects, the center manifold is somewhat quirky. We illustrate the issues in two examples.

6.11.1 First Example

Let us explore more thoroughly the nonhyperbolic equilibrium of (6.110) at $\mathbf{b}_* = \mathbf{0}$. Since $\mathbf{DF}_*$ has no eigenvalues with positive real part, $\mathcal{M}_u$ is trivial; i.e., it contains only the origin. As we saw above, the stable manifold is the y-axis, $\mathcal{M}_s = \{\mathbf{b} \in \mathbb{R}^2 : b_1 = 0\}$. The x-axis is the obvious candidate for the center manifold, and it is indeed an invariant set. However, there are other candidates: either by solving the equations (6.110) and eliminating t or by solving the ODE

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = -\frac{y}{x^2},$$

you can show that for every constant C, the curve

$$y = Ce^{1/x}, \qquad -\infty < x < 0$$

is an orbit of (6.110) that converges to the origin as $t \to \infty$. Thus, for every C, the curve

$$\{\mathbf{b} : b_2 = Ce^{1/b_1},\ -\infty < b_1 < 0\} \bigcup \{\mathbf{b} : b_2 = 0,\ 0 \le b_1 < \infty\} \tag{6.112}$$

is another possible center manifold; the two halves of (6.112) join one another in a $\mathcal{C}^\infty$ manner. (*Be sure to draw some of the curves* (6.112)*!*) In this example it may seem artificial to take $C \neq 0$, but frequently in more complicated examples, such choices cannot be avoided. This infinite multiplicity demonstrates that center manifolds are not unique.

Regarding the asymptotic behavior of trajectories, in the right half-plane, $\mathcal{M}_c$ behaves like an unstable manifold: it contains a single orbit that tends to the equilibrium $\mathbf{0}$ as $t \to -\infty$ but leaves every bounded neighborhood of $\mathbf{0}$ as t increases. On the other hand, in the left half-plane, the curves (6.112) could be viewed as an infinite collection of stable manifolds: each of the orbits comes in from minus infinity and converges to $\mathbf{0}$ as $t \to \infty$. However, the convergence rate is only algebraic, i.e., like an inverse power of t, not exponential.

[24] See [12] for a careful treatment of the subject.

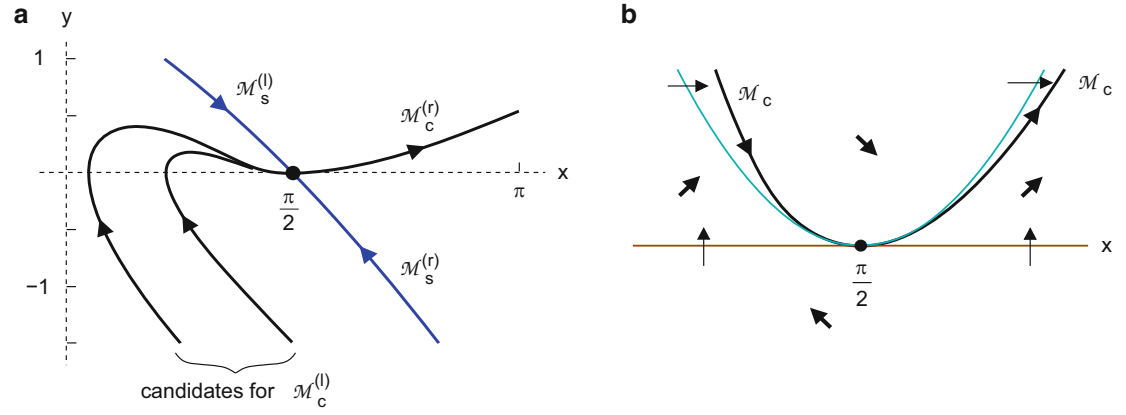

Figure 6.14: *(a) Stable and center manifolds for the critically torqued pendulum system (6.113) with $\beta = 1.3$. (b) Approximation for the local center manifold (black) superimposed on nullclines (brown and cyan), at a greatly expanded scale.*

6.11.2 Second Example

This example comes from the torqued pendulum (4.24) when the torque is set equal to its critical value $\mu = 1$:

$$\begin{aligned} x' &= y, \\ y' &= 1 - \sin x - \beta y, \end{aligned} \tag{6.113}$$

where β is a positive constant. This equation has a nonhyperbolic equilibrium at $\mathbf{b}_* = (\pi/2, 0)$, where its Jacobian is

$$\mathbf{DF}_* = \begin{bmatrix} 0 & 1 \\ 0 & -\beta \end{bmatrix}.$$

As before, there is no unstable manifold. The stable manifold is tangent to $(1, -\beta)$, the eigenvector of $\mathbf{DF}_*$ with eigenvalue $-\beta$. A simple computation locates $\mathcal{M}_s$ as shown in Figure 6.14(a).

In the previous example there was an obvious candidate for $\mathcal{M}_c$, but the situation is less clear in this one. However, there is a standard technique to locate possible center manifolds, which we now apply. Every center manifold must be tangent to the null eigenvector $(1, 0)$ at $\mathbf{b}_*$. Therefore, $\mathcal{M}_c$ may be represented (locally) as a graph

$$\mathcal{M}_c = \{\mathbf{b} \in \mathbb{R}^2 : b_2 = h(b_1),\ \frac{\pi}{2} - \varepsilon < b_1 < \frac{\pi}{2} + \varepsilon\},$$

where h has a Taylor expansion beginning with quadratic terms

$$h(x) = a(x - \pi/2)^2 + b(x - \pi/2)^3 + \dots\ . \tag{6.114}$$

Suppose a trajectory $(x(t), y(t))$ of (6.113) passes through a point $(x_0, h(x_0)) \in \mathcal{M}_c$ at time t_0. Then according to (6.113),

$$\begin{aligned} x'(t_0) &= h(x_0), \\ y'(t_0) &= 1 - \sin x_0 - \beta h(x_0). \end{aligned}$$

But since $\mathcal{M}_c$ is invariant, the velocity $(x'(t_0), y'(t_0))$ must be parallel to the tangent vector of $\mathcal{M}_c$, i.e., to $(1, dh/dx(x_0))$. Being parallel requires that

$$\frac{1 - \sin x - \beta h(x)}{h(x)} = \frac{dh/dx(x)}{1}, \tag{6.115}$$

where we have suppressed the subscript zero on x– this equation holds for all x near zero.

We interpret (6.115) as an ODE for $h(x)$. Clearing the denominator, substituting the series (6.114), and expanding in powers of x, we match coefficients to calculate from (6.115) that

$$a = 1/2\beta, \qquad b = -1/2\beta^3.$$

The graph of (6.114) with quartic and higher-order terms truncated is shown in Figure 6.14(b). Through quadratic terms, the graph follows the y-nullcline

$$\left\{ y = \frac{1 - \sin x}{\beta} \right\},$$

but because of the cubic term in (6.114), the graph veers away from this nullcline—below it in the first quadrant, above it in the second.

This procedure may be continued to calculate arbitrarily many terms in the series (6.114). Every candidate for $\mathcal{M}_c$, such as shown in Figure 6.14(a), matches this series to all orders, but $\mathcal{M}_c$ is nonunique in a fashion similar to the preceding example. Specifically, although there is a unique orbit of (6.113) that escapes from $\mathbf{b}_*$ to the right, there are infinitely many orbits[25] that converge to $\mathbf{b}_*$ from the left.

If the force $F(x) = 1 - \sin x$ in (6.113) is written as the negative gradient of a potential $V(x) = -x - \cos x$ (graphed in Figure 6.15), then the orbits in Figure 6.14(a) can be interpreted in terms of the rolling-marble analogy. For example, the right half of the center manifold, $\mathcal{M}_c^{(r)}$, describes a mass that falls off the equilibrium at time minus infinity. (Note that x *increases* along this orbit, i.e., the pendulum "falls" *upward* from its horizontal equilibrium at $x = \pi/2$, because the torque is stronger than gravity.)

[25]Incidentally, there are more candidates for $\mathcal{M}_c^{(l)}$ than suggested in the figure; see, for example, Figure 9.6(c).

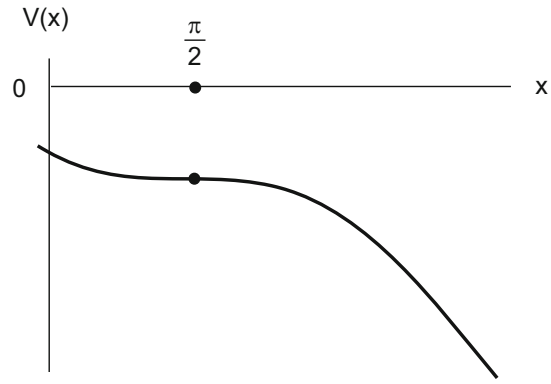

Figure 6.15: *Potential energy function* $V(x) = -x - \cos x$ *for the critically torqued pendulum* (6.113).

Exercise 21 provides a different perspective on the center manifold in this example.

Chapter 7

Oscillations in ODEs

As its title implies, this chapter is concerned with oscillatory solutions of ODEs. Solutions of the van der Pol system (1.36)

$$\begin{aligned} x' &= y, \\ y' &= -\beta(x^2-1)y - x, \end{aligned} \tag{7.1}$$

plotted in Figure 1.7, are representative of the kind of behavior we focus on. Up to now, we have been forced to rely on the computer to study such phenomena. In this chapter, we introduce analytical techniques to predict and describe oscillatory behavior.

This chapter is long, but a significant fraction of it may be skipped without loss of continuity. The essential theory, not to be skipped, is presented in Section 7.1, the first part of Section 7.2, and Section 7.3. The remaining sections, with explicit examples, greatly lengthen the chapter and are not easy reading. On the other hand, they make the topic more vivid by giving concrete descriptions of periodic solutions of nontrivial ODEs—good material, if you have the time for it. Moreover, the asymptotic techniques in Sections 7.5 and 7.6 are applicable in many other contexts.

We discuss the specific contents of this chapter more fully in Section 7.1.3. The interdependence of the various sections is shown in Figure 7.1.

7.1 Periodic Solutions

7.1.1 Basic Issues

A nonconstant solution of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, defined for $-\infty < t < \infty$, is called *periodic* if there exists a real number $T \neq 0$ such that $\mathbf{x}(t+T) = \mathbf{x}(t)$ for all $t \in \mathbb{R}$. Every nonzero T for which this equation holds is called a *period* of $\mathbf{x}$.

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8_7

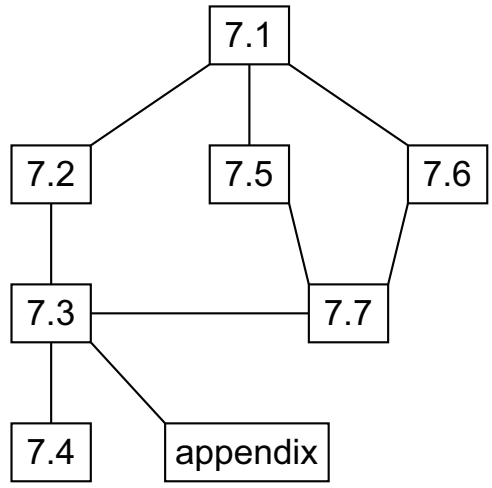

Figure 7.1: *Interdependence of the sections in this chapter.*

The following lemma[1] and its corollary will help in analyzing periodicity.

Lemma 7.1.1. *Let* $\mathbf{x}(t)$ *be a continuous nonconstant function on a closed interval* $[t_1, t_2]$ *that satisfies* $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ *for* $t_1 < t < t_2$ *and moreover*

$$\mathbf{x}(t_1) = \mathbf{x}(t_2), \tag{7.2}$$

this point belonging to the domain of $\mathbf{F}$*. Then (i) the maximal interval of existence for* $\mathbf{x}(t)$ *is* $-\infty < t < \infty$*, and (ii)* $\mathbf{x}(t)$ *is periodic with period* $T = t_2 - t_1$*.*

Proof. Let us define $\mathbf{x}(t)$ for all t by extending $\mathbf{x}$ periodically. Thus for all t,

$$\mathbf{x}(t + T) = \mathbf{x}(t).$$

By (7.2), this extension is unambiguously defined and continuous on $\mathbb{R}$. Since the equation $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ is autonomous, the extension satisfies the ODE on every translate of the original (open) interval, $(t_1 + nT, t_2 + nT)$, where n is an integer. By Lemma 3.2.9, the extension in fact satisfies the equation everywhere. By uniqueness, this periodic extension equals the maximal solution derived from the original solution. □

Corollary 7.1.2. *If the trajectory of a solution* $\mathbf{x}$ *of* $\mathbf{x}' = \mathbf{F}(\mathbf{x})$*, defined for all* t*, is contained in a closed* $\mathcal{C}^1$ *curve* Γ *and if there are no equilibria of* $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ *on* Γ*, then* $\mathbf{x}$ *is periodic.*

Proof. Since Γ is compact and has no equilibria, the minimum speed along Γ, i.e., $\min_\Gamma |\mathbf{F}(\mathbf{x})|$, is positive. Moreover, Γ has finite length. Thus, $\mathbf{x}$ will complete a

[1] In Problem 3.10 we asked you to prove this, but we still give the easy proof here.

circuit of Γ in some time less than the length of Γ divided by this minimum speed. This shows that there is a time T such that $\mathbf{x}(T) = \mathbf{x}(0)$, and we may therefore apply the preceding lemma. □

Here are a couple of simple properties of periodic solutions of an ODE that almost don't require proof.

Proposition 7.1.3. *If $\mathbf{x}(t)$ is a periodic solution of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, then (i) there are no equilibria on the orbit*[2] *of $\mathbf{x}$ and (ii) $\mathbf{x}$ has a minimal period.*

Proof. Regarding Claim (i), suppose that for some t_*, the point $\mathbf{x}(t_*)$ is an equilibrium; let $\mathbf{b}_* = \mathbf{x}(t_*)$. Then $\mathbf{y}(t) \equiv \mathbf{b}_*$ and $\mathbf{x}(t)$ are two different solutions of the IVP

$$\mathbf{y}' = \mathbf{F}(\mathbf{y}), \qquad \mathbf{y}(t_*) = \mathbf{b}_*,$$

contradicting uniqueness. Regarding Claim (ii), define the infimum of the set of periods,

$$p = \inf\{T : T > 0 \quad \text{and} \quad \mathbf{x}(T) = \mathbf{x}(0)\},$$

and consider a sequence T_n of periods that converges to p. If $p = 0$, then

$$\mathbf{x}'(0) = \lim_{n\to\infty} \frac{\mathbf{x}(T_n) - \mathbf{x}(0)}{T_n} = \mathbf{0},$$

which contradicts Claim (i). Thus $p > 0$, and moreover

$$\mathbf{x}(p) = \lim_{n\to\infty} \mathbf{x}(T_n) = \mathbf{x}(0),$$

so p is the minimal period. □

If T is the minimal period of $\mathbf{x}$, then the function $\mathbf{x} : [0, T] \to \mathbb{R}^d$ defines a closed curve that has no self-intersections. In complex analysis (where $d = 2$), such a curve is called a simple closed curve or a *Jordan* curve [11]. Jordan curves can be rather intricate; see, for example, Figure B.2 in Appendix B.

7.1.2 Examples of Periodic Solutions

(a) Examples with a continuum of periodic solutions

Example 1: *(Linear equations)* Recall the system

$$x' = y, \qquad y' = -x,$$

[2]Recall the distinction: *orbit* refers to the set $\{\mathbf{x}(t) : t \in \mathbb{R}\}$ considered as a subset of $\mathbb{R}^d$, independent of any parametrization; *trajectory* refers to the curve with the specific parametrization that satisfies the ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$.

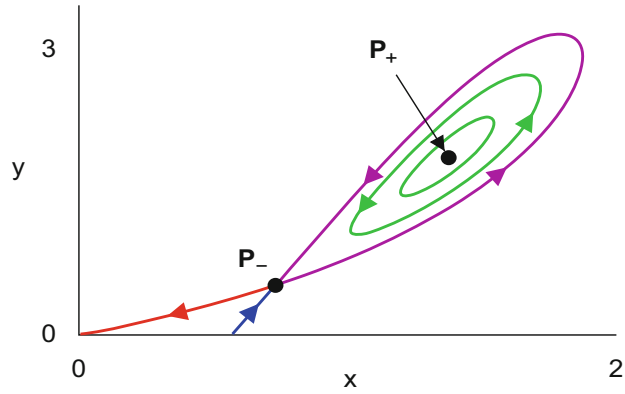

Figure 7.2: *Periodic orbits (green) for the activator–inhibitor system* (7.3) *with* $\rho = 1$ *and* $\sigma = 2.1$.

which comes from writing $x'' + x = 0$, the equation of motion for a simple harmonic oscillator, as a first-order system. Every nonconstant solution of this system is periodic with period 2π, the orbits being circles $\{x^2 + y^2 = C^2\}$. More generally, if A is a $d \times d$ matrix, then the constant-coefficient linear system $\mathbf{x}' = A\mathbf{x}$ has periodic solutions if and only if A has at least one complex-conjugate pair of nonzero pure imaginary eigenvalues. *(Why?)*

Example 2: *(Hamiltonian systems)* Two-dimensional Hamiltonian systems often have a continuum of periodic solutions. Bypassing various mechanical systems, we consider instead the activator–inhibitor system (6.22) with $\rho = 1$:

$$\begin{aligned} (a) \quad x' &= \sigma x^2/(1+y) - x, \\ (b) \quad y' &= x^2 - y. \end{aligned} \tag{7.3}$$

Figure 7.2 shows a one-parameter family of periodic orbits of this system in the case $\sigma = 2.1$. These periodic orbits arise from the fact that, as you showed in Exercise 6.10(f), this system is reparametrized Hamiltonian with

$$H(x, y) = \sigma \ln(1 + y) - \frac{y}{x} - x.$$

The Hamiltonian is constant along trajectories of (7.3), so orbits are contained in level sets of $H(x, y)$. For a range of C, the level set $\{H(x, y) = C\}$ contains a closed loop, and each such loop is a periodic orbit. The loops are bounded by a homoclinic orbit through the saddle point $\mathbf{P}_- \approx (0.730, 0.533)$, shown in purple in the figure. This orbit represents a limiting case of a periodic orbit, i.e., a trajectory that closes up on itself, but only in an infinite amount of time. All the periodic orbits encircle $\mathbf{P}_+ \approx (1.370, 1.877)$, the other nonzero equilibrium of (7.3).

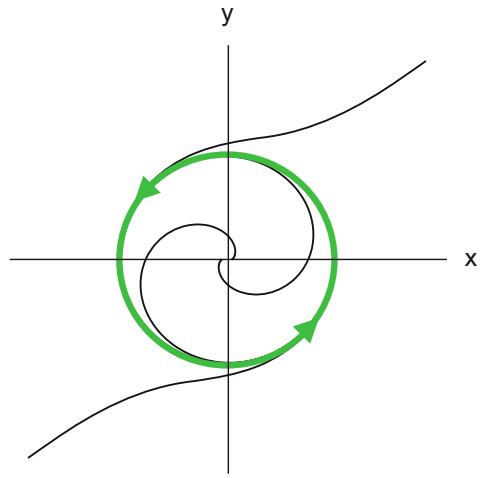

Figure 7.3: *Trajectories of the academic example,* (7.4).

(b) Examples with isolated periodic solutions

In both of the preceding examples, there are infinitely many periodic orbits. Of greater interest to us will be isolated periodic orbits known as *limit cycles*. (This name derives from the fact that for planar systems, nearby trajectories approach the periodic orbit in one of the limits $t \to \pm\infty$.) As our numerics have shown, van der Pol's system (7.1) has such a periodic orbit, and we shall derive this behavior analytically later in this chapter. In the meantime, here are two examples for which we can already show that such an orbit exists.

Example 3: *(An academic example)* Recall the system

$$\begin{aligned} x' &= x - y - (x^2 + y^2)x, \\ y' &= x + y - (x^2 + y^2)y, \end{aligned} \tag{7.4}$$

which we have considered several times above. The circle $\{r = 1\}$ is a limit cycle of this equation. Indeed, rewriting the system in polar coordinates[3]

$$r' = r(1 - r^2), \qquad \theta' = 1, \tag{7.5}$$

we found explicit solutions of this system. Even without the explicit solutions, one may see from (7.5) that the angular variable θ increases at a constant rate, and unless $r(0) = 0$, the radial variable r approaches 1 as $t \to \infty$. Thus, nearby trajectories are attracted to the periodic orbit $\{r = 1\}$ as $t \to \infty$ (see Figure 7.3).

[3]The simpler system $r' = 1 - r$, $\theta' = 1$, has more or less the same behavior. However, unlike (7.5), if we attempt to write this equation in Cartesian coordinates, $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, then $\mathbf{F}(\mathbf{x})$ is singular at the origin. It blows up like $1/r$ there. In other words, the simpler ODE is an equation only on $\mathbb{R}^2 \sim \{\mathbf{0}\}$. Often we shall accept such a singularity in order to simplify an example.

Example 4: *(Phase-locking on the torus)* Consider the system

$$\begin{aligned} \theta_1' &= \omega_1 + K_1 \sin(\theta_2 - \theta_1), \\ \theta_2' &= \omega_2 - K_2 \sin(\theta_2 - \theta_1). \end{aligned} \tag{7.6}$$

Here is a convenient interpretation of the equations (from Strogatz [81]) that makes them seem less abstract.[4] The variables θ_j may be viewed as the angular position of two runners on a circular track, and the equations describe the following assumptions: each runner has a natural "solo" speed ω_j, and these may be different, but they (the runners) place a value on running together, expressed by the terms $\pm K_j \sin(\theta_2 - \theta_1)$, which pull them closer if $|\theta_1 - \theta_2| < \pi$.

We claim that if

$$\left| \frac{\omega_1 - \omega_2}{K_1 + K_2} \right| < 1, \tag{7.7}$$

then there are "phase-locked" or "entrained" solutions of (7.6) of the form

$$\theta_1(t) = \overline{\omega}\, t, \qquad \theta_2(t) = \overline{\omega}\, t + \alpha, \tag{7.8}$$

where $\overline{\omega}, \alpha$ are constants. Note that $\theta_j'(t) = \overline{\omega}$. Thus, (7.6) is satisfied if and only if

$$\begin{aligned} \overline{\omega} &= \omega_1 + K_1 \sin\alpha, \\ \overline{\omega} &= \omega_2 - K_2 \sin\alpha. \end{aligned} \tag{7.9}$$

Subtracting these equations and rearranging, we obtain

$$\sin\alpha = \frac{\omega_2 - \omega_1}{K_1 + K_2}.$$

Condition (7.7) guarantees that this equation has real solutions. Ignoring possible translation by 2π, we pick two of these, say α_c and α_d, where

$$0 < |\alpha_c| < \pi/2, \qquad \pi/2 < |\alpha_d| < \pi. \tag{7.10}$$

(The subscript "c" is a mnemonic for "close," since in this case, the runners are closer to each other; similarly, "d" is a mnemonic for "distant.") Given the value of $\sin\alpha$, we substitute into (7.9) to obtain

$$\overline{\omega} = \frac{K_1\omega_2 + K_2\omega_1}{K_1 + K_2}.$$

This constructs two solutions $\boldsymbol{\gamma}_c(t)$ and $\boldsymbol{\gamma}_d(t)$ of the form (7.8), and these are shown in Figure 7.4.

[4] In Section 12.1 of [96], equations like (7.6) are proposed as a model for coupled oscillations of neurons.

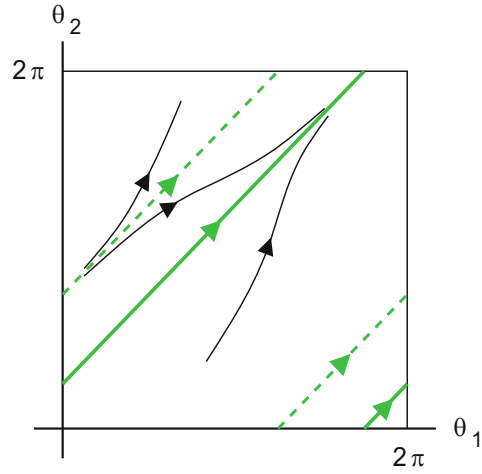

Figure 7.4: *Stable and unstable periodic orbits of the coupled oscillators* (7.6) *correspond to solid and dashed green trajectories. Here,* $K_1 = K_2 = 1/2$, *and* $\omega_2 - \omega_1 = \sqrt{2}/2$, *so that* $\alpha_c = \pi/4$ *and* $\alpha_d = 3\pi/4$.

Considered as ODEs on $\mathbb{R}^2$, (7.6) does not have periodic solutions—the functions $\theta_j(t)$ in (7.8) increase without bound. However, we regard θ_1, θ_2 as coordinates on the *torus* $\mathbb{T}^2 = (\mathbb{R}/2\pi\mathbb{Z}) \times (\mathbb{R}/2\pi\mathbb{Z})$, and in this interpretation,[5] the solutions (7.8) *are* periodic, with period $2\pi/\overline{\omega}$.

Equation (7.6) is the simplest representative of a class of ODEs that model entrainment in biological oscillations; cf. [83]. Incidentally, other entrained solutions of (7.6) are discussed in Section 10.4.

7.1.3 A Leisurely Overview of This Chapter

The issue of asymptotic behavior of solutions of an ODE as $t \to \pm\infty$ provides an instructive perspective on oscillatory solutions. The simplest asymptotic behavior as $t \to \infty$ of a solution that remains bounded is for it to converge to an equilibrium point. For example, in the previous chapter we saw that near an asymptotically stable equilibrium, every solution has this behavior. Limit cycles represent the next level of complexity in the asymptotic behavior of solutions.

Despite numerous analogies, limit cycles are more difficult to analyze than equilibria; even showing that they exist can be challenging. In Section 7.2, we introduce

[5] Already in Section 4.3.3 we found it convenient to reduce a Euclidean coordinate modulo 2π and thereby interpret a planar ODE as one on the cylinder. In a similar spirit, the periodic solution of (7.4) in polar coordinates, $r(t) \equiv 1$, $\theta(t) = t$, may be viewed as a closed curve on the cylinder $(0, \infty) \times S^1$. Incidentally, as we show in Section 7.9.4, every ODE on $\mathbb{T}^2$, in particular (7.6), can be embedded in an ODE on $\mathbb{R}^3$.

a general analytical tool for proving existence of periodic solutions, the Poincaré–Bendixson theorem, and apply it in two examples. Most emphatically, *this theorem is valid only in two dimensions.* The rest of Section 7.2 explores special properties of two-dimensional systems, including Dulac's criterion for nonexistence of periodic solutions.

As with equilibria, there is a notion of stability for limit cycles, driven by the following question: what happens if a trajectory starts from initial conditions that are "close" to a limit cycle? In Section 7.3, we define stability notions for limit cycles and introduce a general theoretical technique for analyzing their stability: *the Poincaré map.* Reminiscent of Theorem 6.1.1, the stability or instability of a limit cycle may be determined from the eigenvalues of a certain matrix derived from the Poincaré map. However, in contrast to Theorem 6.1.1, it is often difficult to calculate this matrix. Nevertheless, the conceptual framework provided by the Poincaré map is invaluable for understanding the behavior of many ODEs.

In Section 7.4 we study existence and stability of a limit cycle in another example, the torqued-pendulum equation.

In the rest of the chapter we turn to *describing* limit cycles as opposed to merely proving that they exist. There are three general techniques for this task:[6]

- numerical computation;
- asymptotic perturbation theory;
- rigorous mathematical analysis.

Virtually every problem is amenable to numerical solution; the limitation of this technique is that one may solve equations only with specific values of the parameters in it, which can make it difficult to get an overview of the behavior of solutions. Asymptotics, which works by deriving simpler, approximate problems that can be solved explicitly, is applicable only if there is a small or large parameter that can be exploited, and the calculations are frequently messy; on the other hand, it often provides an excellent overview of the behavior of solutions. Rigorous analysis is the least general of the three methods—new arguments must be developed for each new problem, and many problems are too complicated for complete analysis. However, you may find the attraction of rigor irresistible.

In this chapter, we *illustrate*[7] the use of the second of these techniques to approximate limit cycles. Specifically, in Section 7.5 we describe periodic solutions of

[6] Ask yourself which method you find most appealing. Your preference provides guidance about possible career choices, or at least specializations within mathematics. If you like numerics best, consider scientific computation; if you like asymptotics best, consider traditional applied mathematics; if you like rigorous methods best, consider mathematical analysis.

[7] Asymptotics is only a secondary focus for this book. We refer you to [98] or [43] to go beyond our limited coverage, or better still, take a course. Asymptotics is a hard subject to learn without

the van der Pol equation in the limit of small β, and in Section 7.6 we do likewise in the opposite limit of large β. Moreover, in Section 7.7 we apply the Poincaré map to prove stability for the solutions of the van der Pol equation constructed in the two preceding sections.

In an appendix we use ideas from *Floquet theory* to fulfill a promise we made in Chapter 1, to explain how vibration can stabilize an inverted pendulum (cf. Exercise 1.12). We include this material in the present chapter because Floquet theory and Poincaré maps are closely related.

7.2 Special Behavior in Two Dimensions . . . Mostly

The topology of the plane, in particular the Jordan curve theorem, greatly constrains the possible dynamical behavior of two-dimensional systems of ODEs. This is captured most fully by the strong version of the Poincaré–Bendixson theorem in Section 7.2.4.

The irreducible core of the two-dimensional theory is a simplified version of the theorem stated in Section 7.2.1 and its application in Section 7.2.2. Section 7.2.3 introduces limit sets, which are meaningful in every dimension; at least skim this material, because limit sets facilitate the discussion in later chapters. The more detailed information on two-dimensional ODEs in Sections 7.2.4–7.2.6 could be omitted with little loss of continuity.

7.2.1 The Poincaré–Bendixson Theorem: Minimal Version

Theorem 7.2.1. *Let* $\mathbf{F} : \mathcal{U} \to \mathbb{R}^2$ *be* $\mathcal{C}^1$ *on the open set* $\mathcal{U} \subset \mathbb{R}^2$*, and suppose that* $\mathcal{K} \subset \mathcal{U}$ *is a compact trapping region for* $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ *that does not contain any equilibria. Then* $\mathcal{K}$ *contains at least one periodic orbit of the ODE.*

At the risk of boring you, we repeat: *This theorem is valid only for planar systems.* No analogous result holds in higher dimensions. The theorem follows from the stronger version in Section 7.2.4.

Note that the trapping region in the theorem might contain several periodic orbits. For a rather contrived example, consider the system on $\mathbb{R}^2 \sim \{\mathbf{0}\}$ in polar coordinates

$$r' \;=\; (1-r)(2-r)(3-r), \qquad \theta' \;=\; 1.$$

The annulus $\{1/2 \le r \le 4\}$ is a trapping region with no equilibria, but it contains three periodic orbits: circles of radii 1, 2, and 3.

---

guidance from a pro.

If you wish to pursue one of the other two techniques, for numerical methods read Section 10.3 of this book and follow the references there, and for rigorous analysis see Section 6.7 of [54].

7.2.2 Applications of the Theorem

(a) The van der Pol equation

Recall that in Section 4.4 we constructed trapping regions for the van der Pol equation (7.1) in order to prove global existence. Let $\mathcal{K}_0$ be one such trapping region, corresponding to a choice of the parameter A in (4.34). (The outer curve in Figure 7.5 bounds such a region.) We cannot apply Theorem 7.2.1 using $\mathcal{K}_0$, because the equilibrium $(x, y) = (0, 0)$ lies inside $\mathcal{K}_0$. However, let us remove a disk of radius ε around the origin from $\mathcal{K}_0$; thus we define

$$\mathcal{K} = \mathcal{K}_0 \sim B(\varepsilon), \tag{7.11}$$

where $\varepsilon < 1$. Now $\partial\mathcal{K} = \partial\mathcal{K}_0 \cup \Gamma$, where Γ is the circle of radius ε. Regarding behavior at $\partial\mathcal{K}$, we know from Section 4.4 that the flow of (7.1) is inward along $\partial\mathcal{K}_0$, so we need consider only Γ. At a point $(\varepsilon\cos\theta, \varepsilon\sin\theta)$ on Γ the inward normal, i.e., pointing into $\mathcal{K}$, is $\mathbf{N}_\theta = (\cos\theta, \sin\theta)$, and we calculate that

$$\langle \mathbf{F}, \mathbf{N}_\theta \rangle = \beta\left(1 - \varepsilon^2\cos^2\theta\right)\varepsilon\sin^2\theta \geq 0.$$

Therefore, $\mathcal{K}$ is a trapping region for (7.1) that contains no equilibria. Hence by Theorem 7.2.1, there must be a periodic orbit of (7.1) inside $\mathcal{K}$.

Although we cannot conclude so from the theorem, there is in fact a *unique* periodic orbit of the van der Pol equation inside $\mathcal{K}$. We have observed this fact in computations; we shall derive it with asymptotics in the limit of small or large β (Sections 7.5 and 7.6, respectively); and we refer you to Meiss [54], p. 224, for an analytical proof for all β.

(b) The torqued pendulum equation

Recall from Section 4.3.3 that the torqued pendulum is described by the system

$$\begin{aligned} x' &= y, \\ y' &= -\sin x - \beta y + \mu. \end{aligned} \tag{7.12}$$

Suppose $\mu > 1$; i.e., suppose the torque is large enough to overcome the pull of gravity, no matter what the angle x of the pendulum may be. On physical grounds it seems intuitively obvious that under this hypothesis, there is a solution $x_*(t), y_*(t)$ of (7.12) such that the pendulum continues to rotate indefinitely in a periodic fashion. Such a solution would satisfy

$$x_*(t+T) = x_*(t) + 2\pi, \qquad y_*(t+T) = y_*(t) \tag{7.13}$$

for an appropriate $T > 0$, so it would be periodic if (7.12) is interpreted as an equation on $S^1 \times \mathbb{R}$.

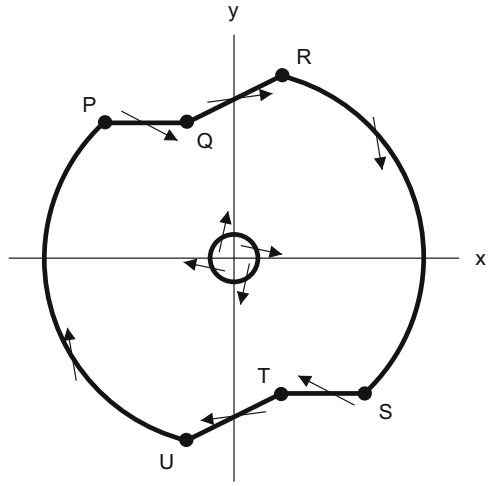

Figure 7.5: *Annular trapping region $\mathcal{K}$ of the form (7.11) used to prove existence of a periodic orbit of the van der Pol equation (7.1). The outer boundary of the region defines a trapping region for the van der Pol equation, as in Section 4.4. The inner boundary quarantines the (unstable) equilibrium $(0,0)$, so that no equilibria are contained in $\mathcal{K}$.*

We shall derive the existence of such a periodic solution with the following generalization of Theorem 7.2.1 to the cylinder. (See Exercise 1(b) for hints on how to prove this result.)

Theorem 7.2.2. *Let $\mathbf{F} : \mathcal{U} \to \mathbb{R}^2$ be $\mathcal{C}^1$ on the open set $\mathcal{U} \subset S^1 \times \mathbb{R}$, and suppose that $\mathcal{K} \subset \mathcal{U}$ is a compact trapping region for $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ that does not contain any equilibria. Then $\mathcal{K}$ contains at least one periodic orbit of the ODE.*

In Chapter 4 we showed that

$$\mathcal{K} = \{(x, y) \in S^1 \times \mathbb{R} : \ y^2/2 - \cos x \leq E_0\} \tag{7.14}$$

is a trapping region for (7.12) when $E_0 \geq (\mu/\beta)^2/2 + 1$. Since $\mu > 1$, this ODE has no equilibria. Thus, it follows from the Poincaré–Bendixson theorem that (7.12) has a periodic solution inside $\mathcal{K}$. (As we will see in Section 7.4, this solution is unique.)

Remark: In the above analysis we assumed that $\mu > 1$ in (7.12). Depending on the value of β, there may still be a periodic orbit even if $\mu < 1$; see Exercise 4.

(c) An auxiliary result, relevant to applying Theorem 7.2.1.

To construct a trapping region for the van der Pol equation with no equilibria, we had to delete a neighborhood of the origin. The following theorem warns you that *for every planar ODE*, only such an annular region can be equilibrium-free.

Theorem 7.2.3. *Let* $\mathbf{F} : \mathcal{U} \to \mathbb{R}^2$ *be* $\mathcal{C}^1$ *on the open set* $\mathcal{U} \subset \mathbb{R}^2$*, and suppose* $\mathcal{U}$ *is simply connected. If* Γ *is a periodic orbit of* $\mathbf{x}' = \mathbf{F}(\mathbf{x})$*, then this equation has at least one equilibrium inside* Γ*.*

This result may be elegantly derived as is a consequence of index theory, which uses algebraic topology to study periodic orbits; see Section 6.5 of [54]. Alternatively, see Theorem 2 in Section 11.5 of [40] for a direct proof of the theorem.

7.2.3 Limit Sets (in Any Dimension)

The following generally useful concept is needed to formulate the strong version of the Poincaré–Bendixson theorem. Unlike the rest of Section 7.2, *here, this idea makes sense in arbitrary dimension.*

Recall from Section 4.5.2 the flow notation $\boldsymbol{\varphi}(t, \mathbf{b})$ for the solution of an IVP

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0) = \mathbf{b}. \tag{7.15}$$

A point $\mathbf{z}$ is called an *ω-limit point* of $\mathbf{b}$ if $\boldsymbol{\varphi}(t, \mathbf{b})$ is defined for all $t \geq 0$ and there exists a sequence $\{t_n\}$ of real numbers tending to infinity such that

$$\lim_{n\to\infty} \boldsymbol{\varphi}(t_n, \mathbf{b}) \;=\; \mathbf{z}.$$

The set of all ω-limit points of $\mathbf{b}$ will be denoted by $\omega(\mathbf{b})$. The *α-limit set*, consisting of points obtained in the limit as $t \to -\infty$, is defined analogously, but we will not make much use of it.

An ω-limit set certainly can be empty, as illustrated by the scalar ODE $x' = x$ with $x(0) = b \neq 0$. (One could generalize the notion of limits to include the point at infinity, but we don't do this.) If the (forward) orbit through $\mathbf{b}$ is contained in a compact set, then $\omega(\mathbf{b})$ is nonempty. Here are some examples of ω-limit sets.

Example 1: *(A single point)* If $\mathbf{b}_*$ is an asymptotically stable equilibrium of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, then there exists a neighborhood $\mathcal{V}$ of $\mathbf{b}_*$ such that $\omega(\mathbf{b}) = \mathbf{b}_*$ for all $\mathbf{b} \in \mathcal{V}$. Similarly, if $\mathbf{b}_*$ is a saddle point and if $\mathbf{b} \in \mathcal{M}_s$, the stable manifold of the saddle, then $\omega(\mathbf{b}) = \mathbf{b}_*$. Equation (6.13), say with the radial equation simplified,

$$r' = 1 - r, \qquad \theta' = 1 - \cos\theta, \tag{7.16}$$

provides another example, $\omega(\mathbf{b}) = \{(1, 0)\}$ for all $\mathbf{b} \in \mathbb{R}^2 \sim \{\mathbf{0}\}$, even though the equilibrium[8] is not stable.

[8] Are we using Cartesian coordinates or polar coordinates to specify the equilibrium? It doesn't matter, because of a convenient mathematical pun: $(1, 0)$ specifies the same point in either coordinates.

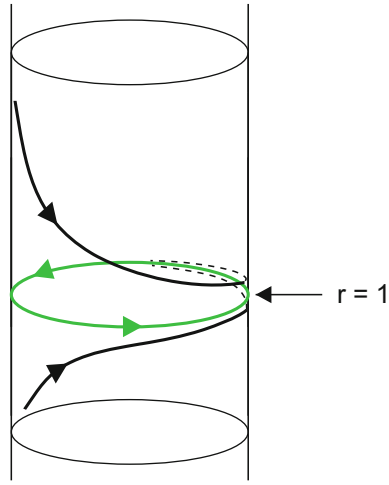

Figure 7.6: *Solutions of* (7.17) *rendered on the cylinder* $(0,\infty)\times S^1$. *The two black trajectories approach the (stable) periodic orbit, shown in green, as* $t\to\infty$. *The periodic orbit is the* ω*-limit of every* $\mathbf{b}\in(0,\infty)\times S^1$.

Example 2: *(A limit cycle)* Consider Example 3 in Section 7.1 (with the radial equation simplified):

$$r' = 1-r, \qquad \theta' = 1. \tag{7.17}$$

The unit circle Γ is a limit-cycle orbit of (7.17), and every solution of (7.17) in $\mathbb{R}^2\sim\{\mathbf{0}\}$ approaches Γ. Thus, in the present terminology, $\omega(\mathbf{b})=\Gamma$ for all $\mathbf{b}\neq 0$. The van der Pol equation exhibits similar behavior.

Remark: Sometimes it will be convenient to abandon the interpretation of an equation such as (7.17) as an ODE on the punctured plane $\mathbb{R}^2\sim\{\mathbf{0}\}$ and regard it as an ODE on the cylinder $(0,\infty)\times S^1$. Figure 7.6 shows a couple of trajectories in the latter interpretation.

Example 3: *(A homoclinic cycle)* New behavior emerges for another tweak of these equations:

$$r' = (1-r)^3, \qquad \theta' = 1-\cos\theta+(r-1)^2. \tag{7.18}$$

In Figure 7.7(a) we indicate the flow of (7.18), interpreted as an ODE on $\mathbb{R}^2\sim\{\mathbf{0}\}$. If $\mathbf{b}\notin\Gamma$, then $\omega(\mathbf{b})=\Gamma$, the unit circle. This behavior may be derived analytically from explicit solutions: if $t\to\infty$, then $r-1$ tends to zero as $t^{-1/2}$, so $\theta'\geq C/t$, which means that θ increases without bound. However, points on Γ behave differently from Example 2: if $\mathbf{b}\in\Gamma$, then $\omega(\mathbf{b})=\{(1,0)\}$. We call Γ a *homoclinic cycle*:

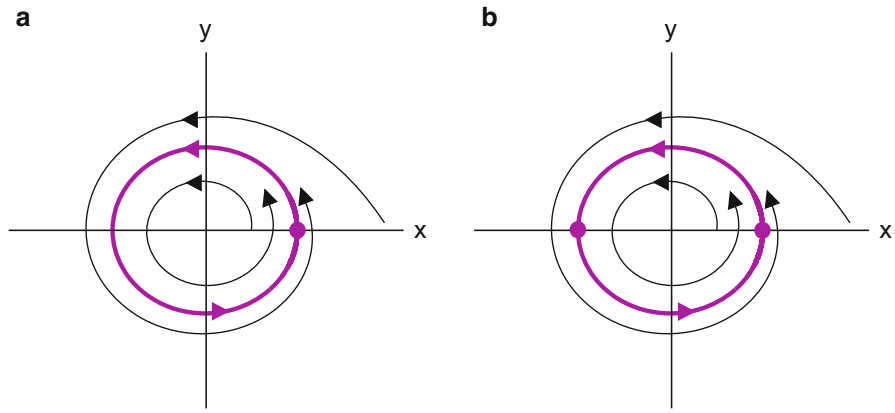

Figure 7.7: *(a) Trajectories for (7.18) and (b) trajectories for (7.19) with $n = 2$. Equilibria are indicated by bold dots. In Panel (a), the circle is a homoclinic cycle, the homoclinic orbit $\{(r,\theta) : r = 1, \quad 0 < \theta < 2\pi\}$ plus the equilibrium $(r,\theta) = (1,0)$. In Panel (b), the circle is a heteroclinic cycle containing two distinct orbits, $\{(r,\theta) : r = 1, \quad 0 < \theta < \pi\}$ and $\{(r,\theta) : r = 1, \quad \pi < \theta < 2\pi\}$, and two equilibria.*

this term describes an equilibrium and a homoclinic orbit connected to it as $t \to \pm\infty$. (This example may seem a little artificial,[9] but in Chapter 9, homoclinic cycles that are ω-limits arise naturally.)

Suppose we further modify the θ-equation in these examples to read

$$r' = (1-r)^3, \qquad \theta' = 1 - \cos(n\theta) + (r-1)^2, \tag{7.19}$$

for some positive integer n. (Figure 7.7(b) shows the flow if $n = 2$.) It is still true that $\omega(\mathbf{b}) = \Gamma$ if $\mathbf{b} \neq 0$ and $\mathbf{b} \notin \Gamma$, but now Γ consists of n equilibria (i.e., $r = 1$, $\theta = 2k\pi/n$, $k = 0, 1, \ldots, n-1$) and *heteroclinic* orbits (i.e., different limits as $t \to \pm\infty$) connecting these equilibria. With an obvious extension of the terminology, we shall call Γ a *heteroclinic cycle*. You should identify limit sets for points on Γ. (Exercise 22 gives more examples of this type.)

Incidentally, the homoclinic orbit in Figure 7.2 has different behavior: this homoclinic orbit is not part of the ω-limit set of any point $\mathbf{b} \in \mathbb{R}^2$. If $\mathbf{b}$ lies inside the homoclinic orbit, then $\omega(\mathbf{b})$ is the periodic orbit on which $\mathbf{b}$ lies; if $\mathbf{b}$ lies on the stable manifold of the saddle point, which includes the homoclinic orbit, then $\omega(\mathbf{b})$ is the saddle point $\mathbf{P}_-$; otherwise, $\omega(\mathbf{b})$ is the origin.

Example 4: *(Tori in higher dimensions)* The ω-limit set of a bounded trajectory for a higher-dimensional system often involves a torus. For example, consider the

[9]To make this more specific, the equilibrium $(1,0)$ of (7.18) is highly degenerate; indeed, all entries of the Jacobian $\mathbf{DF}$ vanish there.

three-dimensional flow on the domain $\{(r, \theta, z) : r > 0\}$ (expressed in cylindrical coordinates)

$$\begin{aligned} \theta' &= \omega_1, \\ \begin{bmatrix} r' \\ z' \end{bmatrix} &= \begin{bmatrix} 1 - 2R^2 - \omega_2, & \\ \omega_2 & 1 - 2R^2 \end{bmatrix} \begin{bmatrix} r - 1 \\ z \end{bmatrix}, \end{aligned} \tag{7.20}$$

where $R(r, z) = \sqrt{(r-1)^2 + z^2}$. Note that R is the radial coordinate in the r, z-plane if the point $(r, z) = (1, 0)$ is taken as the origin, say

$$r - 1 = R\cos\phi, \qquad z = R\sin\phi.$$

With respect to these coordinates, the r, z-subsystem may be rewritten as

$$R' = R - 2R^3, \qquad \phi' = \omega_2. \tag{7.21}$$

This system mimics the academic equation (7.5), with the complication that solutions of (7.21) need not be globally defined, because they may get driven to the set $\{R\cos\phi < -1\}$, for which the coordinate transformation becomes ill defined. However, solutions of (7.20) that start close to the level set $\{R = 1/\sqrt{2}\}$, which we denote by $\mathbb{T}$ because it is a torus in the original Euclidean space, are globally defined and in fact tend to $\mathbb{T}$ as $t \to \infty$. As will follow from Exercise 7, if ω_1/ω_2 is irrational, the ω-limit set of every such initial condition is the entire torus $\mathbb{T}$.

Example 5: *(Complicated limit sets in higher dimensions)* In three or more dimensions, limit sets can be far more complicated than any of the above examples. In fact, in the 1960s, researchers were so perplexed by limit sets they observed that they coined the pejorative phrase "strange attractor" [69] to describe such sets. If your curiosity is piqued, you may peek ahead to see examples of such behavior in Sections 9.6 and 9.7, particularly Figures 9.16(b) and 9.17(a). See [30] for a broader view on these developments.

Although we don't use limit sets all that much, let us nevertheless derive a couple of the most basic properties of such sets. Let $\boldsymbol{\varphi}(t, \mathbf{b})$ be the flow associated with an ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, where $\mathbf{F} : \mathcal{U} \to \mathbb{R}^d$. Recalling the definition from Chapter 6, we say that $\mathcal{V} \subset \mathcal{U}$ is *invariant* with respect to the flow if $\boldsymbol{\varphi}(t, \mathbf{b})$ is defined and belongs to $\mathcal{V}$ for all $\mathbf{b} \in \mathcal{V}$ and all $t \in \mathbb{R}$.

Proposition 7.2.4. *Every ω-limit set $\omega(\mathbf{b})$ is a closed invariant subset of $\mathcal{U}$.*

Proof. If $\omega(\mathbf{b})$ is empty, the assertion is trivial. If $\omega(\mathbf{b})$ is nonempty, let $\mathbf{x}(t) = \boldsymbol{\varphi}(t, \mathbf{b})$. Regarding being closed, suppose that $\{\mathbf{z}_m\}$ is a sequence of points in $\omega(\mathbf{b})$ that converges to $\mathbf{z}$. Then we must show that $\mathbf{z} \in \omega(\mathbf{b})$; i.e., there exists a sequence $\{t_n\}$ tending to infinity such that $\lim_n \mathbf{x}(t_n) = \mathbf{z}$. Since $\mathbf{z}_m \in \omega(\mathbf{b})$, there exist

sequences $\{s_k^{(m)}\}$, all tending to infinity as $k \to \infty$, such that $\lim_k \mathbf{x}(s_k^{(m)}) = \mathbf{z}_m$. For $n = 1, 2, \ldots$, choose $t_n = s_{k(n)}^{(n)}$, with $t_n \geq n$, such that

$$|\mathbf{x}(t_n) - \mathbf{z}_n| < \frac{1}{n}.$$

Then

$$|\mathbf{x}(t_n) - \mathbf{z}| \leq |\mathbf{x}(t_n) - \mathbf{z}_n| \ + \ |\mathbf{z}_n - \mathbf{z}|,$$

and since both terms on the right tend to zero, we see that $\omega(\mathbf{b})$ is closed.

Regarding invariance, suppose $\mathbf{z}_0 \in \omega(\mathbf{b})$; thus there is a sequence t_n tending to infinity such that $\mathbf{z}_0 = \lim_n \boldsymbol{\varphi}(t_n, \mathbf{b})$. Let $\mathbf{z} = \boldsymbol{\varphi}(t_*, \mathbf{z}_0)$ be a point on the trajectory through $\mathbf{z}_0$. If $t_* \geq 0$, consider the sequence $\boldsymbol{\varphi}(t_n + t_*, \mathbf{b})$. By the semigroup property, Proposition 4.5.3,

$$\boldsymbol{\varphi}(t_n + t_*, \mathbf{b}) \ = \ \boldsymbol{\varphi}(t_*, \boldsymbol{\varphi}(t_n, \mathbf{b})),$$

and by continuity,

$$\lim_n \boldsymbol{\varphi}(t_*, \boldsymbol{\varphi}(t_n, \mathbf{b})) = \boldsymbol{\varphi}(t_*, \lim_n \boldsymbol{\varphi}(t_n, \mathbf{b})) = \boldsymbol{\varphi}(t_*, \mathbf{z}_0) = \mathbf{z}.$$

Hence $\mathbf{z} \in \omega(\mathbf{b})$, as claimed.

If $t_* < 0$, then early elements in the sequence $\boldsymbol{\varphi}(t_n + t_*, \mathbf{b})$ might be undefined, because $t_n + t_* < 0$. However, discarding these early problematic elements, we may proceed as above with a subsequence, say $\boldsymbol{\varphi}(t_n + t_*, \mathbf{b})$, where $n \geq N$. □

Limit sets need not be compact; see Exercise 20 for a counterexample. If a limit set is compact, it is connected (see, for example, Lemma 4.16 in Section 4.9 of [54]).

7.2.4 The Poincaré–Bendixson Theorem: Strong Version

In simple language, the strong version of the Poincaré–Bendixson theorem states that most limit sets in two-dimensional systems are no more complicated than Examples 1–3 in the previous subsection.

Theorem 7.2.5. *(Poincaré–Bendixson): Suppose that* $\mathbf{F} : \mathcal{U} \to \mathbb{R}^2$ *is* $\mathcal{C}^1$ *on* $\mathcal{U}$*, where* $\mathcal{U}$ *contains only finitely many equilibria. If the (forward) orbit through* $\mathbf{b}$ *lies in a compact subset of* $\mathcal{U}$*, then* $\omega(\mathbf{b})$ *satisfies one of the following conditions:*

(i) it consists of a single point;

(ii) it is a periodic orbit;

(iii) it is a homoclinic or heteroclinic cycle.

Because the techniques needed to prove this result have limited utility in the rest of the theory of ODEs, we refer you to other sources for a proof, for example Chapter 9 of [95] or Section 16.3 of [15].

Incidentally, the hypothesis that $\mathbf{F}$ has only finitely many equilibria in $\mathcal{U}$ is essential. For example, in the system

$$\begin{aligned} r' &= (1-r)^3, \\ \theta' &= (r-1)^2, \end{aligned} \tag{7.22}$$

the ω-limit set of a trajectory with $r(0) \neq 1$ is the unit circle Γ, and Γ is an infinite union of equilibria. *(Show this by solving the equations!* It's more delicate than you might expect.)

7.2.5 Nonexistence: Dulac's Theorem

The next proposition gives a sufficient condition to exclude periodic solutions. Following its proof, the result is applied in Section 7.2.6.

Proposition 7.2.6. *(Dulac). Suppose that $\mathbf{F} : \mathcal{U} \to \mathbb{R}^2$ is $\mathcal{C}^1$ on the open simply connected set $\mathcal{U} \subset \mathbb{R}^2$. If there exists a $\mathcal{C}^1$ function $g : \mathcal{U} \to \mathbb{R}$ such that the divergence $\nabla \cdot (g\mathbf{F})$ is nonnegative and is not identically zero on any open subset of $\mathcal{U}$, then the equation $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ has no periodic solutions lying entirely within $\mathcal{U}$.*

Remarks: (i) The same conclusion follows if $\nabla \cdot (g\mathbf{F})$ is *nonpositive* and is not identically zero on any open subset of $\mathcal{U}$. (ii) The proof below does not provide much intuition about *why* the proposition is true. As we explore in Section 7.9.1 of the Pearls, such intuition can be derived from considering how the areas of regions evolve under the flow.

Proof of Proposition 7.2.6. Suppose to the contrary that there exists a simple closed orbit Γ, and let Ω denote the interior of Γ. By Green's theorem,

$$\iint_{\Omega} \nabla \cdot (g\mathbf{F})\, dA \;=\; \oint_{\Gamma} (g\mathbf{F}) \cdot \mathbf{N}\, ds,$$

where $\mathbf{N}$, ds, and dA have their usual meanings. By our assumptions regarding $\nabla \cdot (g\mathbf{F})$, the double integral on the LHS is strictly positive. On the other hand, the contour integral on the RHS is zero, since the velocity vector $\mathbf{F}(\mathbf{x}) = \mathbf{x}'$ is tangent to Γ and therefore orthogonal to the normal $\mathbf{N}$. □

7.2.6 Section 1.6 Revisited, Part III

As an application of Proposition 7.2.6, consider the modified Lotka–Volterra model

$$x' = x\left(\frac{x-\varepsilon}{x+\varepsilon}\right) - xy, \qquad y' = \rho(xy - y), \tag{7.23}$$

where $\rho > 0$ and $\varepsilon \geq 0$. (The carrying capacity K is equal to infinity.) If $\varepsilon = 0$, all nonconstant trajectories of (7.23) in the (open) first quadrant are periodic. The seemingly innocent factor $(x - \varepsilon)/(x + \varepsilon)$, which is approximately equal to 1 if x is large, changes the dynamics completely. We claim that (7.23) has no periodic solutions in the biologically meaningful regime $\{x > 0, y > 0\}$. (This behavior is easily verified in computations, but it is a pleasure to be able to derive it analytically.) To prove the claim, we apply Dulac's theorem with $g(x, y) = 1/xy$. (In Section 7.9.1 we explain why this choice is natural.) Since

$$\nabla \cdot (g\mathbf{F}) \;=\; \frac{2\varepsilon}{y(x+\varepsilon)^2}$$

is strictly positive throughout the first quadrant, the claim follows.

In fact, the above information may be combined with the strong Poincaré–Bendixson theorem to show that *virtually all solutions of* (7.23) *converge to total extinction!* (See Exercise 8.)

7.3 Stability of Periodic Orbits and the Poincaré Map

7.3.1 An Eigenvalue Test for Stability

The notion of stability for periodic solutions is completely analogous to stability of equilibria. (For definiteness, you may think of an ODE on $\mathbb{R}^d$, but the same ideas apply for ODEs on the cylinder or the torus.) We say that a periodic solution of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ with orbit Γ is *Lyapunov stable* if for every neighborhood $\mathcal{V}$ of Γ, there is a smaller neighborhood $\mathcal{V}_1$ such that if initial data $\mathbf{b}$ are restricted to belong to $\mathcal{V}_1$, then the IVP is solvable for all positive times, and moreover, $\boldsymbol{\varphi}(t, \mathbf{b}) \in \mathcal{V}$ for all $t \geq 0$. Similarly, we say that a limit cycle is *asymptotically stable* if it is Lyapunov stable and there exists one neighborhood $\mathcal{V}_*$ of Γ such that for all $\mathbf{b} \in \mathcal{V}_*$,

$$\lim_{t\to\infty} \operatorname{dist}(\boldsymbol{\varphi}(t, \mathbf{b}), \Gamma) = 0.$$

For completeness, let us record that the distance from a point $\mathbf{x}$ to a compact set $\mathcal{K}$ is defined by $\operatorname{dist}(\mathbf{x}, \mathcal{K}) = \min_{\mathbf{y}\in\mathcal{K}} |\mathbf{x} - \mathbf{y}|$. For instance, it is easily seen that the limit cycle in Example 3 of Section 7.1 is asymptotically stable.

Let $\boldsymbol{\gamma}(t)$ be a periodic solution[10] of a d-dimensional system $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, say with period T. The theorem below provides a sufficient condition for asymptotic stability, based on the eigenvalues of $\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0))$, the differential of the flow map with respect to its second argument. The following lemma is invoked in enumerating these eigenvalues.

[10]We write lowercase $\boldsymbol{\gamma}(t)$ for the solution and uppercase Γ for the orbit.

Lemma 7.3.1. *One of the eigenvalues of* $\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0))$ *equals unity, with eigenvector* $\boldsymbol{\gamma}'(0)$.

Proof. By periodicity, we have that

$$\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(t)) = \boldsymbol{\gamma}(t).$$

The lemma follows from differentiating this equation with respect to t using the chain rule and evaluating at $t = 0$. □

Here and below we assume that $\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0))$ has eigenvalues $\lambda_1, \ldots, \lambda_{d-1}$, and $\lambda_d = 1$, where λ_d is the trivial eigenvalue associated with $\boldsymbol{\gamma}'(0)$.

Theorem 7.3.2. *With the above notation, if the nontrivial eigenvalues of* $\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0))$ *satisfy* $|\lambda_k| < 1$ *for* $k = 1, \ldots, d-1$, *then* $\boldsymbol{\gamma}$ *is asymptotically stable. If any eigenvalue of* $\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0))$ *satisfies* $|\lambda_k| > 1$, *then* $\boldsymbol{\gamma}$ *is unstable.*

The proof, which will be discussed in Section 7.3.3, is based on the Poincaré map, which we define in the next subsection.

In the following remark, we record that the theorem extends to cases such as the equations

$$\begin{aligned} \theta_1' &= \omega_1 + K_1 \sin(\theta_2 - \theta_1), \\ \theta_2' &= \omega_2 + K_2 \sin(\theta_1 - \theta_2), \end{aligned} \tag{7.24}$$

of Example 4, Section 7.1.2, which has periodic solutions *if* it is regarded as an ODE on the torus $\mathbb{T}^2$. We define the flow map in such cases *without* reducing its components modulo 2π. Of course, such a translation would have no effect on $\mathbf{D}\boldsymbol{\varphi}$, anyway.

Remark 7.3.3. Theorem 7.3.2 is still valid for solutions of an ODE that are periodic only when considered on the cylinder or torus.

By way of illustration, let us apply the extended theorem to (7.24), say focusing on the periodic solution $\boldsymbol{\gamma}_c(t)$ (notation of Section 7.1.2). We may calculate the differential of the flow map $\boldsymbol{\varphi} : \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}^2$ with Theorem 4.6.1. Specifically, $\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}_c(0))$ acting on a vector $\mathbf{b}$ is the solution at time T of the IVP

$$\mathbf{w}' = A(t)\mathbf{w}, \quad \mathbf{w}(0) = \mathbf{b},$$

where $A(t) = \mathbf{DF}(\boldsymbol{\gamma}_c(t))$. Substituting the solution (7.8) into $\mathbf{DF}$, we find that $A(t)$ is the constant matrix

$$A = \begin{bmatrix} -K_1 \cos\alpha_c & K_1 \cos\alpha_c \\ K_2 \cos\alpha_c & -K_2 \cos\alpha_c \end{bmatrix}.$$

Thus $\mathbf{D}\boldsymbol{\varphi}(t, \boldsymbol{\gamma}_c(0)) = e^{tA}$. Now A has eigenvalues $-(K_1 + K_2)\cos\alpha_c$ and 0, so for $t = T$, the differential $\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}_c(0))$ has eigenvalues $\lambda_1 = e^{-T(K_1+K_2)\cos\alpha_c}$ and $\lambda_2 = 1$. It follows from (7.10) that $\cos\alpha_c > 0$, so $|\lambda_1| < 1$, and hence $\boldsymbol{\gamma}_c(t)$ is asymptotically stable. (By a similar argument, $\boldsymbol{\gamma}_d(t)$ is unstable.)

This example was hand picked to yield simple calculations. Usually, Theorem 7.3.2 is harder to apply than Theorem 6.1.1, its close analogue about the stability of equilibria of ODEs. Nevertheless, the result and especially the key idea in its proof—the *Poincaré map*—are of the utmost importance theoretically.

7.3.2 Basics of the Poincaré Map

To construct the Poincaré map[11] of a periodic solution, we first

- choose a base point on the trajectory $\{\boldsymbol{\gamma}(t)\}$, say the position $\boldsymbol{\gamma}(0)$ at time zero, and
- choose a small section Σ of a $(d-1)$-dimensional hyperplane that contains $\boldsymbol{\gamma}(0)$ and is transverse[12] to the periodic orbit there.

We may specify Σ through a level set of a (scalar-valued) linear function, say

$$\Sigma = \{\mathbf{b} \in B(\boldsymbol{\gamma}(0), \eta) \,:\, \langle \mathbf{N}, \mathbf{b}\rangle = \alpha\}\,, \tag{7.25}$$

where $\mathbf{N} \in \mathbb{R}^d$, $\alpha = \langle \mathbf{N}, \boldsymbol{\gamma}(0)\rangle$ is a constant, and $\eta > 0$ specifies the size of Σ. There is no reason to insist on orthogonality of Σ and the periodic orbit at $\boldsymbol{\gamma}(0)$, although this could be accomplished by choosing $\mathbf{N} = \boldsymbol{\gamma}'(0)$. In any case, transversality requires that

$$\langle \mathbf{N}, \boldsymbol{\gamma}'(0)\rangle \neq 0. \tag{7.26}$$

Now for initial conditions $\mathbf{b} \in \Sigma$, consider the IVP

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0) = \mathbf{b}. \tag{7.27}$$

If $\mathbf{b} = \boldsymbol{\gamma}(0)$, then the solution of (7.27) crosses Σ when $t = T$, the minimal period of $\boldsymbol{\gamma}(t)$, precisely at the point $\boldsymbol{\gamma}(0)$. The Poincaré map focuses on the question, *starting from any* $\mathbf{b} \in \Sigma$, *when and especially where does the solution* $\boldsymbol{\varphi}(t, \mathbf{b})$ *of*

[11] It is customary to say *the* Poincaré map, but in fact there are many. "The" is justified in the sense that all of these mappings may be transformed to one another (on appropriately small neighborhoods) by changes of coordinates.

[12] "Transverse" in this context reduces to the condition that $\boldsymbol{\gamma}'(0)$ is not tangent to Σ.

Incidentally, it would not create any mathematical difficulties to assume that Σ was a curved surface, but writing precise equations in this case requires either the abstract language of differential geometry or some heavy-handed notation. Not liking either option, we have assumed that Σ is flat.

(7.27) *next cross* Σ*?* If Σ is given by (7.25), "when" is answered by defining the function $p(\mathbf{x}) = \langle \mathbf{N}, \mathbf{x} \rangle - \alpha$ and solving the equation

$$p(\boldsymbol{\varphi}(t, \mathbf{b})) = 0 \tag{7.28}$$

for t, say $t = \tau(\mathbf{b})$; and "where" is answered by the formula that defines the Poincaré map $\boldsymbol{\Pi}$,

$$\boldsymbol{\Pi}(\mathbf{b}) = \boldsymbol{\varphi}(\tau(\mathbf{b}), \mathbf{b}). \tag{7.29}$$

Theorem 7.3.4. *In the above context, there exists a neighborhood*[13] $\mathcal{N} \subset \mathbb{R}^d$ *of* $\boldsymbol{\gamma}(0)$ *such that (i) if* $\mathbf{b} \in \Sigma \cap \mathcal{N}$*, then* (7.28) *has a unique solution* $\tau(\mathbf{b}) \approx T$ *that is a* $\mathcal{C}^1$ *function of* $\mathbf{b}$*, and (ii) equation* (7.29) *defines a* $\mathcal{C}^1$ *map* $\boldsymbol{\Pi} : \Sigma \cap \mathcal{N} \to \Sigma$*.*

Remark: Since Σ is (a piece of) a hypersurface, $\tau(\mathbf{b})$ may be regarded as a function on a $(d-1)$-dimensional Euclidean space. Thus the statement "τ is a smooth function on Σ" is readily interpreted, which is our reward for assuming that Σ is flat.

Proof of Theorem 7.3.4. (i) Let us apply the implicit function theorem to (7.28). By periodicity, $\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0)) = \boldsymbol{\gamma}(0)$, so if $\mathbf{b} = \boldsymbol{\gamma}(0)$, then $t = T$ solves (7.28). Of course, p is differentiable, and we know from Theorem 4.6.1 that $\boldsymbol{\varphi}$ is also differentiable. From the chain rule,

$$\frac{\partial}{\partial t} p(\boldsymbol{\varphi}(t, \boldsymbol{\gamma}(0)))\Big|_{t=T} = \langle \mathbf{N}, \partial_t \boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0)) \rangle. \tag{7.30}$$

Invoking periodicity again, we note for the second factor in the inner product that

$$\partial_t \boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0)) = \partial_t \boldsymbol{\varphi}(0, \boldsymbol{\gamma}(0)) = \boldsymbol{\gamma}'(0).$$

Thus, by the transversality condition (7.26), we conclude that (7.30) is nonzero. This proves Part (i) of the theorem. Regarding Part (ii), $\boldsymbol{\Pi}$ is $\mathcal{C}^1$, because it is the composition of continuously differentiable functions. □

Remark: If the radius η of the Poincaré section (7.25) is sufficiently small, then $\tau(\mathbf{b})$ represents the *first* time at which the trajectory $\boldsymbol{\varphi}(t, \mathbf{b})$ returns to Σ. Thus, the Poincaré map is sometimes called the *first-return map*. If η were too large, the periodic orbit $\boldsymbol{\gamma}$ could cross Σ "prematurely" without completing a full cycle (see Figure 7.8). However, even if η were too large, the requirement that $\tau(\mathbf{b})$ depend smoothly on $\mathbf{b}$ selects the right solution of (7.28).

Extending the notation, we write

$$T(\Sigma) = \{\mathbf{b} \in \mathbb{R}^d : \langle \mathbf{N}, \mathbf{b} \rangle = 0\}$$

[13] In many examples it might seem as though we could drop the neighborhood $\mathcal{N}$ and just make Σ small. This idea doesn't work when $\boldsymbol{\gamma}$ is unstable; see Exercise 21.

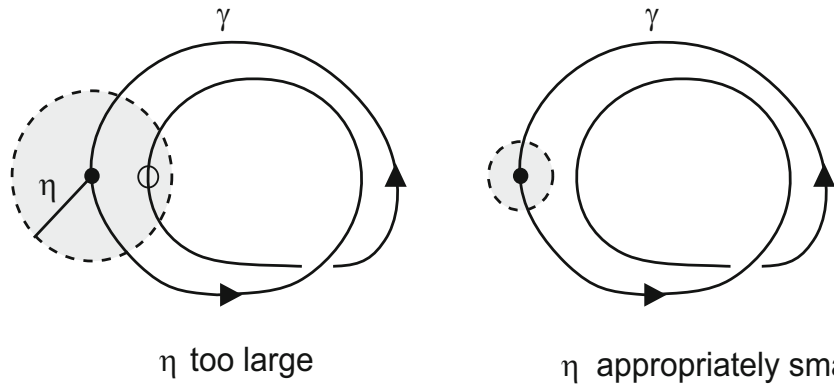

Figure 7.8: *If η is chosen too large, $\boldsymbol{\gamma}$ could cross Σ prematurely, as indicated by an open circle in the left panel. In both panels, the solid black dot represents $\boldsymbol{\gamma}(0)$.*

for the tangent space to Σ. This set differs from Σ in that it is unbounded and has been translated from $\boldsymbol{\gamma}(0)$ to the origin; moreover, $T(\Sigma)$ is a $(d-1)$-dimensional linear subspace of $\mathbb{R}^d$. For every $\mathbf{b} \in \Sigma \cap \mathcal{N}$, the differential of the Poincaré map defines a linear transformation $\mathbf{D\Pi}(\mathbf{b}) : T(\Sigma) \to T(\Sigma)$. The following proposition connects $\mathbf{D\Pi}$ at the base point $\boldsymbol{\gamma}(0)$ to the differential of the flow map.

Proposition 7.3.5. *If $\mathbf{D\Pi}(\boldsymbol{\gamma}(0))$ has eigenvalues $\lambda_1, \ldots, \lambda_{d-1}$, then $\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0))$ has the same eigenvalues $\lambda_1, \ldots, \lambda_{d-1}$ plus the trivial eigenvalue $\lambda_d = 1$.*

Proof of Proposition 7.3.5. Choose a special basis for $\mathbb{R}^d$ as follows. By transversality, $\mathbb{R}^d$ is spanned by $T(\Sigma)$ and the one-dimensional subspace $\mathbb{R}\{\boldsymbol{\gamma}'(0)\}$. Thus, if $\mathbf{v}_1, \ldots, \mathbf{v}_{d-1}$ is a basis for $T(\Sigma)$, then $\mathbf{v}_1, \ldots, \mathbf{v}_{d-1}, \boldsymbol{\gamma}'(0)$ is a basis for $\mathbb{R}^d$. Let M_φ be the $d \times d$ matrix that represents $\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0))$ with respect to the basis $\mathbf{v}_1, \ldots, \mathbf{v}_{d-1}, \boldsymbol{\gamma}'(0)$, and let M_Π be the $(d-1) \times (d-1)$ matrix that represents $\mathbf{D\Pi}(\boldsymbol{\gamma}(0))$ with respect to the basis $\mathbf{v}_1, \ldots, \mathbf{v}_{d-1}$. The proposition follows immediately from the following lemma.

Lemma 7.3.6. *With respect to the above bases, $\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0))$ has the block structure*

$$M_\varphi = \begin{bmatrix} M_\Pi & \mathbf{0} \\ \mathbf{c} & 1 \end{bmatrix}, \tag{7.31}$$

where $\mathbf{c}$ is a $(d-1)$-component row vector.

Proof. In proving Lemma 7.3.1, we saw that $\boldsymbol{\gamma}'(0)$ is an eigenvector of $\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0))$ with eigenvalue 1. It follows from this fact that the last column of M_φ has the structure indicated in (7.31).

Regarding the other columns of M_φ, we compute the directional derivative of (7.29) in an arbitrary direction $\mathbf{v} \in T(\Sigma)$ and evaluate at $\boldsymbol{\gamma}(0)$ to obtain

$$\mathbf{D\Pi}(\boldsymbol{\gamma}(0)) \cdot \mathbf{v} = \partial_t \boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0)) \, \langle \nabla_{\mathbf{b}} \tau(\boldsymbol{\gamma}(0)), \mathbf{v} \rangle + \mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0)) \cdot \mathbf{v}. \tag{7.32}$$

Recognizing that $\partial_t \boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0)) = \boldsymbol{\gamma}'(0)$ and rearranging terms, we obtain

$$\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0)) \cdot \mathbf{v} = \mathbf{D\Pi}(\boldsymbol{\gamma}(0)) \cdot \mathbf{v} - \langle \nabla_{\mathbf{b}} \tau(\boldsymbol{\gamma}(0)), \mathbf{v} \rangle \, \boldsymbol{\gamma}'(0).$$

This equation verifies the first $d-1$ columns of (7.31) and moreover determines that the last row of M_φ is given by $\mathbf{c} = -\nabla_{\mathbf{b}} \tau(\boldsymbol{\gamma}(0))$. This proves both the lemma and Proposition 7.3.5. □

7.3.3 Discrete-Time Dynamics and the Proof of Theorem 7.3.2

The main idea of the proof is easily conveyed through a trivial example, the equation given in polar coordinates by

$$r' = 1 - r, \qquad \theta' = 1. \tag{7.33}$$

Of course, $r(t) \equiv 1$, $\theta(t) = t$ satisfies (7.33), or in more geometric terms, the unit circle is a periodic orbit of this equation. Relative to this periodic solution, we choose the transverse section

$$\Sigma = \{(b_1, b_2) : 1/2 < b_1 < 3/2, \;\; b_2 = 0\} \;=\; (1/2, 3/2) \times \{0\}.$$

To define the Poincaré map, given $b \in (1/2, 3/2)$, we must solve (7.33) with initial conditions

$$r(0) = b, \qquad \theta(0) = 0; \tag{7.34}$$

this yields $(r(t), \theta(t)) = (1 + (b-1)e^{-t}, t)$. Solving the equation for the time of first return, $\theta(t) = 0 \pmod{2\pi}$, we find that $\tau(b) \equiv 2\pi$; thus,

$$\mathbf{\Pi}(b) = r(2\pi) = 1 + e^{-2\pi}(b-1). \tag{7.35}$$

Regarding Theorem 7.3.2, we ask, does the trajectory that originates from $\mathbf{b}_0 = (b, 0)$ converge to the unit circle? Well, after one circuit of the origin, it crosses Σ at $\mathbf{b}_1 = \mathbf{\Pi}(\mathbf{b}_0)$, which by (7.35) is a factor $e^{-2\pi}$ closer to the circle than $\mathbf{b}_0$; after two circuits, it crosses at $\mathbf{b}_2 = \mathbf{\Pi}(\mathbf{b}_1)$, which is a factor $(e^{-2\pi})^2$ closer to the circle; and so on. So we have reproved the obvious, that the trajectory does converge. But this strategy generalizes.

Suppose $\boldsymbol{\gamma}(t)$ is a periodic trajectory of an ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ with Poincaré map $\mathbf{\Pi} : \Sigma \cap \mathcal{N} \to \Sigma$. If $\mathbf{b}_0 \in \Sigma \cap \mathcal{N}$, we follow the trajectory $\boldsymbol{\varphi}(t, \mathbf{b}_0)$ forward in time. When $t = \tau(\mathbf{b}_0)$, the trajectory makes its first return to Σ, crossing the Poincaré

section at the point $\mathbf{b}_1 = \mathbf{\Pi}(\mathbf{b}_0)$. If $\mathbf{b}_1 \in \Sigma$ happens to belong to $\Sigma \cap \mathcal{N}$, then the trajectory crosses Σ a second time at $\mathbf{b}_2 = \mathbf{\Pi}(\mathbf{b}_1)$. Continuing for as long as these iterates remain in $\Sigma \cap \mathcal{N}$, we may recursively define a sequence of subsequent crossings, $\mathbf{b}_{n+1} = \mathbf{\Pi}(\mathbf{b}_n) = \mathbf{\Pi}^{n+1}(\mathbf{b}_0)$. Thus, the Poincaré map allows us to recast "continuous time" questions about the behavior of trajectories as $t \to \infty$ in terms of "discrete time" questions about the convergence of this sequence. In more picturesque language, the Poincaré map lets us examine trajectories under a strobe light.

It's time to formalize concepts. By a *discrete-time dynamical system*[14] we mean a mapping $\mathbf{\Psi} : \mathcal{U} \to \mathbb{R}^d$, where $\mathcal{U} \subset \mathbb{R}^d$ is open. "Time" for such a system counts the number of iterations of $\mathbf{\Psi}$; i.e., we define $\mathbf{\Psi}^0(\mathbf{z}) = \mathbf{z}$ and

$$\mathbf{\Psi}^{n+1}(\mathbf{z}) = \mathbf{\Psi}(\mathbf{\Psi}^n(\mathbf{z}))$$

for as long as the iterates remain in $\mathcal{U}$. Of course, if $\mathbf{\Psi}(\mathcal{U}) \subset \mathcal{U}$, the iteration continues indefinitely.

A *fixed point* of a discrete dynamical system is a point $\mathbf{z}_* \in \mathcal{U}$ such that $\mathbf{\Psi}(\mathbf{z}_*) = \mathbf{z}_*$. This concept is analogous to an equilibrium of an ODE. *Lyapunov stability*, *asymptotic stability*, and *instability* of a fixed point are defined with the obvious modifications of the definitions for an equilibrium. For example, consider the one-dimensional discrete dynamical system $\mathbf{\Psi} : \mathbb{R} \to \mathbb{R}$ given by $\mathbf{\Psi}(z) = z^2$. The map has two fixed points, 0 and 1; the former is asymptotically stable, and the latter is unstable. *(Show this directly, and relate it to the next proposition.)*

The eigenvalues of the Jacobian of $\mathbf{\Psi}$ provide a convenient test for asymptotic stability of a fixed point.

Theorem 7.3.7. *Let $\mathbf{z}_*$ be a fixed point of the $\mathcal{C}^1$-map $\mathbf{\Psi} : \mathcal{U} \to \mathbb{R}^d$. If every eigenvalue of the Jacobian $\mathbf{D\Psi}(\mathbf{z}_*)$ satisfies $|\lambda| < 1$, then $\mathbf{z}_*$ is asymptotically stable. If $|\lambda| > 1$ for some eigenvalue, then $\mathbf{z}_*$ is unstable.*

If $\mathbf{\Psi}$ were linear, say $\mathbf{\Psi}(\mathbf{z}) = A\mathbf{z}$ for some matrix A, the theorem would reduce to Exercise 2.13(b). The proof for nonlinear $\mathbf{\Psi}$, which we pose as Exercise 13(a) below, is based on approximation by the linearization of $\mathbf{\Psi}$ at the fixed point. This proof is similar in spirit to, but technically simpler than, the analogous test for stability of equilibria of an ODE, Theorem 6.1.1.

Note that $\mathbf{D\Psi}(\mathbf{z}_*)$ might have complex eigenvalues $\lambda = \alpha + i\beta$, in which case $|\lambda|$ means $\sqrt{\alpha^2 + \beta^2}$, but the theorem still holds.

[14] Discrete-time dynamical system are important and interesting in their own right, independent of any connection to Poincaré maps. Chapter 10 of [81] has a brief, readable introduction to the behavior of one-dimensional maps, and [17] gives a more thorough, but still readable, treatment. (See also Sections 9.9.2 and 10.6(a) below.)

Proof of Theorem 7.3.2 (assuming Theorem 7.3.7). If $\boldsymbol{\gamma}(t)$ is a periodic solution of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, then $\boldsymbol{\gamma}(0)$ is a fixed point of the Poincaré map $\boldsymbol{\Pi}$. In Exercise 13(b) we ask you to sharpen the correspondence between continuous-time and discrete-time systems by showing that $\boldsymbol{\gamma}(0)$ is a Lyapunov stable, asymptotically stable, or unstable fixed point of the Poincaré map iff the trajectory $\boldsymbol{\gamma}(t)$ has the corresponding stability behavior with respect to the ODE. Regarding $\boldsymbol{\Pi}$ as a discrete dynamical system on a $(d-1)$-dimensional Euclidean space, we can relate the stability of $\boldsymbol{\gamma}(0)$ to eigenvalues of $\mathbf{D}\boldsymbol{\Pi}(\boldsymbol{\gamma}(0))$ through Theorem 7.3.7. Of course, by Proposition 7.3.5 the eigenvalues of $\mathbf{D}\boldsymbol{\Pi}(\boldsymbol{\gamma}(0))$ are just the nontrivial eigenvalues of $\mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0))$. Thus, Theorem 7.3.2 follows from Theorem 7.3.7. □

A reminder: please reexamine this proof to check that it works equally well for ODEs that have a periodic solution on the cylinder or torus, as asserted in Remark 7.3.3.

7.4 Stability of the Limit Cycle in the Torqued Pendulum

In Section 7.2.2(b) we applied the Poincaré–Bendixson theorem to show that the torqued pendulum equation (7.12) has a periodic solution (as an ODE on the cylinder) if $\mu > 1$. Now we prove that this periodic solution is asymptotically stable. In proving this, we give an independent proof of its existence and we show that it is unique.

To begin, we eliminate time from (7.12): y as a function of x satisfies the scalar ODE $dy/dx = F(y, x)$, where

$$F(y, x) = \frac{y'}{x'} = \frac{\mu - \sin x}{y} - \beta. \tag{7.36}$$

We restrict y to an interval $\varepsilon \le y \le M$, where we choose ε such that

$$0 < \varepsilon < \frac{\mu - 1}{\beta}, \quad \text{which implies that } F(\varepsilon, x) > 0,$$

and choose M such that

$$M > \frac{\mu + 1}{\beta}, \quad \text{which implies that } F(M, x) < 0.$$

Thus, as illustrated in Figure 7.9(a), if $\varepsilon \le b \le M$, then the solution of the IVP

$$\frac{dy}{dx} = F(y, x), \qquad y(0) = b, \tag{7.37}$$

is trapped between the lines $y = \varepsilon$ and $y = M$. More formally, we have the following:

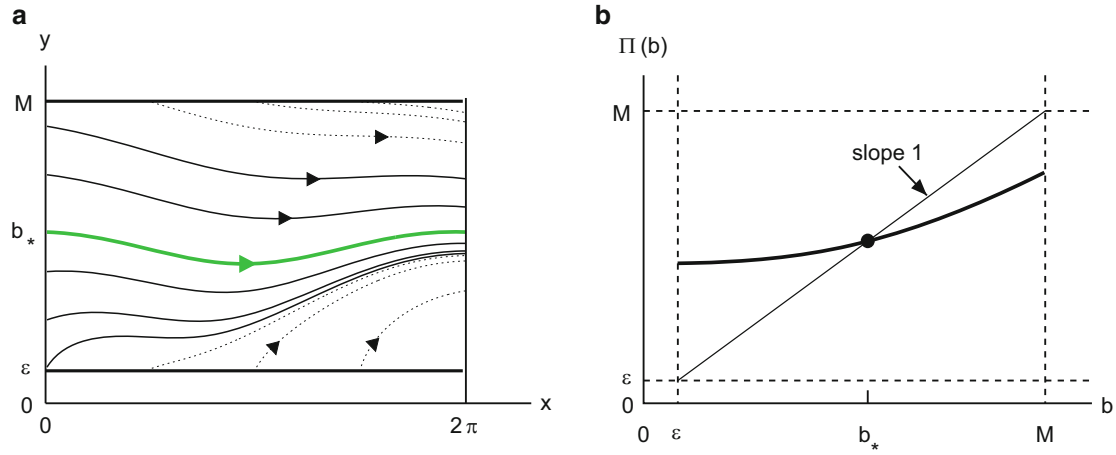

Figure 7.9: *(a) Trajectories for (7.36), an ODE for orbits of the torqued pendulum equation (7.12), with $\beta = 1/2$, $\mu = 3/2$. Flow is restricted to the strip $\{\varepsilon < y < M\}$, where $\varepsilon = 1/2$ and $M = 6$. The trajectory shown in green, which returns to the same height (i.e., $y(2\pi) = y(0) = b_*$), picks out a periodic solution of (7.12). (b) Graph of Π, the Poincaré-like map $b \mapsto \varphi(2\pi, b)$ for (7.36). The intersection of the graph with the diagonal locates the periodic solution of the torqued-pendulum equation (7.12).*

Claim 1: If $\varepsilon \le b \le M$, then the solution $\varphi(x, b)$ of the IVP (7.37) exists for all $x \ge 0$ and moreover satisfies $\varepsilon \le \varphi(x, b) \le M$.

Leaving it to the dedicated reader to construct a careful proof of this claim, we move on:

Claim 2: The derivative $\partial\varphi/\partial b$ satisfies the estimate

$$0 < \frac{\partial \varphi}{\partial b}(x, b) < 1.$$

Proof of Claim 2. According to the generalization of Theorem 4.6.1 to ODEs with variable coefficients, the solution of (7.37) depends differentiably on the initial condition b, and moreover, $\partial\varphi/\partial b(x, b)$ satisfies the linear IVP

$$\frac{dw}{dx} = -\left(\frac{\mu - \sin x}{\varphi^2(x, b)}\right) w, \qquad w(0) = 1,$$

where the RHS of the equation was obtained by differentiation of (7.36) with respect to y. The claim follows from the observation that the coefficient of w in this equation is negative. □

Using the solution of (7.37), define a map[15] $\Pi : [\varepsilon, M] \to [\varepsilon, M]$ by the formula $\Pi(b) = \varphi(2\pi, b)$. As illustrated in Figure 7.9(b), it follows from our claims above that there is a unique point $b_* \in (\varepsilon, M)$ where the graph of Π crosses the diagonal in $[\varepsilon, M] \times [\varepsilon, M]$.

Now let $x_*(t), y_*(t)$ be the solution of (7.12) with initial conditions $x_*(0) = 0$, $y_*(0) = b_*$. As shown in Section 4.3.3, this IVP has a solution for all $t \geq 0$. Indeed, by Claim 1, $y_*(t) \geq \varepsilon$. Since $x_*'(t) = y_*(t) \geq \varepsilon$, it follows that $x_*(t)$ increases steadily, and there is a positive time $T \leq 2\pi/\varepsilon$ such that $x_*(T) = 2\pi$. Since $\mathbf{b}_*$ is a fixed point of Π, we have $y_*(T) = b_*$. Now

$$x_*(t+T) - 2\pi,\ y_*(t+T) \tag{7.38}$$

also satisfies (7.12) and has the same initial conditions as $x_*(t), y_*(t)$. Thus, by uniqueness, (7.38) coincides with $x_*(t), y_*(t)$, which therefore provides our desired periodic solution of (7.12). Examining the above construction, including solutions outside the strip $\{\varepsilon < y < M\}$, you may show that $x_*(t), y_*(t)$ is the only periodic solution.

Finally, we deduce that $x_*(t), y_*(t)$ is asymptotically stable from Theorem 7.3.2, because $|\Pi'(b_*)| < 1$. Indeed, $|\Pi'(b)| < 1$ for all $b \in (\varepsilon, M)$.

7.5 Van der Pol with Small β: Weakly Nonlinear Analysis

The Poincaré–Bendixson theorem asserts the existence of a periodic orbit without giving any specific information about it. We turn now to methods for calculating periodic (and other) solutions approximately. In this section we use perturbation theory to describe the limit cycle of the van der Pol system

$$\begin{aligned} x' &= y, \\ y' &= -x - \beta(x^2 - 1)y, \end{aligned} \tag{7.39}$$

in the limit of small β. Specifically, we show that a circle of radius 2, in symbols

$$x(t) \approx 2\cos t, \quad y(t) \approx -2\sin t, \tag{7.40}$$

is an approximate solution (to lowest order in β) of (7.39). Although we focus on the van der Pol system, these techniques are generally applicable to what are called *weakly nonlinear* equations. Other examples are posed in Exercises 15–17; see also [81] (Section 7.6, including the associated exercises) or look online.

To prepare the way for analysis of the van der Pol equation, we first introduce perturbation theory through two examples. You will probably feel that these calcu-

[15] Although we use Poincaré-map notation, we have not yet related this map to any periodic solution.

lations are rather long. We urge you to persevere. Developing tolerance for slogging through longer calculations is part of the maturation process for mathematicians.

7.5.1 Two Illustrative Examples of Perturbation Theory

Example 1: Consider the one-parameter family of initial value problems

$$x' = -x + \varepsilon x^2, \quad x(0) = 1, \tag{7.41}$$

where ε is a small parameter. If $\varepsilon = 0$, then (7.41) has the solution $x(t) = e^{-t}$. Even if $\varepsilon \neq 0$, (7.41) may be solved exactly, because the equation is separable. However, let us temporarily ignore this exact solution and use perturbation theory to obtain the $\mathcal{O}(\varepsilon^2)$-approximation

$$x(t) \approx e^{-t} + \varepsilon(e^{-t} - e^{-2t}). \tag{7.42}$$

This perturbation-theory result estimates how much the small positive term εx^2 on the RHS of (7.41) slows down the decay of the solution.

In perturbation theory, in attacking a one-parameter family of problems like (7.41), one considers *all small values of* ε *simultaneously*. To emphasize this point of view, we write $x(t;\varepsilon)$ for the solution, indicating the dependence on ε, and we suppose $x(t;\varepsilon)$ has a power-series expansion[16]

$$x(t;\varepsilon) = x_0(t) + \varepsilon x_1(t) + \varepsilon^2 x_2(t) + \dots \ . \tag{7.43}$$

Inserting the expansion into (7.41) yields

$$x_0' + \varepsilon x_1' + \varepsilon^2 x_2' + \dots \ = \ -[x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \dots] \ + \varepsilon[x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \dots]^2,$$

where the dots indicate terms that are of order ε^3 or higher. Expanding out the squared term, we obtain

$$x_0' + \varepsilon x_1' + \varepsilon^2 x_2' + \dots \ = \ -x_0 + \varepsilon[-x_1 + x_0^2] + \varepsilon^2[-x_2 + 2x_1 x_0] + \dots \ .$$

For each t, the LHS and RHS of this equation are functions of ε, and for them to be the *same* functions, the coefficient of each power of ε on the left must equal the corresponding coefficient on the right. This principle may be used to calculate ODEs for every coefficient $x_n(t)$ in (7.43). In particular, matching terms of corresponding orders through ε^2 generates the ODEs

[16] In some cases this series may not converge—it could be only asymptotic (cf. Section 1.4 of [43] or Section 6.2 of [98]). Whether the series converges or not doesn't really have much impact on the application of the method. In this example, one may deduce from the exact solution (7.44) that the series does actually converge in some neighborhood of zero.

$$\begin{aligned}\mathcal{O}(\varepsilon^0): &\quad x_0' + x_0 = 0,\\ \mathcal{O}(\varepsilon^1): &\quad x_1' + x_1 = x_0^2,\\ \mathcal{O}(\varepsilon^2): &\quad x_2' + x_2 = 2x_1x_0.\end{aligned}$$

Since the initial condition $x(0,\varepsilon) = 1$ holds for all ε, it follows by a similar matching argument that

$$x_0(0) = 1, \quad x_1(0) = 0, \quad x_2(0) = 0, \quad \dots .$$

We attack the equations sequentially. First, $x_0(t) = e^{-t}$ satisfies the $\mathcal{O}(\varepsilon^0)$-IVP. Given $x_0(t)$, the $\mathcal{O}(\varepsilon)$-problem is an inhomogeneous IVP whose solution is $x_1(t) = e^{-t} - e^{-2t}$. The $\mathcal{O}(\varepsilon^2)$-problem may be solved similarly.

The first two terms of (7.43) yield the approximation (7.42). To assess the accuracy of this approximation, we solve (7.41) explicitly via separation of variables, obtaining

$$x(t,\varepsilon) = \frac{1}{\varepsilon + (1-\varepsilon)e^t}. \tag{7.44}$$

Please check that for each t,

$$x(t,\varepsilon) = x_0(t) + \varepsilon x_1(t) + \mathcal{O}(\varepsilon^2), \tag{7.45}$$

as expected.

In point of fact, (7.45) holds *uniformly for* $0 \le t < \infty$. Such uniformity is rare. One would expect errors in the approximation to accumulate as time increases. Thus, normally an expansion like (7.45) would be uniform only over finite intervals, say $0 \le t \le T$. However, problem (7.41) was hand-picked so that all coefficients $x_n(t)$ in (7.43) decay as $t \to \infty$. In other words, the estimate (7.45) is uniform over $[0,\infty)$ *but only because both sides tend to zero for large* t.

Example 2: Our next example illustrates the accumulation of errors in a power-series approximation as t increases *and* shows how to cope with this difficulty. Consider the one-parameter family of initial value problems

$$x'' + (1+\varepsilon)x = 0, \qquad x(0) = 1,\ x'(0) = 0. \tag{7.46}$$

The exact solution of the IVP is $x(t) = \cos(\sqrt{1+\varepsilon}\, t)$, which is periodic; in particular, it does *not* decay as $t \to \infty$. Ignoring the exact solution and seeking an approximate solution as above, we suppose $x(t,\varepsilon) = x_0(t) + \varepsilon x_1(t) + \dots$ and insert this expansion into the ODE,

$$x_0'' + \varepsilon x_1'' + \dots + (1+\varepsilon)[x_0 + \varepsilon x_1 + \dots] = 0.$$

Multiplying out the product and grouping like powers of ε, we obtain ODEs

$$\mathcal{O}(\varepsilon^0): \quad x_0'' + x_0 = 0, \qquad \mathcal{O}(\varepsilon^1): \quad x_1'' + x_1 = -x_0,$$

subject to initial conditions

$$x_0(0) = 1, \ x_0'(0) = 0, \qquad x_1(0) = x_1'(0) = 0.$$

The solution of the leading-order problem is $x_0(t) = \cos(t)$. Substitution of $x_0(t)$ into the $\mathcal{O}(\varepsilon)$-equation leads to an inhomogeneous ODE for x_1 with a resonant[17] forcing term:

$$x_1'' + x_1 = -\cos t.$$

Imposing the initial conditions, we obtain

$$x_1(t) = -\frac{1}{2}\, t \sin t,$$

which yields the two-term asymptotic approximation

$$x(t,\varepsilon) \approx x_0(t) + \varepsilon x_1(t) = \cos t - \varepsilon \frac{t}{2} \sin t. \tag{7.47}$$

The error in (7.47) is $\mathcal{O}(\varepsilon^2)$, uniformly for t in every finite interval $0 \le t \le T$. However, this approximation fails miserably as $t \to \infty$ (see Figure 7.10). Indeed, $x_1(t)$, the supposedly small first correction, in fact becomes *large* compared to $x_0(t)$!

The problem with the simple ansatz above, $x(t,\varepsilon) = x_0(t) + \varepsilon x_1(t) + \ldots$, is that both $x(t,\varepsilon)$ and $x_0(t)$ are periodic, but *they have different periods*, with the period of $x(t,\varepsilon)$ depending on ε. Because of this mismatch, there's no way the two functions could remain close to one another indefinitely.

A momentary interlude for terminology: a term in an expansion $x(t,\varepsilon) = x_0(t) + \varepsilon x_1(t) + \ldots$ with a coefficient $x_n(t)$ that grows without bound as $t \to \infty$ is sometimes called a *secular* term. Such terms, which typically come from resonant forcing as above, are incompatible with periodic behavior.

The *Poincaré–Lindstedt method*[18] allows one to obtain an approximate solution of (7.46) and similar problems with periodic solutions that holds for arbitrarily large times. In this method one introduces a scaled time

$$\tau(t,\varepsilon) = \omega(\varepsilon)t = (1 + \omega_1\varepsilon + \omega_2\varepsilon^2 + \ldots)t \tag{7.48}$$

[17] That is, the period of the forcing term equals the period of solutions of the homogeneous equation, $x_1'' + x_1 = 0$. The term *resonance* derives from Exercise 1.13.

[18] There are more general techniques for finding accurate large-time approximations, e.g., the method of multiple scales (a.k.a. two-timing); see [43] (Chapter 3) or [81] (Section 7.6). We present a less general method here, because in our estimation, it is less confusing to apply.

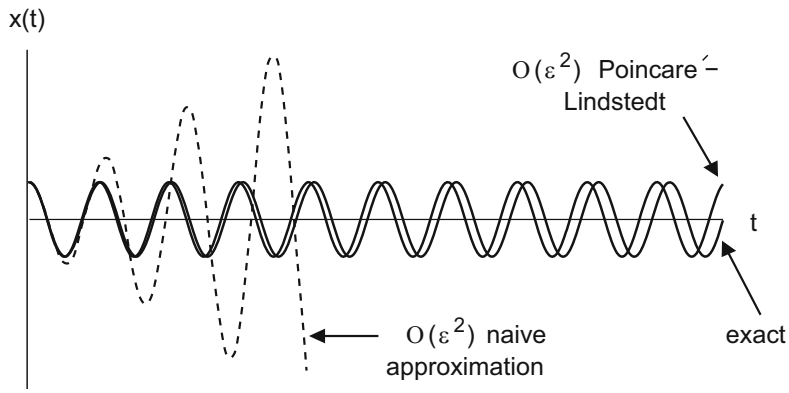

Figure 7.10: *Comparison of the exact solution of* (7.46) *with its "naive" two-term regular perturbation expansion approximation* (7.47) *and Poincaré–Lindstedt approximation* (7.50) *for* $0 \le t \le 50$, *assuming* $\varepsilon = 0.5$. *(The graph of the naive approximation stops when it grows out of the scale of the viewing window.)*

and seeks a power-series expansion of $x(t, \varepsilon)$,

$$x(t,\varepsilon) = x_0(\tau(t,\varepsilon)) + \varepsilon x_1(\tau(t,\varepsilon)) + \varepsilon^2 x_2(\tau(t,\varepsilon)) + \dots, \tag{7.49}$$

in which the coefficients $x_n(\tau)$ depend on the scaled time. We will show that *if the scaling factor* $\omega(\varepsilon)$ *in* (7.48) *is chosen cleverly, it can compensate for the mismatch between the periods of the exact and approximate solutions.* Of course, the \$64 question is how to choose the scaling factor. In the calculation below it will be seen that if we require at each order that no secular terms arise, then both the undetermined coefficients ω_n in (7.48) and the terms $x_n(\tau)$ in the series are uniquely determined.

Let's get on with it. Invoking the chain rule to derive $d/dt = \omega(\varepsilon)\, d/d\tau$, we may rewrite the ODE in (7.46) as

$$\omega^2(\varepsilon)\frac{d^2x}{d\tau^2} + (1+\varepsilon)x = 0.$$

Inserting the expansions for $\omega(\varepsilon)$ and $x(t, \varepsilon)$, we obtain

$$(1 + 2\omega_1\varepsilon + \dots)\left[\frac{d^2x_0}{d\tau^2} + \varepsilon\frac{d^2x_1}{d\tau^2} + \dots\right] + (1+\varepsilon)[x_0 + \varepsilon x_1 + \dots] = 0.$$

On expanding the products and grouping terms according to their order in ε, the leading-order and next-order correction terms obey the equations

$$\mathcal{O}(\varepsilon^0): \quad \frac{d^2x_0}{d\tau^2} + x_0 = 0,$$
$$\mathcal{O}(\varepsilon^1): \quad \frac{d^2x_1}{d\tau^2} + x_1 = -2\omega_1 \frac{d^2x_0}{d\tau^2} - x_0.$$

Since τ is proportional to t, these functions must satisfy the initial conditions at $\tau = 0$,

$$x_0(0) = 1, \ \frac{dx_0}{d\tau}(0) = 0, \qquad x_1(0) = \frac{dx_1}{d\tau}(0) = 0.$$

The solution of the leading-order equation is $x_0(\tau) = \cos\tau$. Given $x_0(\tau)$, the $\mathcal{O}(\varepsilon)$-equation becomes

$$\frac{d^2x_1}{d\tau^2} + x_1 = (2\omega_1 - 1)\cos\tau.$$

Now here is the key point: *to avoid a secular term in $x_1(\tau)$, we must require that* $\omega_1 = 1/2$, so that the RHS of this equation vanishes. The solution of the (now homogeneous) IVP for x_1 is the trivial function $x_1(\tau) \equiv 0$. Thus, modulo errors that are of order ε^2 or higher, our approximation of the true solution $x(t) = \cos(\sqrt{1+\varepsilon}\,t)$ is given by

$$x(t,\varepsilon) \ \approx \ \cos\left[(1+\varepsilon/2)t\right]. \tag{7.50}$$

This estimate is far more satisfactory than (7.47), since the only discrepancy between (7.50) and the exact solution of (7.46) is the $\mathcal{O}(\varepsilon^2)$ difference between their periods. Indeed, in Figure 7.10 we made the deliberately "large" choice of $\epsilon = 0.5$ so we could see the discrepancy without needing a vastly longer time scale.

To assess the error in (7.50) more quantitatively, note that $\sqrt{1+\varepsilon} = 1 + \varepsilon/2 + \mathcal{O}(\varepsilon^2)$; thus, (7.50) may be derived from the exact solution $\cos(\sqrt{1+\varepsilon}\,t)$ by neglecting terms that are $\mathcal{O}(\varepsilon^2 t)$ inside the argument of the cosine. Therefore, the error in (7.50) is small if $t \ll 1/\varepsilon^2$; by contrast, the error in (7.47) is small only if $t \ll 1/\varepsilon$, a much more stringent condition.

By carrying the Poincaré–Lindstedt approximation to successively higher orders, one can obtain approximations that are accurate if $t \ll 1/\varepsilon^n$ for every integer n. However, these higher-order calculations are tedious, and we suggest that you save your energy for higher-order calculations with the van der Pol equation, where the results hold greater interest.

A pessimist might complain that the above calculation seems a little mysterious, and we would agree. On the other hand, an optimist might exclaim how wonderfully it all works out, and we would again agree. Speaking personally, we have found that over time, the mystery in asymptotics seems to recede into the background, while the wonder remains and even grows.

7.5.2 Application to the van der Pol Equation

Let us now apply the Poincaré–Lindstedt method to a real problem, the van der Pol equation

$$x'' + \varepsilon(x^2 - 1)x' + x = 0, \tag{7.51}$$

where we write ε for the friction coefficient as a reminder that it is small. (The calculations are slightly simpler if we work with the second-order scalar equation rather than the equivalent first-order system (7.39).) In notable contrast with the linear equation (7.46), for which all solutions are periodic, the periodic solution of (7.51) is unique up to translation (in time). Without loss of generality, we may perform a translation in t such that a local maximum of the periodic solution of (7.51) is located at $t = 0$. Then this periodic solution will satisfy initial conditions

$$x(0) = b(\varepsilon), \quad x'(0) = 0, \tag{7.52}$$

where *the unknown amplitude* $b(\varepsilon) > 0$ *must be determined* along with the solution itself.

Here goes. Defining scaled time as in (7.48), we may rewrite (7.51) as

$$\omega^2(\varepsilon)\frac{d^2x}{d\tau^2} + \varepsilon(x^2 - 1)\omega(\varepsilon)\frac{dx}{d\tau} + x = 0.$$

Inserting expansions for $\omega(\varepsilon)$ and $x(t, \varepsilon)$ into the equation, we obtain

$$[1 + 2\omega_1\varepsilon + \cdots]\left[\frac{d^2x_0}{d\tau^2} + \varepsilon\frac{d^2x_1}{d\tau^2} + \cdots\right]$$

$$+\varepsilon[(x_0^2 - 1) + \cdots][1 + \cdots]\left[\frac{dx_0}{d\tau} + \cdots\right]$$

$$+[x_0 + \varepsilon x_1 + \cdots] = 0,$$

where we have retained only those terms that contribute to orders ε^0 and ε^1. Grouping terms of like order, we calculate the equations

$$\begin{aligned}
\mathcal{O}(\varepsilon^0): &\quad \frac{d^2x_0}{d\tau^2} + x_0 = 0, \\
\mathcal{O}(\varepsilon^1): &\quad \frac{d^2x_1}{d\tau^2} + x_1 = -2\omega_1\frac{d^2x_0}{d\tau^2} - (x_0^2 - 1)\frac{dx_0}{d\tau}.
\end{aligned}$$

Regarding initial conditions, we expand $b(\varepsilon)$ in a series

$$b(\varepsilon) = b_0 + b_1\varepsilon + \ldots \tag{7.53}$$

and deduce from (7.52) that

$$x_0(0) = b_0, \quad \frac{dx_0}{d\tau}(0) = 0, \quad \text{and} \quad x_1(0) = b_1, \quad \frac{dx_1}{d\tau}(0) = 0. \tag{7.54}$$

The solution of the lowest-order problem is $x_0(\tau) = b_0 \cos\tau$, where b_0 is yet to be determined. To determine b_0 we need to start the calculation at the next order!

We substitute $x_0(\tau)$ into the $\mathcal{O}(\varepsilon)$ equation to obtain

$$\frac{d^2x_1}{d\tau^2} + x_1 = 2\omega_1 b_0 \cos\tau - \left(b_0^2 \cos^2\tau - 1\right)(-b_0 \sin\tau).$$

We seek to annihilate all resonant forcing terms that would cause secular growth in $x_1(t)$. Doing so will determine both b_0 and ω_1. The problematic terms $2\omega_1 b_0 \cos\tau$ and $b_0 \sin\tau$ are easy to spot, but there is another troublemaker lurking here as well. By use of the trigonometric identity[19]

$$\cos^2\theta \, \sin\theta = \frac{1}{4}[\sin\theta + \sin 3\theta], \tag{7.55}$$

the ODE for x_1 can be rewritten:

$$\frac{d^2x_1}{d\tau^2} + x_1 = 2\omega_1 b_0 \cos\tau + \left(\frac{b_0^3}{4} - b_0\right)\sin\tau + \frac{b_0^3}{4}\sin 3\tau. \tag{7.56}$$

The $\sin 3\tau$ term is harmless; i.e., it has the particular solution $(-b_0^3/32)\sin 3\tau$, which is periodic. To avoid secular terms, we must require that

$$\omega_1 b_0 = 0 \quad \text{and} \quad b_0^3/4 - b_0 = 0.$$

Since $b_0 > 0$ by assumption, the expansion can proceed only if $\omega_1 = 0$ and $b_0 = 2$. Hence

$$x_0(\tau) = 2\cos\tau, \qquad \tau(t,\varepsilon) = [1 + \mathcal{O}(\varepsilon^2)]t,$$

and so

$$x(t,\varepsilon) = 2\cos t + \mathcal{O}(\varepsilon) \quad \text{for } t \ll 1/\varepsilon^2, \tag{7.57}$$

which is a more quantitative reformulation of (7.40).

Disappointingly, the $\mathcal{O}(\varepsilon)$-correction to the period vanishes.[20] To obtain a nonzero correction, you must complete the calculation of the $\mathcal{O}(\varepsilon)$ correction and start the calculation at the next highest order. We guide you through this task in Exercise 14.

---

[19] This identity may also be interpreted as the Fourier series expansion of $\cos^2\theta \sin\theta$.

[20] Here is an independent derivation of this fact: If $x(t)$ is a periodic solution of van der Pol's equation (7.51), then $x(-t)$ is a periodic solution of (7.51) with ε replaced by $-\varepsilon$, and it has the same period. Therefore, the period of the limit cycle must be an *even* function of ε.

Challenge: Can you predict, even without calculation, whether the $\mathcal{O}(\varepsilon^2)$-correction to the frequency will make the period longer or shorter?[21]

7.6 Van der Pol with Large β: Singular Perturbation Theory

Complementing the small-β analysis of the previous section, in the present section we describe approximately the periodic solution of van der Pol's equation

$$x'' + \beta(x^2 - 1)x' + x = 0 \tag{7.58}$$

when β is large. Again, we focus on this one specific example, but the general method—called *singular perturbation theory*—is widely applicable (cf. Chapters 1 and 2 of [43] or Chapters 6, 7, and 10 of [98]). Singular perturbations already occurred in the equation $\varepsilon x'' + x' + x = 0$ in Problem 1.14. The adjective "singular" refers to the fact that the order of the equation is reduced if the perturbing term $\varepsilon x''$ is dropped.[22] Since new phenomena appear with singular perturbations, it is wise to arm ourselves with as much preparatory information as possible.

7.6.1 Two Sources of Guidance

(a) A linear analogy

The closest linear analogue of (7.58),

$$x'' + \beta x' + x = 0, \qquad (\beta \gg 1), \tag{7.59}$$

may be interpreted as a spring–mass system with enormous friction.[23] The general solution of (7.59) is a linear combination of exponentials $e^{\lambda t}$, where

$$\lambda = \lambda_\pm = \frac{-\beta \pm \sqrt{\beta^2 - 4}}{2}.$$

In the limit of large β,

$$\lambda_+ = -1/\beta + \mathcal{O}(\beta^{-3}), \quad \lambda_- \approx -\beta + \mathcal{O}(\beta^{-1}).$$

---

[21] *Hint:* It may be useful to reflect on the information from the next section that solutions slow down as the coefficient of the frictional term tends to infinity.

[22] We also refer to a fast–slow system such as (4.36) as a singular perturbation. In that case, the *dimension* of the system is reduced if $\varepsilon = 0$. The behavior of solutions of that problem was analyzed in Exercise 5.12.

[23] We discussed such equations in the Pearls of Chapter 5. You may find it useful to review Section 5.9.1, but our discussion here is self-contained.

Formulas for exact solutions are rather messy. Suppose we seek an approximate solution of (7.59) with the ansatz

$$x(t) = C_1 e^{-t/\beta} + C_2 e^{-\beta t}.$$

Imposing initial conditions

$$x(0) = b_1, \qquad x'(0) = b_2,$$

we calculate that

$$C_1 = b_1 + \frac{b_2}{\beta} + \mathcal{O}(\beta^{-2}), \qquad C_2 = -\frac{b_2}{\beta} + \mathcal{O}(\beta^{-2}),$$

yielding the approximate solution

$$x(t) \approx \left(b_1 + \frac{b_2}{\beta}\right) e^{-t/\beta} - \left(\frac{b_2}{\beta}\right) e^{-\beta t}. \tag{7.60}$$

Comparison with the exact solution shows that this approximation is correct to $\mathcal{O}(\beta^{-2})$.

Let us note the structure of the approximation (7.60). This function starts with a rapidly decaying small-amplitude transient (represented by the second term in (7.60)), followed by slow decay (represented by the first term). Note that the larger the friction, the slower the decay in the long-lived term. Figure 7.11 compares the exact and approximate solutions of the IVP, assuming $\beta = 10$, for initial conditions

$$x(0) = 5, \qquad x'(0) = 1. \tag{7.61}$$

(b) Numerical solutions

In Figure 7.12 we graph the solution of van der Pol's equation (7.58) with the same initial conditions (7.61) and $\beta = 10$, obtained numerically. Initially, the solution resembles the solution for the linear analogue in that after a small-amplitude rapidly decaying transient (too small to be seen in the figure), there is a period of slow decay.[24] However, after this initial period, the two solutions go completely separate ways: while the linear solution decays all the way to zero, the solution of (7.58) settles into periodic behavior. In other words, the solution converges to exactly the *periodic orbit that we seek*. Moreover, in this periodic solution, intervals of slow evolution are punctuated by intervals of rapid, almost jump-like, evolution. In the literature such oscillations are termed *relaxation* oscillations. As in an earthquake, stress builds up slowly until it reaches a breaking point and relaxes precipitously.

[24]The decay for van der Pol's equation is *much slower* than for the linear equation. For these initial conditions, the effective coefficient of friction in van der Pol's equation, $\beta(x^2-1)$, is enhanced by a large factor, initially almost 25.

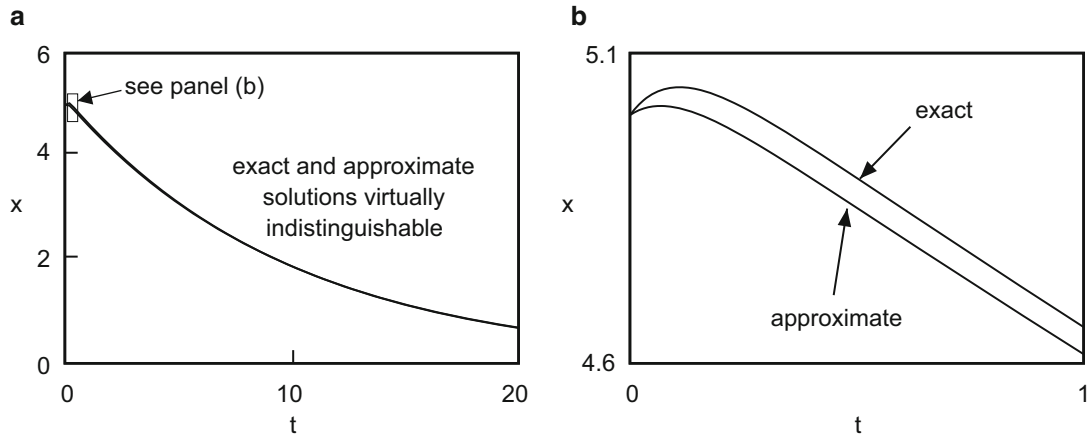

Figure 7.11: *Exact and approximate solutions of linear analogue* (7.59) *with* $\beta = 10$ *and initial conditions* (7.61). *With the viewing scale of Panel (a), the two curves are almost indistinguishable. Panel (b) shows a blowup of the solutions at small times.*

7.6.2 Approximation of the Initial Decay in (7.58)

The initial slow decay in Figure 7.12 can be described with the fast-equations-to-equilibrium approximation in an appropriately scaled version[25] of (7.58). Specifically, consider $\tau = t/T$, where $T = \beta^p$ for some power p; if $p = 1$, we obtain the equation

$$\frac{1}{\beta^2}\frac{d^2x}{d\tau^2} + (x^2 - 1)\frac{dx}{d\tau} + x = 0. \tag{7.62}$$

Neglecting the second-order term gives the reduced problem

$$(x^2 - 1)\frac{dx}{d\tau} + x = 0, \qquad x(0) = 5, \tag{7.63}$$

where for a first-order equation we impose only one initial condition.

[25]For pedagogical reasons let's explore the consequences of not scaling and ignoring the principle of dominant balance from Section 5.9.1; i.e., let us attempt to derive an approximate equation from (7.58) by dropping the first and third terms on the grounds that the middle term, which has the large coefficient, should dominate. This leads to a completely sterile equation, $(x^2-1)x' = 0$, which has only trivial constant solutions.

Recall that the principle of dominant balance tells us to scale the equation so that two terms have comparable magnitude and dominate the third. The scaling $\tau = t/\beta$ yields (7.62), which has this structure. The scaling $\tau = \beta t$ would balance two different terms; although our discussion doesn't employ it, it is used in deriving the results quoted in Section 7.6.4.

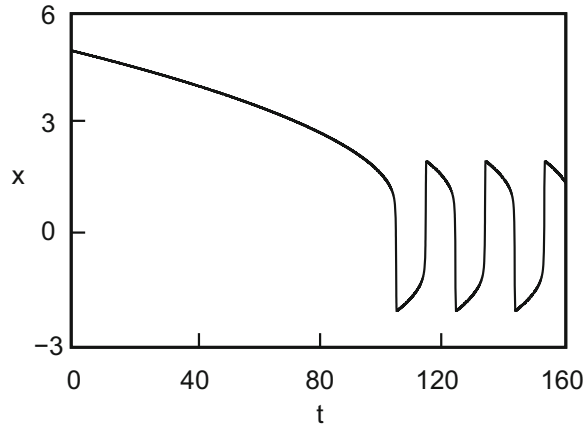

Figure 7.12: *Numerical solution of the IVP* (7.58), (7.61) *for van der Pol's equation with* $\beta = 10$.

By separability, the general solution of the ODE in (7.63) is

$$\frac{x^2}{2} - \ln|x| = C - \tau \tag{7.64}$$

for an arbitrary constant C. Although we can't solve (7.64) for x as a function of τ, we may understand the solution graphically. In Figure 7.13 we have plotted the LHS of (7.64). To satisfy the initial condition we choose $C = C_0$ so that the horizontal line $\{x = C_0\}$ intersects this curve at $x = 5$, i.e., $C_0 = 25/2 - \ln 5$. At later times, $x(\tau)$ is the value of x where the curve intersects a horizontal line of height $C_0 - \tau$; thus, $x(\tau)$ gradually decays. As Figure 7.14 shows, this decay of the approximate solution closely matches the initial decay of the solution of the full problem.[26]

However, this construction eventually fails. The function $x^2/2 - \ln|x|$ has a minimum at $x = 1$, where it equals $1/2$. Thus if $\tau > C_0 - 1/2$, the horizontal line $\{x = C_0 - \tau\}$ no longer intersects the curve in Figure 7.13. In other words, we are encountering a new phenomenon here: although the fast-equation-to-equilibrium approximation works well for a while, *it (correctly) predicts its own breakdown.* This behavior serves as a reminder to be cautious when making such a drastic approximation.

[26]In comparing Figures 7.12 and 7.14, keep in mind that scaled time $\tau = t/\beta$ is plotted in the latter figure.

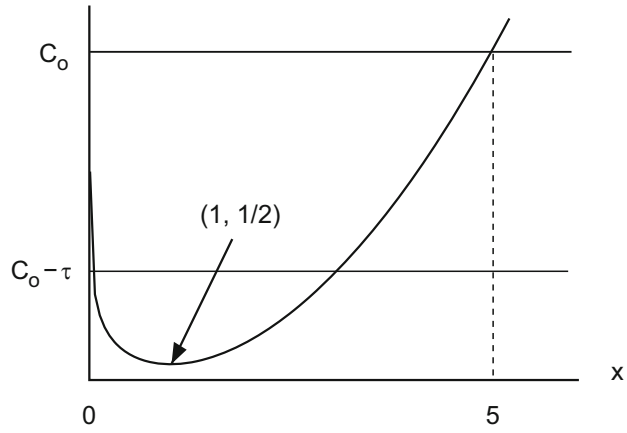

Figure 7.13: *Graph of* (7.64)*, the implicit solution of the reduced van der Pol equation* (7.63)*.*

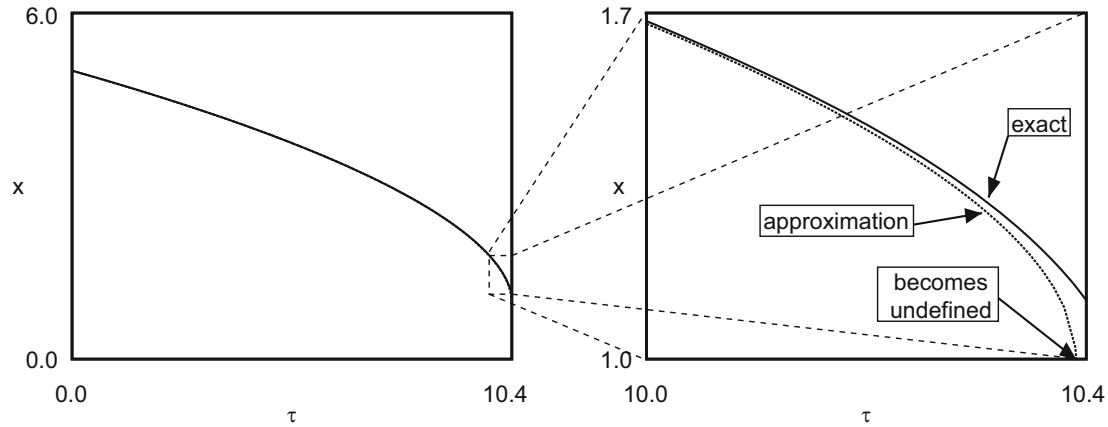

Figure 7.14: *Comparison of the exact solution of the IVP* (7.62) *for van der Pol's equation with the asymptotic approximation* (7.64)*, which solves the reduced equation* (7.63)*. For early times (Panel (a)), the two functions are indistinguishable. Again,* $\beta = 10$*.*

7.6.3 Phase-Plane Analysis of a Related Equation

The solution of (7.62) beyond this breakdown can be better understood through study of a related ODE,

$$\frac{1}{\beta^2}\frac{d^2z}{d\tau^2} + \left[\frac{1}{3}\left(\frac{dz}{d\tau}\right)^3 - \frac{dz}{d\tau}\right] + z = 0, \tag{7.65}$$

a scaled version of Rayleigh's equation. You can easily check that if $z(\tau)$ satisfies (7.65), then $x(\tau) = dz/d\tau(\tau)$ satisfies the ODE in (7.62). Conversely, if $x(\tau)$ satisfies the ODE in (7.62), then for an appropriate constant C,

$$z(\tau) = \int_0^\tau x(\tau')\,d\tau' + C$$

satisfies (7.65).

We want to describe periodic solutions[27] of these ODEs. The problem with (7.62) is that if $\beta \to \infty$, the slope $dx/d\tau$ tends to infinity during the relaxation part of the oscillations. By contrast, as we shall see, in the periodic solution of (7.65), both z and $dz/d\tau$ remain bounded as $\beta \to \infty$.

The geometry of the phase plane clarifies the behavior of (7.65). In the usual way, (7.65) can be rewritten as a first-order system,

$$\begin{array}{lrcl} \text{(a)} & dz/d\tau & = & x, \\ \text{(b)} & \varepsilon\, dx/d\tau & = & -z - (x^3/3 - x), \end{array} \tag{7.66}$$

where $\varepsilon = \beta^{-2}$ and we have retained the notation $x = dz/d\tau$ from the derivation of (7.65). Making the fast-equation-to-equilibrium approximation yields

$$\begin{array}{lrcl} \text{(a)} & dz/d\tau & = & x, \\ \text{(b)} & z + (x^3/3 - x) & = & 0. \end{array} \tag{7.67}$$

For orientation, let's reinterpret our approximate solution of (7.62) in the previous section in terms of (7.67). To give the discussion a touch of whimsy, imagine an ant, as depicted in Figure 7.15(a), crawling along the nullcline (7.67b), starting at $x = 5$ and adjusting its vertical speed to the current value of x. All goes well until it reaches $x = 1$, when the rules break down—it cannot continue to climb without leaving the nullcline. Of course, this is just the breakdown we saw in (7.63) in another guise.

To understand what happens after the reduced equations break down, we return to the full problem (7.66). The vector field associated with (7.66) is sketched in Figure 7.15(b), where double arrowheads indicate the fast flow in the x-direction

[27] In Exercise 3 we ask you to use the Poincaré–Bendixson theorem to prove that (7.65) has a periodic solution.

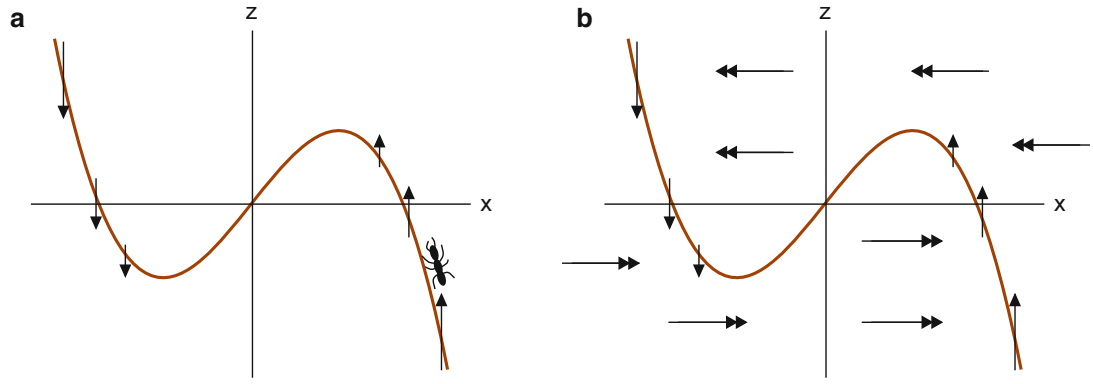

Figure 7.15: *(a) An animation of the reduced system* (7.67)*. The cubic has a local maximum at* $x = 1$ *at height* $z = 2/3$*. (b) The vector field defined by the fast–slow system* (7.66)*.*

that occurs at points off the nullcline. Enlarging on our whimsical imagery, we may interpret (7.66b) as a strong wind that sweeps the ant along whenever it (the ant) cannot get a foothold on the nullcline. With this interpretation it is easy to follow the solution after the breakdown—the ant climbs slightly above the peak of the nullcline at $(x, z) = (1, 2/3)$ and is blown quickly to a different portion of the nullcline. This prediction matches qualitatively the behavior seen in Figure 7.12 after the initial transient, and we claim that it even matches quantitatively. Since the horizontal motion to the left is very rapid, the value of z hardly changes during the flight; thus $z \approx 2/3$ where the ant lands on the nullcline; we compute from (7.67b) that $x = -2$ when $z = 2/3$; and in Figure 7.12 we see that $x \approx -2$ when rapid evolution stops.

The subsequent relaxation oscillations in Figure 7.12 also emerge from these ideas. After the ant reaches $(-2, 2/3)$, it can once again get a foothold on the nullcline and follow (7.67b). Since $x < 0$, the ant now crawls downward on the nullcline till it reaches $x = -1$. Here the approximation (7.67) again breaks down, and the ant is swept over to the right branch of the nullcline at approximately $(2, -2/3)$. Then it climbs upward, retracing part of its original journey, etc.

In conclusion, this discussion has led us to an approximate description of a periodic solution[28] of (7.66) in four phases (see Figure 7.16a):

- *Phase 1:* A piece that follows the x-nullcline upward from $(2, -2/3)$ to the local maximum of the nullcline at $(1, 2/3)$. Here the speed is $\mathcal{O}(1)$.

[28]By differentiating and undoing scaling, we may convert this into a description of a periodic solution of (7.58). *Question:* Does the maximum slope of the periodic solution of (7.58) tend to infinity as $\beta \to \infty$?

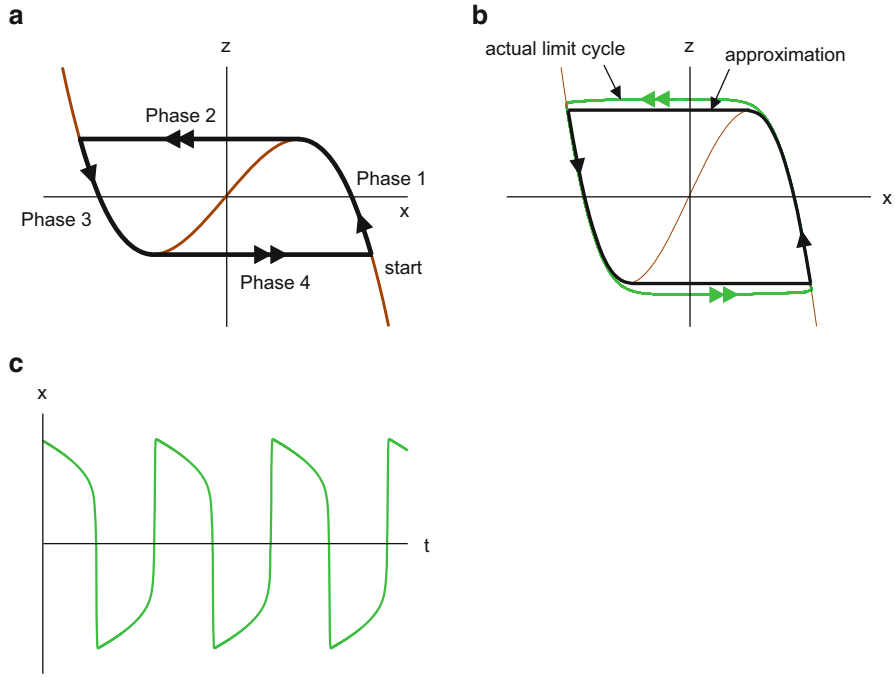

Figure 7.16: *(a) Illustrating the four phases in the approximate description of the limit cycle of the van der Pol equation for large β. (b) Comparison of the approximate solution (black curve) and numerical solution (green curve) of (7.66) for $\beta = 10$. The vertical scale has been stretched relative to that of Panel (a) in order to make these curves more distinguishable. (c) Corresponding relaxation oscillations of $x(t)$. For $x(t)$, the asymptotic approximation is indistinguishable from the numerical solution.*

- *Phase 2:* A horizontal piece from $(1, 2/3)$, intersecting the x-nullcline at $(-2, 2/3)$. Here the speed is $\mathcal{O}(\varepsilon^{-1})$.
- *Phase 3:* A piece that follows the x-nullcline downward from $(-2, 2/3)$ to the local minimum of the nullcline at $(-1, -2/3)$. Here the speed is $\mathcal{O}(1)$.
- *Phase 4:* A horizontal piece from $(-1, -2/3)$, returning to the nullcline at $(2, -2/3)$. Here the speed is $\mathcal{O}(\varepsilon^{-1})$.

Panels (b) and (c) of Figure 7.16 compare this approximate solution with a numerical solution of (7.66). Can you believe that such a simplistic argument gives such a good approximation of complicated behavior? As we said before, we continue to be awed by the power of asymptotics.

7.6.4 Concluding Remarks

Although our discussion above has been purely heuristic, with effort this analysis can be made completely rigorous. The pioneering work was done by Fenichel; [46] gives a comprehensive introduction to the theory, including references.

Let us estimate the period of the oscillations. Most of the time required for the solution of (7.66) to complete the cycle around the origin is spent in Phases 1 and 3. Recalling the solution (7.64) of the reduced equation, we calculate that while x decreases from $x = 2$ to $x = 1$ along the nullcline, τ increases by $3/2 - \ln 2$. Thus, Phase 1 lasts approximately this long, and by symmetry, Phase 3 has the same duration. Doubling this estimate and undoing the scaling $\tau = t/\beta$, we obtain the approximation

$$T(\beta) \approx (3 - 2\ln 2)\beta \tag{7.68}$$

for the period of the relaxation oscillations of (7.58). In particular, the period gets large as $\beta \to \infty$. This is to be expected—the first-order term in (7.58) involves friction, and as friction gets large, all motion slows down. We invite you to check (7.68) numerically (as well as its refinement (7.69) below).

With a tour de force application of higher-order asymptotics,[29] it has been shown that

$$T(\beta) = (3 - 2\ln 2)\beta + 3\zeta_1\beta^{-1/3} + \mathcal{O}(\beta^{-1}\ln\beta), \tag{7.69}$$

where the constant $\zeta_1 \approx 2.33811$ is the first zero of the Airy function $Ai(\zeta)$. The $\mathcal{O}(\beta^{-1/3})$-term is perhaps surprising, since it seems like the neglected durations of Phases 2 and 4 are only $\mathcal{O}(\beta^{-1})$ (in the original, unscaled, time) as $\beta \to \infty$. The subtle point is that a substantial amount of time is required to make the transitions from Phase 1 to Phase 2 and from Phase 3 to Phase 4, i.e., the transitions at which the solution is pushed off the nullcline. Both insight and patience are required to derive (7.69); and with sloppy analysis you might easily miss the $\mathcal{O}(\beta^{-1/3})$-term altogether.

7.7 Stability of the van der Pol Limit Cycles

In this section we prove that the van der Pol limit cycles constructed above are stable. The proofs are fairly short but uninspiring. We include them for completeness, expecting that most readers will omit them on the first pass through the book.

[29] In regular perturbation theory, such as in Section 7.5, higher-order calculations are routine, although often tedious. By contrast, the derivation of (7.69) is anything but routine. The fractional powers and logarithmic terms in (7.69), which are common in singular-perturbation expansions, warn of hidden complexities. The derivation of this result is based on a sophisticated application of *matched asymptotic expansions*. Exercise 5.12 gives a simple illustration of this technique; the general method is developed in [98] and in [43]; and the specific formula (7.69) is derived in [13].

Incidentally, some authors give a formula for $T(\beta)$ with a different coefficient for $\beta^{-1/3}$; we assure you that equation (7.69) is correct.

7.7.1 Case 1: Small β

We will calculate an approximate Poincaré map with perturbation theory. Superficially, the calculation resembles the Poincaré–Lindstedt method in Section 7.5, but there is a crucial difference: the solution of the IVP for the first-order correction *will* have secular terms. Indeed, the secular terms are precisely what make generic solutions converge to the periodic solution as $t \to \infty$.

As in Section 7.5, we work with the second-order scalar equation rather than a first-order system, adapting the Poincaré map accordingly. We constructed an approximate periodic solution $\boldsymbol{\gamma}(t)$ of the van der Pol equation with initial conditions $\boldsymbol{\gamma}(0) \approx 2$ and $\boldsymbol{\gamma}'(0) = 0$. Given an initial value b close to $\boldsymbol{\gamma}(0)$, let $x(t, b, \varepsilon)$ be the solution of the zero-initial-velocity IVP

$$\frac{d^2x}{dt^2} + \varepsilon(x^2 - 1)\frac{dx}{dt} + x = 0, \qquad x(0) = b, \ \frac{dx}{dt}(0) = 0. \tag{7.70}$$

Using the time derivative of this solution, define a return time[30] $\tau(b, \varepsilon)$ that satisfies

$$\frac{dx}{dt}(\tau(b, \varepsilon), b, \varepsilon) = 0 \tag{7.71}$$

with $\tau(b, \varepsilon) \approx 2\pi$. Then the Poincaré map is given by

$$\Pi(b) = x(\tau(b, \varepsilon), b, \varepsilon), \tag{7.72}$$

where our notation suppresses the dependence of Π on ε.

We look for an expansion of $x(t, b, \varepsilon)$ with the usual form

$$x(t, b, \varepsilon) = x_0(t, b) + \varepsilon x_1(t, b) + \dots. \tag{7.73}$$

Substituting the series into the equation, we obtain ODEs

$$\begin{aligned} \mathcal{O}(\varepsilon^0): \quad & \frac{d^2x_0}{dt^2} + x_0 = 0, \\ \mathcal{O}(\varepsilon^1): \quad & \frac{d^2x_1}{dt^2} + x_1 = -(x_0^2 - 1)\frac{dx_0}{dt}, \end{aligned}$$

subject to initial conditions

$$x_0(0, b) = b, \ \frac{dx_0}{dt}(0, b) = 0; \qquad x_1(0, b) = \frac{dx_1}{dt}(0, b) = 0.$$

[30] This τ has no connection to the scaled time in the Poincaré–Lindstedt method. We do not scale time in the present calculation.

The solution of the leading-order IVP is

$$x_0(t,b) = b\cos t.$$

Given this we calculate, as in Section 7.5, that

$$-(x_0^2-1)\frac{dx_0}{dt} = \left(\frac{b^3}{4} - b\right)\sin t + \frac{b^3}{4}\sin 3t,$$

and the solution of the first-order correction IVP is

$$x_1(t,b) = \left(\frac{b^3}{4} - b\right)\left[-\frac{t}{2}\cos t + \frac{1}{2}\sin t\right] - \frac{b^3}{32}\left[\sin 3t - 3\sin t\right].$$

Substitution into (7.71) gives the equation

$$-b\sin\tau(b,\varepsilon) + \mathcal{O}(\varepsilon) = 0,$$

from which we conclude that $\tau(b,\varepsilon) = 2\pi + \mathcal{O}(\varepsilon)$. Note that $x_0(\tau(b,\varepsilon),b) = x_0(2\pi,b) + \mathcal{O}(\varepsilon^2)$, because $dx_0/dt(2\pi,b) = 0$. Thus, substitution into (7.72) yields

$$\Pi(b) = x_0(2\pi,b) + \varepsilon x_1(2\pi,b) + \mathcal{O}(\varepsilon^2) = b - \pi\varepsilon\left(\frac{b^3}{4} - b\right) + \mathcal{O}(\varepsilon^2). \tag{7.74}$$

We differentiate and evaluate Π' at $b = \boldsymbol{\gamma}(0)$, the starting point of the periodic trajectory. Since $\boldsymbol{\gamma}(0) = 2 + \mathcal{O}(\varepsilon)$, we obtain $\Pi'(\boldsymbol{\gamma}(0)) = 1 - 2\pi\varepsilon + \mathcal{O}(\varepsilon^2)$. In particular, $\Pi'(\boldsymbol{\gamma}(0)) < 1$, provided ε is sufficiently small, so the limit cycle is asymptotically stable.

Remarks: (i) We reject the solution of $dx/dt(t,b,\varepsilon) = 0$ with $t \approx \pi$. At this time, the trajectory in the (x,x')-phase plane crosses the negative x-axis, which is the kind of premature crossing of Σ illustrated in Figure 7.8. (ii) Let's use (7.74) to look for fixed points of Π. Neglecting the $\mathcal{O}(\varepsilon^2)$-term, we see that $\Pi(b) = b$ if $b = 2$. This reproduces the Poincaré–Lindstedt result of Section 7.5 that the periodic solution, to lowest order, is a circle of radius 2 in the x, x'-phase space.

7.7.2 Case 2: Large β

When β is large, (7.66) is a convenient reduction of the van der Pol equation to a first-order system. In Section 7.6, we argued that this system has a periodic orbit Γ, as sketched in Figure 7.16. To determine the stability of this orbit, consider the Poincaré section

$$\Sigma = \{(x,z) : x = 1,\ -1 < z < 0\}.$$

The orbit Γ crosses Σ in "Phase 4" of the asymptotic description of the orbit, at the base point $(1, b_*)$, where $b_* \approx -2/3$. Given a nearby point $(1,b) \in \Sigma$, to evaluate the

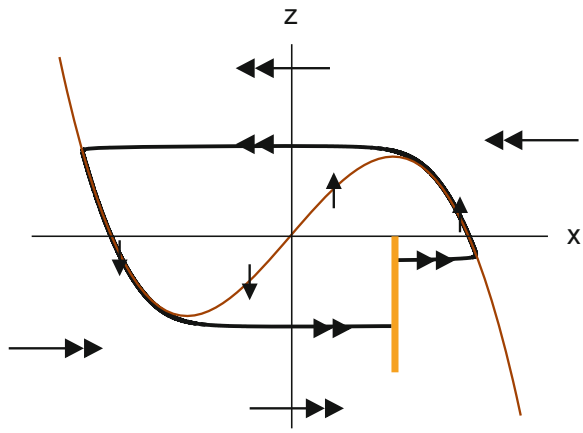

Figure 7.17: *The Poincaré map for the periodic solution of* (7.66)*, with* $\varepsilon = 1/100$*, together with the* x*-nullcline. The Poincaré section* $\{(1, b) : -1 < b < 0\}$ *is shown in bold orange. One trajectory, starting from the initial conditions* $(x(0), y(0)) = (1, -0.2)$*, is illustrated.*

Poincaré map $\Pi(b)$ we must solve the IVP for (7.66) with initial conditions

$$x(0) = 1, \qquad z(0) = b.$$

This IVP may by analyzed with the same asymptotic argument as was used in constructing Γ. (See Figure 7.17.) Indeed, after racing over to the nullcline (7.67b), the solution closely follows Γ and returns to cross Σ very near to the base point $(1, b_*)$. Thus for all b we have $\Pi(b) \approx b_*$. Indeed, there is great compression, and $|\Pi'(b_*)| \ll 1$. Hence the orbit is stable.

7.8 Exercises

After the core exercises, there are subsections on the Poincaré–Lindstedt method and changing coordinates in an ODE.

7.8.1 Core Exercises

The core exercises address the following issues:

Unfinished business	1
Existence of periodic solutions	2–4
Nonexistence of periodic solutions	5
ω-limits of trajectories	6, 7
The Poincaré–Bendixson theorem	8
A combination of techniques	9
Eigenvalues of a Poincaré map	10, 11
Discrete dynamical systems	12, 13

1. (a) Derive Theorem 7.2.1, the simpler version of the Poincaré–Bendixson theorem, from Theorem 7.2.5.

   Remark: This exercise is easy, *provided* you get started along right road—choose any point $\mathbf{b} \in \mathcal{U}$ and consider the trajectory that starts at $\mathbf{b}$.

   (b) Prove Theorem 7.2.2, the Poincaré–Bendixson theorem generalized to ODEs on the cylinder.

   Hint: Using coordinates (θ, z) on $S^1 \times \mathbb{R}$, you may identify the cylinder with $\mathbb{R}^2 \sim \{\mathbf{0}\}$ via the mapping

   $$(\theta, z) \mapsto (e^z \cos\theta, e^z \sin\theta). \tag{7.75}$$

   Using this map, transform from an ODE on the cylinder

   $$\begin{bmatrix} \theta' \\ z' \end{bmatrix} = \mathbf{F}(\theta, z)$$

   to one on the plane, and apply the Poincaré–Bendixson theorem to the latter.

   Remark: As follows from Exercise 7, a Poincaré–Bendixson-type result does *not* hold for the torus $\mathbb{T}^2$.

2. Solve the ODE on $\mathbb{R}^2 \sim \{\mathbf{0}\}$

   $$\begin{aligned} r' &= 1 + \varepsilon \sin\theta - r, \\ \theta' &= 1, \end{aligned}$$

   to show that if $|\varepsilon| < \sqrt{2}$, it has a periodic solution.

   Remark: The condition on ε is needed so that r remains positive.

3. Construct an appropriate trapping region for and apply the Poincaré–Bendixson theorem to (7.66), thereby showing that (7.65) has a periodic solution.

   Remark: The usual version of Rayleigh's equation is

   $$x'' + \beta \left[\frac{(x')^3}{3} - x' \right] + x = 0;$$

   (7.65) resulted from a special scaling appropriate for large β.

4. Show that the torqued-pendulum equation (7.12) has a periodic solution if $\beta < \mu/2$.

   Hint: Mimicking (7.14), consider a region of the form

   $$\mathcal{K} = \{(x, y) \in S^1 \times (0, \infty) : \ 1 + \varepsilon < y^2/2 - \cos x \leq E_0\}.$$

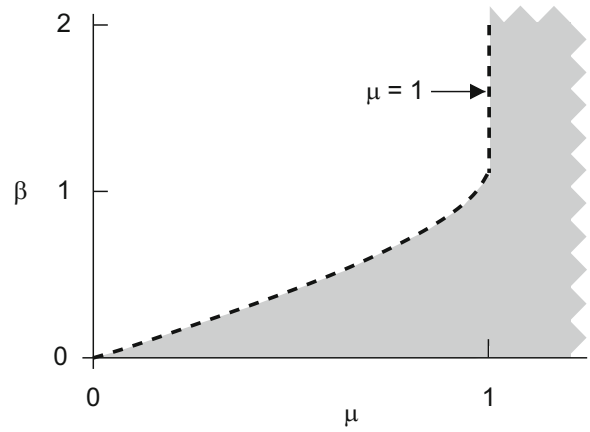

Figure 7.18: *The shaded area indicates the region of parameter space for which (7.12) has a periodic solution. The boundary of the region becomes vertical at $\mu = 1$, $\beta \approx 1.19$. For parameters in the shaded region with $\mu < 1$, (7.12) has bistable behavior—both a stable equilibrium and a stable periodic solution are possible ω-limits.*

Show that if $E_0 \geq (\mu/\beta)^2/2 + 1$ and $\varepsilon > 0$ is sufficiently small, then $\mathcal{K}$ is an equilibrium-free trapping region. Apply Theorem 7.2.2.

Discussion: Figure 7.18 displays the region in parameter space in which (7.12) has a periodic solution. Reflecting our result from Section 7.2.2(b), this region includes the set $\{\mu > 1\}$, but it also includes the wedge $\{\beta < \mu/2\}$.

5. (a) Show that a scalar autonomous ODE $x' = f(x)$ cannot have periodic solutions.

Hint: Note that the equation $x' = f(x)$ is posed on the line $\mathbb{R}^1$, *not* the circle—on S^1, even the trivial equation $x' = 1$ has periodic solutions. Suppose indirectly that there exists a periodic solution of $x' = f(x)$, and let T denote its period. Multiply both sides of the ODE by dx/dt, integrate over an interval of length T, and derive a contradiction.

(b) Given an ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, show that if there is a function $L : \mathcal{U} \to \mathbb{R}$ such that

$$\text{for all } \mathbf{x} \in \mathcal{U} \quad \langle \nabla L(\mathbf{x}), \mathbf{F}(\mathbf{x}) \rangle < 0,$$

then this equation cannot have a periodic solution in $\mathcal{U}$.

(c) *Introduction:* The system

$$x' = x - \frac{xy}{1+Sx} \qquad y' = \rho\left(\frac{xy}{1+Sx} - y\right)$$

is another modification of the Lotka–Volterra system: consumption saturates when the prey are plentiful. It is the limit of the Rosenzweig–MacArthur model in Exercise 5.10 as the carrying capacity for the prey tends to infinity.

Apply Dulac's theorem with $g(x, y) = 1/xy$ to show that this system has no periodic orbits in the first quadrant.

6. Assuming that condition (7.7) holds, determine the ω-limit sets for all points $\mathbf{b} \in \mathbb{T}^2$ under the flow of (7.6).

Remark: If (7.7) does not hold, then ω-limit sets may vary with the parameters in complicated ways (cf. Section 10.4).

7. *Introduction:* By *linear flow on the torus*, we mean the system

$$\begin{aligned} \theta_1' &= \omega_1, \\ \theta_2' &= \omega_2, \end{aligned} \tag{7.76}$$

where θ_1 and θ_2 are reduced modulo 2π.

(a) Show that if ω_1/ω_2 in (7.76) is rational, then (i) every orbit is periodic and (ii) for every initial condition, the omega limit set $\omega(\mathbf{b})$ is the periodic orbit on which $\mathbf{b}$ lies.

(b) Show that if ω_1/ω_2 in (7.76) is irrational, then (i) there are no closed orbits; (ii) every orbit $\boldsymbol{\varphi}(t, \mathbf{b})$ is dense in the torus; and (iii) for every initial condition, $\omega(\mathbf{b})$ is the entire torus.

Hint: Consider the flow as a Poincaré-like map with respect to the Poincaré section

$$\Sigma = \{(\theta_1, \theta_2) \in \mathbb{T}^2 : \theta_1 = b_1 \pmod{2\pi}\}.$$

The orbit $\boldsymbol{\varphi}(t, \mathbf{b})$ passes through Σ at times $t = 2\pi n/\omega_1$, where n is an integer. To prove Claim (ii), it suffices to show that $\{\boldsymbol{\varphi}(2\pi n/\omega_1, \mathbf{b}) : n = 1, 2, \ldots\}$ is dense in Σ. *(Why?)* Given a desired level of accuracy $\varepsilon > 0$, choose $N > 2\pi/\varepsilon$ and divide Σ into N "boxes"

$$\Sigma_k = \left\{ (\theta_1, \theta_2) \in \mathbb{T}^2 : \theta_1 = b_1,\ \frac{k-1}{N} \le \frac{\theta_2}{2\pi} < \frac{k}{N} \pmod{2\pi} \right\},$$

where $k = 1, 2, \ldots, N$. Now the set $\{\boldsymbol{\varphi}(2\pi n/\omega_1, \mathbf{b}), \quad n = 0, 1, \ldots, N\}$ contains $N + 1$ points that belong to Σ, so at least two of them (say n_1, n_2 where $n_1 > n_2$) must lie in the same box (see Figure 7.19(a)); this means that the distance between them is less than ε, or the second component φ_2 of $\boldsymbol{\varphi}$ satisfies

$$0 < |\varphi_2(2\pi n_1/\omega_1, \mathbf{b}) - \varphi_2(2\pi n_2/\omega_1, \mathbf{b})| < \varepsilon,$$

the difference being reduced mod 2π. Thus, if $\Delta n = n_1 - n_2$, the subsequence $\boldsymbol{\varphi}(2\pi\ell\Delta n/\omega_1, \mathbf{b}), \quad \ell = 0, 1, \ldots$ starts at (b_1, b_2) and marches along Σ in uniform steps that are smaller than ε, so it eventually comes within ε of every point in Σ (Figure 7.19(b)).

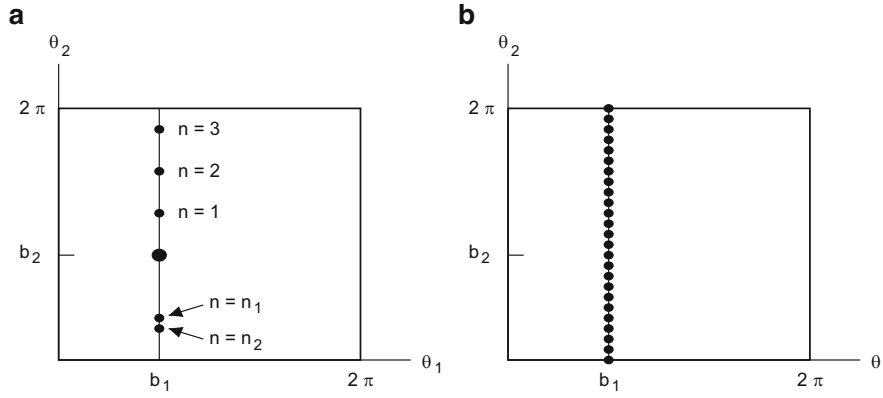

Figure 7.19: *(a) Some points in the set* $\{\boldsymbol{\varphi}(2\pi n/\omega_1, \mathbf{b}), \quad n = 0, 1, \ldots, N\}$ *(see Exercise 7). (b) Some points in the set* $\{\boldsymbol{\varphi}(2\pi\ell\Delta n/\omega_1, \mathbf{b}), \quad \ell = 0, 1, \ldots\}$.

8. For equation (7.23), the Lotka–Volterra system plus the Allee effect, show that for generic initial conditions in the first quadrant, the solution converges to extinction. Identify the exceptional initial conditions.

   Hint: Strictly speaking, you should begin by constructing trapping regions to derive global existence. This step is not trivial, because for small ε, trajectories tend to hug a periodic solution of the Lotka–Volterra equations; thus, a simple polygonal region is too crude to trap solutions.

   You can convince yourself that generically, solutions tend to extinction by locating equilibria and stable/unstable manifolds, using a generous dose of *what-else-can-it-be* logic. Invoke the strong version of the Poincaré–Bendixson theorem to make a rigorous argument.

9. *Introduction:* In preparation for a bifurcation example in Chapter 9, consider the effects of perturbing the van der Pol system by adding a nonlinear term to the restoring force, say

$$\begin{aligned} x' &= y, \\ y' &= -\beta(x^2 - 1)y - x - \varepsilon x^2. \end{aligned} \tag{7.77}$$

   (a) Apply the Poincaré–Bendixson theorem to deduce that provided $|\varepsilon|$ is sufficiently small, (7.77) still has a periodic solution.

   (b) Show that this system has a saddle point at $(-1/\varepsilon, 0)$.

   (c) Show that if $\varepsilon > 1$, then one half of the global stable manifold through the saddle point connects (as $t \to -\infty$) to the equilibrium at the origin. *(Draw this!)*

   Hint: Without the frictional term $-\beta(x^2 - 1)y$, equation (7.77) would be Hamiltonian with the Hamiltonian

$$L(x, y) = x^2/2 + y^2/2 + \varepsilon x^3/3.$$

The frictional term causes $L(x,y)$ to *increase* in the strip $|x| < 1$, which for $\varepsilon > 1$ contains both the saddle point and the origin. To derive the result, use $L(x,y)$ as a Lyapunov function for the *backward* equation (cf. Exercise 6.12(a)).

(d) Argue that (7.77) has no periodic solutions if $\varepsilon > 1$.

Hint: The stable manifold in Part (c) blocks the periodic orbit of (7.77) that for small ε encircles the origin, and you could use this fact to solve the problem. However, an easier and more elegant approach, which we discuss along with the necessary background in the problem solutions on our website, is to invoke *index theory*. Hey, take the easy way out!

(e) Check your conclusions in Parts (c) and (d) by computer simulations.

10. *Introduction:* Consider a spring–mass system subjected to a trigonometric periodic force, say $\Gamma \cos \omega t$. After scaling, we may describe this situation by the three-dimensional first-order autonomous system

$$\begin{aligned} x' &= y, \\ y' &= -x - \beta y + \Gamma \cos z, \\ z' &= \omega, \end{aligned} \tag{7.78}$$

where ω, β, Γ are real constants with $\beta > 0$. Because the RHS is periodic in z, we may regard (7.78) as an ODE on $\mathbb{R}^2 \times S^1$, where $S^1 = \mathbb{R}/2\pi\mathbb{Z}$. In this problem you identify a periodic orbit of the system, you find its Poincaré map, and you use this map to show that the periodic orbit is asymptotically stable.

(a) Show that (7.78) has a $2\pi/\omega$-periodic solution, say $x_p(t), y_p(t), z_p(t)$.

Hint: In looking for explicit solutions of (7.78), it is convenient to recall the formulation as a scalar equation,

$$x'' + \beta x' + x = \Gamma \cos \omega t. \tag{7.79}$$

Construct a periodic solution of (7.78) with $z_p(t) = \omega t$ from the particular solution of (7.79) discussed in Problem 1.13(b).

(b) Relative to the transverse section

$$\Sigma = \{(x, y, z) \in \mathbb{R}^2 \times (\mathbb{R}/2\pi\mathbb{Z}) : z = 0 \pmod{2\pi}\},$$

which we identify with $\mathbb{R}^2$, show that there is a 2×2 matrix such that the Poincaré map $\mathbf{\Pi} : \mathbb{R}^2 \to \mathbb{R}^2$ of this periodic solution has the form[31]

$$\mathbf{\Pi}(\mathbf{b}) = A(\mathbf{b} - \mathbf{b}_0) + \mathbf{b}_0,$$

where $\mathbf{b}_0 = (x_p(0), y_p(0))$.

[31] This map is an example of what is called an *affine* map: more precisely, it equals the sum of a linear map (i.e., $\mathbf{b} \mapsto A\mathbf{b}$) and a constant term (i.e., $\mathbf{b}_0 - A\mathbf{b}_0$).

Hint: The matrix A relates to homogeneous solutions of(7.79), but for this part of the problem you need not determine the precise matrix in this formula—the form of $\mathbf{\Pi}$ results from the linearity of (7.79). Interpret the fact that $\mathbf{b}_0$ is a fixed point of $\mathbf{\Pi}$.

(c) Show that the eigenvalues of $\mathbf{D\Pi}$ satisfy $|\lambda_j(\mathbf{D\Pi})| < 1,\ j = 1, 2$, so the periodic solution is asymptotically stable.

Hint: Prove that if $e^{\mu t}$ is a solution of the homogeneous equation $x'' + \beta x' + x = 0$, then $e^{2\pi\mu/\omega}$ is an eigenvalue of $\mathbf{D\Pi}$.

11. *Introduction:* Consider the system of ODEs

$$r' = z, \quad \theta' = re^z, \quad z' = r - 1. \tag{7.80}$$

You may think of it either as an equation on $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 > 0\}$ written in cylindrical coordinates or as an equation on $\mathbb{R} \times (\mathbb{R}/2\pi\mathbb{Z}) \times \mathbb{R}$. With respect to the latter coordinates, the system has the 2π-periodic solution $\boldsymbol{\gamma}(t) = (1, t, 0)$. Define a Poincaré map with respect to a Poincaré section Σ contained in the plane $\{\theta = 0\}$.

Find the eigenvalues of $\mathbf{D\Pi}$ to assess the stability of this periodic solution.

Hint: Because the θ-equation is so messy, the time for a trajectory starting in Σ to return to Σ—the solution of (7.28)—depends on the initial condition $\mathbf{b}$ in a totally impenetrable way. Differentiating (7.29), including the dependence on $\tau(\mathbf{b})$, to find $\mathbf{D\Pi}$ directly would discourage even the most stalwart investigator. Fortunately, according Proposition 7.3.5, you need only calculate the eigenvalues of $\mathbf{D}\boldsymbol{\varphi}(2\pi, \boldsymbol{\gamma}(0))$, which is quite benign.

12. *Introduction:* This exercise offers routine practice with a discrete dynamical system, defined by

$$\begin{aligned} x_{n+1} &= x_n - y_n + x_n y_n \\ y_{n+1} &= x_n + (1 - a)y_n + y_n^2 \end{aligned}$$

where a is a real parameter.

(a) For all a, the origin is a fixed point of this mapping. Apply Proposition 7.8.1 to determine for what range of a the fixed point is stable.

Hint: If you want to look ahead, the result of Exercise 25 would be helpful here.

(b) There are also fixed points in the plane $\{x = 1\}$. Find these, make a graph of y vs. a.

13. (a) Prove Theorem 7.3.7.

Remark: As in Proposition 6.2.1, the instability part of the result seems obvious. Although the proof is not completely trivial, it may not be worth bothering with this issue a second time.

(b) Show that the fixed point $\boldsymbol{\gamma}(0)$ of the Poincaré map of a periodic solution of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ is Lyapunov stable, asymptotically stable, or unstable iff the trajectory $\boldsymbol{\gamma}(t)$ has the corresponding stability behavior with respect to the ODE.

7.8.2 The Poincaré–Lindstedt Method

The exercises in this section show that with the Poincaré–Lindstedt method you can get an analytical hold on many interesting, but otherwise intractable, problems involving periodic solutions, both old and new, provided they can be cast in the appropriate perturbation-theory language. (See also Section 7.10.2.) The calculations are longish, but we hope you feel that the interest of the results justifies such efforts, and we give you copious hints along the way.

14. (a) Show that given $\omega_1 = 0$ and $b_0 = 2$, the solution of (7.56) subject to initial conditions (7.54) is

$$x_1(\tau) = -\frac{1}{4}(\sin 3\tau - 3\sin\tau) + b_1 \cos\tau,$$

where b_1 is the as yet undetermined next coefficient in the expansion of $b(\varepsilon)$.

(b) Solve for the next-order coefficients, ω_2 in (7.48) and b_1 in (7.53).

Hint: Derive an equation for $x_2(t)$. Without actually solving this equation, determine ω_2 and b_1 by requiring that no secular terms arise in solving this equation. To recognize dangerous terms you will need to expand powers of trig functions in terms of multiple-angle trig functions, as in (7.55). The relevant formulas are most easily derived using complex exponentials.

(c) Determine the $\mathcal{O}(\varepsilon^2)$-correction to the period of the limit-cycle solution of van der Pol's equation (7.51).

15. *Introduction:* For every value of $\varepsilon > 0$, the IVP with the normalized initial conditions

$$x'' + x + \varepsilon x^3 = 0, \qquad x(0) = 1,\ x'(0) = 0,$$

has a periodic solution; if $\varepsilon < 0$, it still has a periodic solution, provided ε is not too large. In either case the period is $2\pi + \mathcal{O}(\varepsilon)$.

Use a Poincaré–Lindstedt expansion to determine the $\mathcal{O}(\varepsilon)$-correction to the period caused by the cubic nonlinearity.

Remark: With this equation you get nontrivial information more quickly than with the van der Pol equation. Specifically, to get the desired correction, you need not solve the ODE for the first-order term $x_1(\tau)$, you need only ensure that no secular terms would arise in solving it.

16. *Introduction:* This problem gives an analytical handle on the exponentially growing solutions of Mathieu's equation that you found numerically in Exercise 3.13. We include a frictional term, since this doesn't complicate the calculations significantly.

Apply the Poincaré–Lindstedt method to argue that for small $\varepsilon > 0$, the equation

$$x'' + \varepsilon\beta x' + (1/4 + 2\varepsilon\cos t)x = 0 \tag{7.81}$$

has an exponentially growing solution unless $\beta > 2$.

Hint: Floquet theory (cf. Section 7.10.1) suggests looking for a solution in the form

$$x(t,\varepsilon) = e^{(\lambda_1\varepsilon+\lambda_2\varepsilon^2+\ldots)t}\,[x_0(t)+\varepsilon x_1(t)+\varepsilon^2 x_2(t)+\ldots],$$

where each term $x_k(t)$ is periodic. As in the proof of Proposition 7.10.3, substitute this ansatz into (7.81) and separate terms of different orders to derive ODEs for $x_0(t)$ and $x_1(t)$. The x_0-equation has the general solution

$$x_0(t) = A\cos(t/2) + B\sin(t/2), \tag{7.82}$$

where A and B are as yet undetermined. Formulate the equation for $x_1(t)$; without actually solving it, derive the conditions on A and B such that no secular terms would be generated by solving it. You will find that the only choice without secular terms is $A = B = 0$ *unless* λ_1 satisfies a certain quadratic equation, whose roots give the Floquet exponents. Show that one root of this equation leads to a growing exponential unless $\beta > 2$.

Discussion: Both this problem and Proposition 7.10.3 study exponential growth of solutions of Mathieu's equation with the Poincaré–Lindstedt method. Despite this similarity, interpretation of the results differs greatly in the two problems. In (7.81), without the periodic forcing of the cosine term, solutions of this equation would exhibit slowly decaying oscillations characterized approximately by (7.82); the takeaway lesson is that the effect of periodic forcing can build up over time, leading to exponential growth. By contrast, in (7.100), without the cosine term, solutions would suffer exponential growth $e^{\lambda t}$, where $\lambda = \varepsilon(\sqrt{\beta^2+4\gamma}-\beta)/2$; the takeaway lesson is that oscillatory forcing can cancel such exponential growth.

Two other, lesser, differences between the two problems: (i) In analyzing (7.81), the Floquet exponent emerges already at first order in ε, while with (7.100), the calculation has to be carried to second order. (ii) Although the forcing in (7.81) is 2π-periodic, $x_0(t)$ *in* (7.82) *and the subsequent terms in the expansion have period* 4π. In other words, the Floquet multiplier in this problem is (real and) negative. By contrast, the Floquet multiplier in (7.100) is positive.

Remark: With only slightly more effort one can analyze (7.81) with some detuning, i.e., periodic forcing at a mismatched frequency. Specifically, for small ε, the equation

$$x'' + \varepsilon\beta x' + (1/4 + \varepsilon\kappa_1 + 2\varepsilon\cos t)x = 0 \tag{7.83}$$

has an exponentially growing solution if $\beta^2/4 + \kappa_1^2 < 1$.

17. (a) *Introduction:* In Exercise 1.13 you analyzed resonance in the periodic forcing of a linear spring. In this problem you study how nonlinearity can modify the phenomena. The general equation would be

$$x'' + \beta x' + x + \alpha x^3 = \Gamma\cos\omega t,$$

but we can't find explicit solutions of this equation. We want to isolate a perturbation-theory problem that we can solve but that still raises the main issues. Thus we assume (i) friction, the nonlinearity, and forcing are all small and (ii) the forcing frequency ω is nearly resonant, $\omega \approx 1$. These assumptions lead to the equation

$$x'' + \varepsilon\beta x' + x + \varepsilon\alpha x^3 = \varepsilon\cos((1+\varepsilon\Omega)t), \tag{7.84}$$

where we have used the strength of the forcing as the smallness parameter ε.

Apply the Poincaré–Lindstedt method to show that, to leading order, the amplitude A of the periodic response of (7.84) satisfies

$$(2\Omega A - 3\alpha A^3/4)^2 = 1 - (\beta A)^2. \tag{7.85}$$

Hint: In the usual way, seek an expansion of the solution

$$x_0(\tau) + \varepsilon x_1(\tau) + \ldots,$$

where $x_j(\tau)$ is a 2π-periodic function of the scaled time $\tau = (1 + \varepsilon\Omega)t$. The leading-order equation implies that $x_0(\tau)$ has the form $A\cos(\tau + \phi)$, where you may assume $A \geq 0$. The amplitude A and the phase angle ϕ cannot be determined at his level; two coupled equations for A and ϕ may be derived by requiring that no secular terms be generated in solving the x_1-equation. Eliminate ϕ from these two equations to obtain (7.85).

(b) *Introduction:* In Figure 7.20(a), assuming $\alpha = 0.5, \beta = 0.35$, the positive solution A of (7.85) is plotted as a function of the detuning parameter Ω, which measures the separation between the forcing frequency and the frequency of linear vibrations. This graph can be better understood by solving (7.85) for Ω as a function of A,

$$\Omega = \pm\frac{\sqrt{1 - (\beta A)^2}}{2A} + \frac{3}{8}\alpha A^2. \tag{7.86}$$

The two terms in this expression are graphed separately (for $A > 0$) in Figure 7.20(b). The first term is meaningful only if $0 < A \leq 1/\beta$; it follows from this observation that $A = 1/\beta$ is the maximum amplitude in Figure 7.20(a), a maximum that tends to infinity as $\beta \to 0$. The second term causes the asymmetry around $\Omega = 0$ of the response curve in Figure 7.20(a). The following calculation shows that if the nonlinearity is strong enough, this asymmetry can force A to be a multivalued function of Ω (as shown in the figure).

Choosing the plus/minus sign in (7.86) so that the two terms have the same sign, calculate $d\Omega/dA$ and show that this derivative has two zeros in the open interval $0 < A < 1/\beta$ if α is sufficiently large (in absolute value).

Remark: Although the above analysis does not show it, this periodic solution of (7.84) is asymptotically stable. Thus, you may check the rather surprising theoretical results with simulation.

7.8.3 Changes of Variables

18. (a) Verify that writing (7.4), the system

$$\begin{aligned} x_1' &= x_1 - x_2 - (x_1^2 + x_2^2)x_1, \\ x_2' &= x_1 + x_2 - (x_1^2 + x_2^2)x_2, \end{aligned} \tag{7.87}$$

in polar coordinates leads to (7.5).

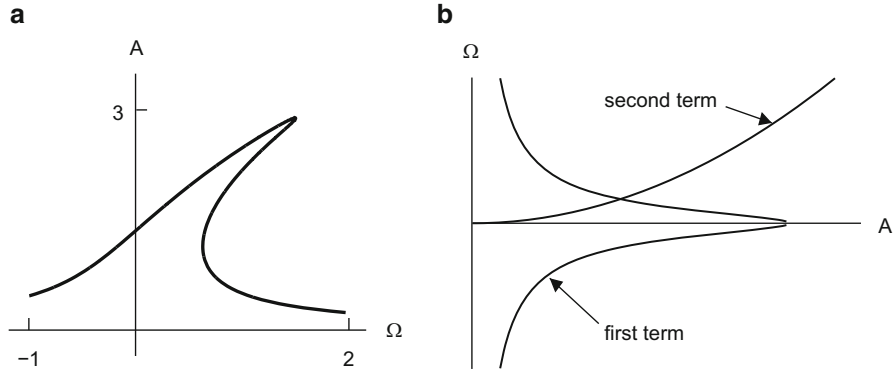

Figure 7.20: *(a) The solution of (7.85), the amplitude of the periodic solution of (7.84) as a function of the detuning parameter Ω, assuming $\alpha = 0.5, \beta = 0.35$. (b) Graphs of the two terms in (7.86), used in analyzing (7.85).*

> *Remark:* The main point of this exercise is to let you practice the slick technique described in Section 7.9.4; i.e., define $\mathbf{\Phi}(r, \theta)$ by
>
> $$\begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{\Phi}(r, \theta) = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix}$$
>
> and work out (7.95).

(b) Write the system

$$\begin{aligned} r' &= 1 - r, \\ \theta' &= 1, \end{aligned} \tag{7.88}$$

in rectangular coordinates.

> *Discussion:* For making the transformation in the reverse direction, you need to define $\mathbf{\Phi}(x, y)$ in (7.95) by
>
> $$\begin{bmatrix} r \\ \theta \end{bmatrix} = \mathbf{\Phi}(x, y) = \begin{bmatrix} \sqrt{x^2 + y^2} \\ \arctan(y/x) \end{bmatrix}.$$
>
> You will find that the equations in Cartesian coordinates have a $1/r$ singularity at the origin; i.e., your equation will be defined only on $\mathbb{R}^2 \sim \{\mathbf{0}\}$.

19. Suppose the two-dimensional ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ is $\mathcal{C}^1$ and has an equilibrium at the origin. Show that writing this equation in polar coordinates leads to a system such that for small r,

$$r' = rg(\theta) + o(r), \qquad \theta' = h(\theta) + o(1), \tag{7.89}$$

where

$$\begin{aligned} g(\theta) &= c_1 \cos^2\theta + c_2 \cos\theta \sin\theta + c_3 \sin^2\theta, \\ h(\theta) &= d_1 \cos^2\theta + d_2 \cos\theta \sin\theta + d_3 \sin^2\theta, \end{aligned}$$

and the coefficients satisfy

$$c_1 - c_3 + d_2 = 0, \qquad c_2 - d_1 + d_3 = 0. \tag{7.90}$$

Hint: Expand $\mathbf{F}(\mathbf{x})$ in a Taylor series

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} \alpha_{11}x_1 + \alpha_{12}x_2 \\ \alpha_{21}x_1 + \alpha_{22}x_2 \end{bmatrix} + o(|\mathbf{x}|),$$

apply (7.95), express c_k, d_k in terms of α_{jk}, and verify that equations (7.90) hold.

Remark: You may use this exercise to recognize when an ODE in polar coordinates comes from a nonsingular ODE on $\mathbb{R}^2$. For example, (7.5) has the form (7.89) with

$$c_1 = c_3 = d_1 = d_3 = 1, \qquad c_2 = d_2 = 0,$$

and conditions (7.90) are satisfied. By contrast, (7.88) does not have the required form.

20. (a) Suppose $\mathbf{x}(t)$ is a solution of (7.87) such that $|\mathbf{x}(0)| < 1$. Define the transformed unknown

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{\sqrt{1 - x_1^2}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{7.91}$$

and derive the ODE that $\mathbf{y}(t)$ satisfies.

Hint: This calculation gives the dedicated reader an opportunity for more practice using (7.95) to change variables in an ODE. Invert the transformation (7.91) to write

$$\mathbf{x} = \mathbf{\Phi}(\mathbf{y}) = \frac{1}{\sqrt{1 + y_1^2}} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

(b) Argue that for every nonzero initial condition $\mathbf{b}$ with $|b_2| < 1$, the ω-limit set of the solution of your equation for $\mathbf{y}$ equals the parallel lines

$$\{(y_1, y_2) : y_2 = \pm 1\},$$

an ω-limit set that is neither compact nor connected.

Hint: You needn't actually calculate the y-equation to do this part of the problem. The point is that the unit circle $\{x_1^2 + x_2^2 = 1\}$ is the ω-limit set of solutions of (7.87) with initial conditions inside the unit disk, and the transformation (7.91) maps the unit disk onto the strip $\{|y_2| < 1\}$.

7.8.4 PHD Exercises

21. *Introduction:* The ODE on $\mathbb{R}^2 \sim \{\mathbf{0}\}$,

$$r' = r - 1, \qquad \theta' = 1,$$

has an obvious periodic solution, say with $\boldsymbol{\gamma}(0) = (1, 0)$. Consider a Poincaré section

$$\Sigma = \{(r, \theta) : b_1 < r < b_2,\ \theta = 0\},$$

where say $b_1 = 1/2$ and $b_2 = 2$.

Find a neighborhood $\mathcal{N}$ of $(1, 0)$ such that a Poincaré map $\mathbf{\Pi} : \Sigma \cap \mathcal{N} \to \Sigma$ may be defined.

Remark: The point of this problem is that without the neighborhood $\mathcal{N}$, no Poincaré map of Σ to itself can be defined.

22. *Introduction:* In this exercise you construct an interesting ω-limit set from a Hamiltonian system modified with a carefully chosen gradient term:

$$\begin{aligned} x' &= \partial_y H - \varepsilon \partial_x (H^2), \\ y' &= -\partial_x H - \varepsilon \partial_y (H^2). \end{aligned} \tag{7.92}$$

(a) First consider (7.92) with $H(x, y) = (1 - x^2)(1 - y^2)$ and $\varepsilon = 0$. Show that (i) equation (7.92) has a one-parameter family of periodic orbits (which are level sets of $H(x, y)$) inside the square $\{|x| < 1, |y| < 1\}$ and (ii) the boundary of the square is a heteroclinic cycle, each vertex being a hyperbolic equilibrium.

(b) Then consider (7.92) with the same Hamiltonian and $\varepsilon > 0$. Show that (i) the square is still a heteroclinic cycle and (ii) it is the ω-limit of every trajectory starting from a nonzero point inside the square. (See Figure 7.21.)

Hint: The square is part of the zero set of $H(x, y)$, and the gradient term pushes the flow toward this set.

23. (a) Use a computer to calculate the period $T(\beta)$ of the limit-cycle solution of van der Pol's equation (7.58) for $\beta = 10, 20, 40, 80, 160, 320$.

Remark: You may find it challenging to obtain an accurate approximation for large β, since there are two obstacles working against you: (i) The period of the limit cycle increases with β. (ii) Large β may force you to compensate by choosing a *very* small time step in your numerical ODE solver.

(b) Make a log-log plot of

$$T(\beta) - [3 - 2\ln 2]\beta \quad \text{vs.} \quad \beta$$

and deduce from its slope that the first correction term in (7.69) is indeed of order $\beta^{-1/3}$.

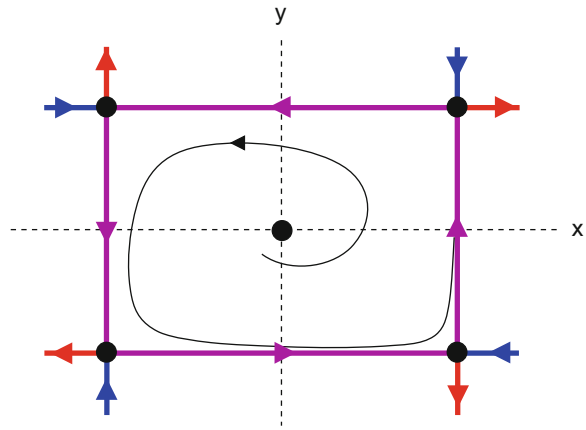

Figure 7.21: *An ω-limit set that is a square heteroclinic cycle. See Exercise 22.*

(c) Use curve-fitting software to estimate the best value of the coefficient C in fitting $T(\beta)$ by a curve of the form

$$[3 - 2\ln 2]\beta + C\beta^{-1/3}.$$

24. *Introduction:* Recall your computation in Exercise 6.23 that illustrated excitability in the FitzHugh–Nagumo equations 6.88.

Explain the behavior observed in your computation with fast-slow analysis, as in Section 7.6

25. Prove the following:

Proposition 7.8.1. *If A is a 2×2 matrix, then its eigenvalues have modulus less than 1 if and only if*

$$\begin{array}{cc} (i) & \mathrm{tr}A - \det A < 1, \\ (ii) & \mathrm{tr}A + \det A > -1, \\ (iii) & \det A < 1. \end{array}$$

Hint: Consider the eigenvalues of A as functions of the entries. The eigenvalues of the zero matrix have modulus less than 1, and the zero matrix satisfies Conditions (i)–(iii). As A varies, its eigenvalues will continue to have modulus less than 1 unless (i) a real eigenvalue passes through $+1$, (ii) a real eigenvalue passes through -1, or (iii) a pair of complex conjugate eigenvalues cross the unit circle. Show that conditions (i)–(iii) test for exactly these three occurrences.

Incidentally, drawing the set defined by Conditions (i)–(iii) in the plane with coordinates $\mathrm{tr}A$, $\det A$, which is a triangle, may help you remember these conditions.

7.9 Pearls of Wisdom

7.9.1 Area and Dulac's Theorem

Consider an ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ on a domain $\mathcal{U}$ with flow $\boldsymbol{\varphi}$. If $\mathcal{K} \subset \mathcal{U}$ is a region with a piecewise smooth boundary, write $\mathcal{K}_t$ for the image of $\mathcal{K}$ under $\boldsymbol{\varphi}(t, \cdot)$. Although the following proposition holds in every dimension (with area being replaced by volume), we restrict to the only case we need, $d = 2$.

Proposition 7.9.1. *If $\mathcal{A}$ indicates the area of a subset of $\mathbb{R}^2$, then*

$$\frac{d}{dt}\mathcal{A}(\mathcal{K}_t) = \iint_{\mathcal{K}_t} \operatorname{div}\ \mathbf{F}(\mathbf{x})\, dx_1 dx_2.$$

Proof. The calculations are clearer if we rewrite the equation in the equivalent form

$$\frac{d}{ds}\mathcal{A}(\mathcal{K}_{s+t})\Big|_{s=0} = \iint_{\mathcal{K}_t} \operatorname{div}\ \mathbf{F}(\mathbf{x})\, dx_1 dx_2.$$

Now by the semigroup property $\boldsymbol{\varphi}(s, \cdot) : \mathcal{K}_t \to \mathcal{K}_{s+t}$, so the change-of-variable formula for integrals (see Theorem 9.3.1 in [51], quoted in (B.26)) tells us that

$$\mathcal{A}(\mathcal{K}_{s+t}) = \iint_{\boldsymbol{\varphi}(s,\mathcal{K}_t)} dx_1 dx_2 = \iint_{\mathcal{K}_t} \det \mathbf{D}\boldsymbol{\varphi}(s, \mathbf{x})\, dx_1 dx_2.$$

We will differentiate under the integral, for which we need the following:

Claim 7.9.2.

$$\frac{\partial}{\partial s} \det \mathbf{D}\boldsymbol{\varphi}(s, \mathbf{x})\Big|_{s=0} = \operatorname{div}\ \mathbf{F}(\mathbf{x}).$$

Proof. With the order notation, this proof reduces to a sequence of self-explanatory formulas:

$$\boldsymbol{\varphi}(s, \mathbf{x}) = \mathbf{x} + s\mathbf{F}(\mathbf{x}) + o(s),$$

$$\mathbf{D}\boldsymbol{\varphi}(s, \mathbf{x}) = I + s\,\mathbf{DF}(\mathbf{x}) + o(s),$$

$$\det \mathbf{D}\boldsymbol{\varphi}(s, \mathbf{x}) = 1 + s\ \operatorname{tr} \mathbf{DF}(\mathbf{x}) + o(s).$$

(Regarding the third formula, see Exercises 3.9 and 4.21.) Of course, $\operatorname{tr} \mathbf{DF} = \operatorname{div} \mathbf{F}$. Thus, differentiation with respect to s proves both the claim and the proposition. □

Now let's apply the proposition to obtain a more intuitive proof of Dulac's theorem, first in the case that the function $g(\mathbf{x})$ is the constant function, $g(\mathbf{x}) \equiv 1$. Suppose $\operatorname{div} \mathbf{F}(\mathbf{x}) \geq 0$ and $\operatorname{div} \mathbf{F}(\mathbf{x})$ does not vanish identically on any open subset

of $\mathcal{U}$; then by the proposition, for every region, we have $(d/dt)\mathcal{A}(\mathcal{K}_t) > 0$. We claim that $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ cannot have a periodic orbit. If there were such an orbit, consider the region $\mathcal{K}$ inside the orbit and observe that $\mathcal{K}_t = \mathcal{K}$, so $(d/dt)\mathcal{A}(\mathcal{K}_t) = 0$, a contradiction.

Essentially the same proof works when $g(\mathbf{x})$ is nonconstant. Let $\boldsymbol{\psi}(t, \mathbf{x})$ be the flow associated with the ODE $\mathbf{x}' = g(\mathbf{x})\mathbf{F}(\mathbf{x})$, and for an arbitrary region let $\mathcal{K}_t = \boldsymbol{\psi}(t, \mathcal{K})$. If $\operatorname{div}(g\mathbf{F})(\mathbf{x}) \geq 0$ and $\operatorname{div}(g\mathbf{F})(\mathbf{x})$ does not vanish identically on any open subset of $\mathcal{U}$, then by the proposition, $(d/dt)\mathcal{A}(\mathcal{K}_t) > 0$. Suppose that $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ has a periodic solution. If $g(\mathbf{x})$ is nonvanishing, then the orbits of $\mathbf{x}' = g(\mathbf{x})\mathbf{F}(\mathbf{x})$ are exactly the same as those of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, so $\boldsymbol{\psi}$ has the same periodic orbit. Even if $g(\mathbf{x})$ does vanish, $\mathbf{x}' = g(\mathbf{x})\mathbf{F}(\mathbf{x})$ may have additional equilibria that could stop the flow from completing the circuit around the periodic orbit, *but* the region $\mathcal{K}$ inside this periodic orbit of $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ does not move under the flow of $\boldsymbol{\psi}(t, \cdot)$. Thus $\mathcal{K}_t = \mathcal{K}$, so $(d/dt)\mathcal{A}(\mathcal{K}_t) = 0$, a contradiction as before.

Here is a consequence of Proposition 7.9.1: *if* $\operatorname{div}\mathbf{F}(\mathbf{x}) \equiv 0$, *e.g., if* $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ *is Hamiltonian, then areas are preserved under the flow.* This result is true in any dimension.

These ideas also explain why in Section 7.2.6 in applying Dulac's theorem to (7.23) we use the factor $g(x, y) = 1/xy$: equation (7.23) is a perturbation of the Lotka–Volterra equations that have reparametrized Hamiltonian form

$$x' = xy\frac{\partial H}{\partial y}, \qquad y' = -xy\frac{\partial H}{\partial x},$$

and $1/xy$ cancels the reparametrization factor.

7.9.2 Poincaré-Like Maps in Constructing Periodic Solutions

If an autonomous system is subjected to periodic forcing, often the resulting system has a periodic solution with the same period. In Exercise 1.13 and Exercises 16 and 17 of the present chapter we analyzed resonance problems where such solutions could be found explicitly. Even in the absence of explicit solutions, there is a general construction for proving that such solutions exist, which we illustrate on the forced van der Pol equation,

$$x'' + \beta(x^2 - 1)x' + x = \Gamma\cos\omega t. \tag{7.93}$$

As in Exercise 10, we may rewrite this equation as a first-order autonomous system

$$\begin{aligned} x' &= y, \\ y' &= -x - \beta(x^2 - 1)y + \Gamma\cos z, \\ z' &= \omega. \end{aligned}$$

Given $\mathbf{b} \in \mathbb{R}^2$, let $\boldsymbol{\varphi}(t, \mathbf{b}) \in \mathbb{R}^3$ be the solution of the IVP for this system with initial conditions

$$(x(0), y(0), z(0)) = (\mathbf{b}, 0).$$

With respect to the Poincaré section $\{z = 0\}$, define a Poincaré-like map $\boldsymbol{\Pi}(\mathbf{b}) = \tilde{\boldsymbol{\varphi}}(2\pi/\omega, \mathbf{b})$, where $\tilde{\boldsymbol{\varphi}} = (\varphi_1, \varphi_2)$ denotes the first two components of $\boldsymbol{\varphi}$. Trapping regions can be constructed to show that $\boldsymbol{\Pi}$ maps a large ball of initial conditions in $\mathbb{R}^2$ continuously into itself. It then follows from the Brouwer fixed-point theorem[32] that $\boldsymbol{\Pi}$ has a fixed point, say $\boldsymbol{\Pi}(\mathbf{b}_*) = \mathbf{b}_*$. The solution of (7.93) with initial conditions $(\mathbf{b}_*, 0)$ is periodic with period $2\pi/\omega$.

Remarks: (i) This existence argument says nothing about uniqueness or stability. Especially for large forcing, (7.93) may exhibit complicated behavior that includes many periodic solutions, mostly unstable. (Cf. Section 2.1 of [33].) *(ii)* Perhaps, buoyed by our successes with the Poincaré–Lindstedt method elsewhere in the chapter, you are tempted to seek explicit solutions of (7.93) through an expansion in Γ. *Beware!* The conflict between oscillations at the natural period of the equation and the period of the forcing makes this treacherous territory.

7.9.3 Stable/Unstable Manifolds in Other Contexts

The concept of stable/unstable manifolds extends to hyperbolic periodic orbits of an ODE, where this concept is defined as follows: a periodic solution $\boldsymbol{\gamma}(t)$ with Poincaré map $\boldsymbol{\Pi}$ is called *hyperbolic* if the eigenvalues of $\mathbf{D}\boldsymbol{\Pi}$ satisfy

$$|\lambda_k(\mathbf{D}\boldsymbol{\Pi})| \neq 1.$$

For example, the periodic orbit in Exercise 11 is hyperbolic. For this particular orbit we may identify the stable and unstable manifolds explicitly, because the r, z-subsystem is linear. Specifically,

$$\mathcal{M}_s = \{(r, \theta, z) : z = -(r-1)\}, \qquad \mathcal{M}_u = \{(r, \theta, z) : z = r - 1\}. \tag{7.94}$$

(In these formulas, we are interpreting (7.80) as an equation on $\mathbb{R} \times (\mathbb{R}/2\pi\mathbb{Z}) \times \mathbb{R}$; if (7.80) is interpreted as an equation on Euclidean space, effectively restricting the variables to $(0, \infty) \times (\mathbb{R}/2\pi\mathbb{Z}) \times \mathbb{R}$, the condition $r > 0$ must be added to the definitions (7.94).)

These concepts also extend to fixed points of a mapping. In fact, stable/unstable manifolds for a periodic orbit of an ODE may be derived from stable/unstable manifolds for the fixed point of the Poincaré map of the periodic orbit. We recommend Section 2.6 of [17] for a user-friendly treatment of stable manifolds for maps.

[32] The best reference for this theorem may be to look online. It is covered in many topology texts, but usually it's near the end of the book where more of an investment is needed to extract the essential ideas.

7.9.4 Miscellaneous

On many occasions we have made a linear change of coordinates to simplify an ODE. Nonlinear changes of coordinates, such as polar coordinates, offer even more opportunities to simplify. It is always possible to derive the transformed equations with brute force, but there is a very slick way to derive them: Given an ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$, write a new coordinate, say $\mathbf{y}$, backward from what you might expect, in the form $\mathbf{x} = \mathbf{\Phi}(\mathbf{y})$. Then the $\mathbf{y}$-equation may be written with elegant compactness

$$\mathbf{y}' = \mathbf{D\Phi}(\mathbf{y})^{-1} \cdot (\mathbf{F} \circ \mathbf{\Phi})(\mathbf{y}). \tag{7.95}$$

Exercises 18–20 offer practice with this formula.

Every flow on a torus can be embedded in higher-dimensional Euclidean space. For example, in (7.20), linear flow on $\mathbb{T}^2$ is embedded in a three-dimensional ODE. More generally, the flow

$$\theta_1' = f(\theta_1, \theta_2), \qquad \theta_2' = g(\theta_1, \theta_2)$$

on $\mathbb{T}^2$ is embedded in the three-dimensional system

$$\begin{array}{ccc} \theta' & = & f(\theta, \phi(r,z)) \\ \begin{bmatrix} r' \\ z' \end{bmatrix} & = & \begin{bmatrix} 1-2R^2 & -g(\theta,\phi(r,z)) \\ g(\theta,\phi(r,z)) & 1-2R^2 \end{bmatrix} \begin{bmatrix} r-1 \\ z \end{bmatrix} \end{array}$$

where $\phi(r,z) = \arctan(z/(r-1))$ and $R(r,z) = \sqrt{(r-1)^2+z^2}$. This system is defined on $\{(r,\theta,z) : R(r,z) < 1\}$, which is a solid torus.

7.10 Appendix: Stabilizing an Inverted Pendulum

In Exercise 1.12 we found with computation that an inverted pendulum can be stabilized by rapid vibration of the pivot. In this appendix we make a perturbation-theory calculation that predicts this phenomenon.

7.10.1 A Smidgen of Floquet Theory

Floquet theory is concerned with a variable-coefficient $d \times d$ linear system[33]

$$\mathbf{x}' = A(t)\mathbf{x}, \tag{7.96}$$

[33]Such systems arise, among other contexts, in describing the linearization of the Poincaré map of a periodic solution of an ODE; this connection suggests the notation Π for the matrix in (7.97).

where $A(t)$ is T-periodic, i.e., $A(t+T) = A(t)$. The solution $\boldsymbol{\varphi}(t, \mathbf{b})$ of the IVP for (7.96) depends linearly on the initial condition $\mathbf{b}$. In particular, there is a (real) matrix Π such that at time T, after one period of the coefficient matrix $A(t)$,

$$\boldsymbol{\varphi}(T, \mathbf{b}) = \Pi\, \mathbf{b}, \tag{7.97}$$

which is called the *monodromy* matrix. The structure of solutions of (7.96) depends on this matrix. The following lemma expresses the simplest case of this connection.

Lemma 7.10.1. *If μ is an eigenvalue of Π, then (7.96) has a solution of the form*

$$\mathbf{x}(t) = e^{\lambda t}\mathbf{w}(t), \tag{7.98}$$

where $\lambda = T^{-1} \ln \mu$ and $\mathbf{w}(t)$ is T-periodic.

Remarks: The eigenvalue μ is called a *Floquet multiplier*, and λ is called a *Floquet exponent*. The Floquet multiplier μ cannot vanish *(why?)*, so its logarithm may be taken. If μ is real and positive, then $\lambda = \ln \mu$ is real, and the function $\mathbf{w}(t)$ can be assumed real-valued. Otherwise, the logarithm of μ must be taken in the complex sense, and then both λ and $\mathbf{w}(t)$ are complex-valued. This behavior is hardly surprising for a genuinely complex eigenvalue, but it is unnatural if μ is real and negative. A better representation for the solution in the latter case is given in Lemma 7.10.2 below.

Proof. Let $\mathbf{v}$ be an eigenvector of Π with eigenvalue μ, and let

$$\mathbf{w}(t) = e^{-\lambda t}\boldsymbol{\varphi}(t, \mathbf{v}).$$

Obviously, $e^{\lambda t}\mathbf{w}(t) = \boldsymbol{\varphi}(t, \mathbf{v})$ satisfies the ODE (7.96). To show that $\mathbf{w}(t)$ is periodic, we first apply the semigroup property to conclude that

$$\boldsymbol{\varphi}(t+T, \mathbf{v}) = \boldsymbol{\varphi}(t, \boldsymbol{\varphi}(T, \mathbf{v})) = \boldsymbol{\varphi}(t, \Pi\, \mathbf{v}) = \boldsymbol{\varphi}(t, \mu\mathbf{v}) = \mu\boldsymbol{\varphi}(t, \mathbf{v}),$$

where we have used the linearity of $\boldsymbol{\varphi}(t, \cdot)$ at the last step. Thus,

$$\mathbf{w}(t+T) = e^{-\lambda(t+T)}\boldsymbol{\varphi}(t+T, \mathbf{v}) = e^{-\lambda t}e^{-\lambda T}\mu\boldsymbol{\varphi}(t, \mathbf{v}) = e^{-\lambda t}\boldsymbol{\varphi}(t, \mathbf{v}) = \mathbf{w}(t),$$

since $e^{-\lambda T}\mu = 1$. □

Please check that essentially the same proof gives the following more natural representation when μ is real and negative.

Lemma 7.10.2. *If $\mu < 0$ is an eigenvalue of Π, then (7.96) has a (real) solution of the form*

$$\mathbf{x}(t) = e^{\lambda t}\mathbf{w}(t),$$

where $\lambda = T^{-1} \ln |\mu|$ and $\mathbf{w}(t)$, which is $2T$-periodic, satisfies $\mathbf{w}(t+T) = -\mathbf{w}(t)$.

If Π is diagonalizable, then there are d linearly independent solutions of (7.96) of the form (7.98), and the general solution of (7.96) is a linear combination of these basic solutions. If Π has repeated eigenvalues and a deficiency of eigenvectors, the general solution of (7.96) may still be described, but the construction is a little more complicated. We refer to Section 2.8 of [54] for details.

7.10.2 Some Stable Solutions of Mathieu's Equation

The Mathieu equation is a second-order linear ODE with a sinusoidal coefficient; flipping a sign, we may write it as

$$y'' + (-\kappa + 2\varepsilon \cos t)y = 0. \tag{7.99}$$

We are interested in the case $\kappa > 0$, which means that if $\varepsilon = 0$, generic solutions of the equation will grow like $e^{\sqrt{\kappa}t}$. However, if $\varepsilon \neq 0$ and κ is not too large, the periodic forcing makes all solutions of the equation remain bounded as $t \to \infty$. This statement can be proved with the same technique that we use to prove the following result for a modified equation that also includes a frictional term.

Proposition 7.10.3. *According to lowest-order perturbation theory in the parameter ε, all solutions of the equation*

$$y'' + \beta\varepsilon y' + (-\gamma\varepsilon^2 + 2\varepsilon \cos t)y = 0 \tag{7.100}$$

decay exponentially, provided $\gamma < 2$.

Proof. We calculate with the scalar equation (7.100), without reducing it to a first-order system as in (7.96). Despite this mismatch, we invoke Lemma 7.10.1 to justify looking for solutions of (7.100) in the form

$$y(t, \varepsilon) = e^{(\lambda_1\varepsilon + \lambda_2\varepsilon^2 + \ldots)t} \, [y_0(t) + \varepsilon y_1(t) + \varepsilon^2 y_2(t) + \ldots],$$

where each term $y_k(t)$ is 2π-periodic. By linearity, without loss of generality we may impose an initial condition that $y(0, \varepsilon) = 1$. A calculation shows that

$$e^{-(\lambda_1\varepsilon + \lambda_2\varepsilon^2 + \ldots)t} y'' = \varepsilon^0[y_0''] + \varepsilon[y_1'' + 2\lambda_1 y_0'] + \varepsilon^2[y_2'' + 2\lambda_1 y_1' + 2\lambda_2 y_0' + \lambda_1^2 y_0] + \ldots \, ;$$

y' has a simpler expansion, of which only the $\mathcal{O}(\varepsilon^0)$ and $\mathcal{O}(\varepsilon)$ terms are needed.

We substitute these formulas into (7.100) and collect terms at various orders:

$$\begin{array}{lrcl} \mathcal{O}(\varepsilon^0): & y_0'' & = & 0, \\ \mathcal{O}(\varepsilon^1): & y_1'' + 2\lambda_1 y_0' + \beta y_0' + 2\cos t \, y_0 & = & 0, \\ \mathcal{O}(\varepsilon^2): & y_2'' + 2\lambda_1 y_1' + 2\lambda_2 y_0' + \lambda_1^2 y_0 & & \\ & +\beta[y_1' + \lambda_1 y_0] - \gamma y_0 + 2\cos t \, y_1 & = & 0. \end{array} \tag{7.101}$$

The solution of the leading-order equation, including the initial condition, is $y_0(t) \equiv 1$; we reject t, the other linearly independent solution of the equation, because this has secular growth, which is not consistent with periodicity. Using this leading-order solution, we reduce the first-order equation to $y_1'' = -2\cos t$, whose solution, including initial condition, is

$$y_1(t) = 2(\cos t - 1).$$

No information about the growth rate is obtained at this level.

We rewrite the next-order equation, for $y_2(t)$, with the inhomogeneous terms isolated on the right:

$$-y_2'' = 2\lambda_1 y_1' + 2\lambda_2 y_0' + \lambda_1^2 y_0 + \beta y_1' + \beta\lambda_1 y_0 - \gamma y_0 + 2\cos t\, y_1. \tag{7.102}$$

We do not need to solve this equation, only to require that the inhomogeneity not force any aperiodic terms on the particular solution of the equation.

Claim 7.10.4. *Modulo terms that are periodic, the expression*

$$y_{\text{partic}}(t) = -(\lambda_1^2 + \beta\lambda_1 - \gamma + 2)t^2/2 \tag{7.103}$$

is a particular solution of (7.102).

Proof. Let us account for the seven terms on the RHS of (7.102) one by one. The second term, $2\lambda_2 y_0'$, vanishes. The first and fourth terms, $2\lambda_1 y_1'$ and $\beta y_1'$, are each proportional to $\sin t$; thus, their contributions to the particular solution are periodic. The third, fifth, and sixth terms, which are constant, are recognizable in (7.103). Finally, for the last term we have

$$2\cos t\, y_1 = 2 + 2\cos 2t - 4\cos t.$$

The constant term 2 appears in (7.103), and the contributions of the two cosine terms to the particular solution are periodic. □

The secular contribution (7.103) to the particular solution will vanish only if

$$\lambda_1^2 + \beta\lambda_1 - \gamma + 2 = 0. \tag{7.104}$$

The leading-order Floquet exponents are determined by this equation. By hypothesis, $\beta > 0$. Thus, both roots of (7.104) have negative real parts iff $-\gamma + 2 > 0$. □

The perturbation-theory proof of the proposition gives information only for small ε. Thus, we conclude from the proposition that there is a positive ε_0 such that if

$$\text{(a) } \varepsilon < \varepsilon_0 \qquad \text{and} \qquad \text{(b) } \gamma < 2, \tag{7.105}$$

then all solutions of (7.100) decay exponentially. In principle, ε_0 might vary with β, but computations suggest that this is not a problem.

7.10.3 Application to the Inverted Vibrated Pendulum

After scaling, the inverted vibrated pendulum is governed by the equation (1.53):

$$x'' + \beta x' + (1 - \alpha\omega^2 \cos \omega t) \sin x = 0. \tag{7.106}$$

Let us define the rescaled time $\tau = \omega t$ and linearize the equation about the unstable equilibrium $x = \pi$, which yields the equation

$$\frac{d^2y}{d\tau^2} + \frac{\beta}{\omega}\frac{dy}{d\tau} + \left(-\frac{1}{\omega^2} + \alpha \cos \tau\right) y = 0. \tag{7.107}$$

To connect (7.107) to (7.100), we ignore friction[34] and make the identifications

$$\gamma\varepsilon^2 = 1/\omega^2, \qquad 2\varepsilon = \alpha. \tag{7.108}$$

Thus, condition (7.105) for stability translates to

$$\text{(a) } \alpha < 2\varepsilon_0 \qquad \text{and} \qquad \omega^2 > \frac{2}{\alpha^2}. \tag{7.109}$$

In words, solutions of (7.107) will decay, provided the amplitude α of forcing is not too great and the frequency ω is sufficiently large. This behavior of the linearization carries over to the full equation (7.106) near the equilibrium at $x = \pi$.

[34]Equation (7.109a) imposes a limit on the amplitude of vibration for stabilization, and because of our cavalier treatment of friction, it is unclear how this limit depends on β. This issue could be explored numerically, but of course if $\beta = 0.1$, we know from Exercise 1.12 that $\alpha = 0.1$ is below this limit.

Chapter 8

Bifurcation from Equilibria

In Chapter 6 we studied the behavior of solutions of an ODE near a hyperbolic equilibrium point. In this chapter we turn to behavior near nonhyperbolic equilibria. Both for theoretical reasons and applications, it is natural to consider this problem in the context of a one-parameter family of ODEs, say

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}, \mu), \tag{8.1}$$

where $\mathbf{F} : \mathcal{U} \times \mathcal{I} \to \mathbb{R}^d$ is a vector-valued function on an open subset of $\mathbb{R}^d \times \mathbb{R}$. Suppose that for all μ near some fixed value μ_* in the interval $\mathcal{I}$, (8.1) has a smoothly varying equilibrium[1] $\mathbf{x}_{\text{eq}}(\mu)$ that is nonhyperbolic for $\mu = \mu_*$ but hyperbolic on either side of μ_*. Bifurcation theory seeks to characterize the behavior of solutions of (8.1) for μ near μ_*. Unlike the behavior near a hyperbolic point, this behavior *depends crucially on nonlinear terms* in the expansion of $\mathbf{F}$ at the equilibrium point.

Specific examples taken from applications play a central role in our study of bifurcation. The phenomena will be more or less familiar, but we describe them with a different focus that provides an invaluable new method to investigate ODEs. Indeed, bifurcation theory provides access to an amazing variety of mathematical, physical, biological, and engineering phenomena, more than we can describe or even list here. Just the examples considered in the text cover a rather wide range of applications. To highlight this point, and hopefully to enliven the exposition, in several of the supporting figures we include a cartoon suggestive of the application, utilizing space that otherwise would be blank.

After presenting examples of the best-known type of bifurcation in Section 8.1, we summarize the remainder of the chapter in Section 8.2.

[1] Up until now, a subscript star has been used to designate an isolated equilibrium of an ODE. In the present context, we have a curve of equilibria, and star designates a particular equilibrium of interest on that curve.

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8_8

8.1 Examples of Pitchfork Bifurcation

8.1.1 Bead on a Rotating Hoop

In Chapter 5 (see Section 5.1.1 and Exercise 5.5) we introduced ODEs for a bead sliding on a wire hoop that is rotating at constant speed, illustrated in Figure 8.1(a). After scaling, this led to the system

$$\begin{aligned} x' &= y, \\ y' &= -\beta y - \sin x + \mu \sin x \cos x. \end{aligned} \tag{8.2}$$

For every value of μ, (8.2) has the trivial equilibrium[2] $x = y = 0$, in which the bead is located at rest at the bottom of the hoop. To investigate its stability, we calculate the 2×2 Jacobian matrix[3] of (8.2) at $(0, 0, \mu)$:

$$\mathbf{DF}(\mathbf{0}, \mu) = \begin{bmatrix} 0 & 1 \\ -1 + \mu & -\beta \end{bmatrix}.$$

The determinant of this matrix equals $1 - \mu$, which vanishes if $\mu = 1$, so in this case the equilibrium is nonhyperbolic.

In fact, this example possesses additional structure that is typical for bifurcation problems. Specifically, if $\mu < 1$, then the equilibrium $(0, 0)$ is asymptotically stable; while if $\mu > 1$, it is unstable (more precisely, a saddle point). Colloquially, we say that the equilibrium loses its stability when μ crosses the point $\mu_* = 1$.

The central message of bifurcation theory is this: *When an equilibrium loses stability as a parameter is varied, expect new solutions of some type to appear.*[4] Acting on this message, we look for steady-state solutions of (8.2). The velocity y vanishes at every equilibrium. Substituting $y = 0$ into the second equation yields the condition

$$(-1 + \mu \cos x) \sin x = 0. \tag{8.3}$$

The sine factor vanishes if $x = 0$ or $x = \pi$; i.e., this factor gives the obvious two equilibria of (8.2). The other factor vanishes if

$$\cos x = \frac{1}{\mu}. \tag{8.4}$$

[2]We ignore the equilibrium at $x = \pi$, which is unstable for all $\mu \geq 0$.

[3]In the context of a parametrized family of ODEs like (8.1), the notation $\mathbf{DF}$ denotes the matrix of derivatives of $\mathbf{F}$ with respect to *the state variables* $x_1, \ldots, x_d$ *only*. We write out derivatives with respect to parameters explicitly, such as $\partial \mathbf{F}/\partial \mu$.

[4]The occurrence of additional solutions does not conflict with the result from Chapter 3 that under minimal hypotheses, the solution of the IVP is unique. The new solutions here are *equilibrium* solutions, possible ω-limits of trajectories.

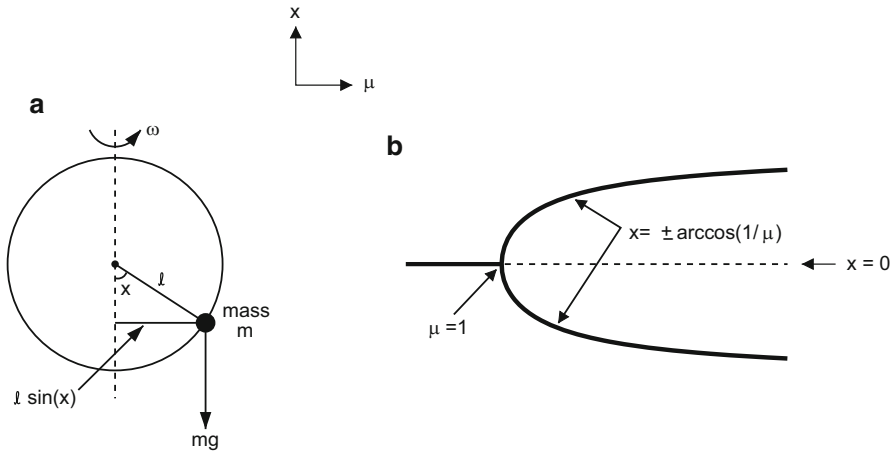

Figure 8.1: *(a) Schematic diagram of the bead on a rotating wire hoop, equation* (8.2)*. The nondimensional bifurcation parameter is* $\mu = \ell\omega^2/g$*. (b) Bifurcation diagram for this system. Here and below, stable equilibria are shown as solid curves; unstable, dashed. In particular, the trivial equilibrium* $x = 0$ *loses stability when the nondimensionalized rotation speed* μ *exceeds* 1*, and simultaneously two new, stable equilibria appear. (Unstable equilibria at* $x = \pm\pi$ *are not shown in the figure.)*

This equation has no real solutions if $0 \le \mu < 1$, but two real solutions appear as soon as μ crosses the critical value $\mu_* = 1$. (Can you hear the spirit of bifurcation theory whispering smugly, "I told you so"?)

Figure 8.1(b), known as a *bifurcation diagram*, shows a graph of these various equilibrium solutions in the x, μ-plane. Intervals of μ where the equilibria are asymptotically stable are indicated by a solid curve; unstable, by a dotted curve. (You may either calculate now or recall from Exercise 6.2(c) that the new equilibria given by (8.4) are stable.)

Bifurcation diagrams are usually interpreted in the context of what is called *quasistatic variation of parameters*. Imagine that, starting from the equilibrium $x = 0$ with $\mu < 1$, we increase μ by a small increment and wait until the system returns to equilibrium; then increase μ by another small increment and again wait for reequilibration, etc. Nothing will happen as long as μ stays smaller than 1; the system will remain at its stable equilibrium at $x = 0$. However, when μ crosses $\mu_* = 1$, we expect the system to move away from this equilibrium. Strictly speaking, $x = 0$ is still an equilibrium when $\mu > 1$, but since it is now unstable, if the system is subjected to the slightest bit of noise, the solution will evolve away from $x = 0$. Using Exercise 6.12 as a guide, you can show that for $\mu > 1$, virtually all solutions will tend to one of the equilibria (8.4). The solution may evolve to either equilibrium, $x = \pm\arccos(1/\mu)$, when μ first crosses $\mu_* = 1$; which case occurs depends on accidents

in initial conditions and/or noise. However, once one of the two branches has been selected, the system will follow that branch under further quasistatic increases of μ.

Remarks: (i) The origin of the term *bifurcation* may be seen in Figure 8.1(b): as μ is increased, the unique stable solution $x = 0$ is replaced by the two[5] stable solutions, $x = \pm \arccos(1/\mu)$. *(ii)* In bifurcation theory we often abbreviate "asymptotically stable" to "stable" (as in the previous remark). Lyapunov stability doesn't play much of a role in this subject. *(iii)* The particular bifurcation diagram in Figure 8.1(b) is known as a *pitchfork*. As will be discussed in Section 8.5.5, pitchfork bifurcations are common in systems that exhibit reflectional symmetry. Illustrating this behavior, nonzero equilibria of (8.2) occur in symmetric pairs that are mapped into one another by the reflection $x \mapsto -x$, while the trivial solution $x = 0$ is invariant under the reflection. In the dynamics, this symmetry is expressed as follows: if $(x(t), y(t))$ is a solution of (8.2), then so is $(-x(t), -y(t))$. *(iv)* It follows from a general result, called the *principle of exchange of stability*, that the bifurcating solutions in Figure 8.1(b) are stable. In other words, although it was pedagogically useful, the specific calculation to derive this behavior was actually unnecessary. This principle will be discussed in Section 8.5.4.

8.1.2 The Lorenz Equations

As a second example of a pitchfork bifurcation, recall from Exercise 4.3(a) the Lorenz equations

$$\begin{array}{lrcl} (a) & x' & = & \sigma(y - x), \\ (b) & y' & = & \rho x - y - xz, \\ (c) & z' & = & -\beta z + xy, \end{array} \tag{8.5}$$

where σ, ρ, and β are positive (dimensionless) parameters. We reverse the order of presentation from the previous example. Here we first look for equilibrium solutions of (8.5), and then we make the connection with a loss of stability. The first equation implies that $x = y$ at equilibrium, the third equation then implies that $z = y^2/\beta$, and substitution into the second yields the equation

$$y(\rho - 1 - y^2/\beta) = 0. \tag{8.6}$$

This equation can be satisfied by virtue of either factor's vanishing, which yields a pitchfork bifurcation diagram as shown in Figure 8.2(a), where ρ is taken as the bifurcation parameter. As is conventional, we plot only the one variable y in the figure. This variable is sufficient to determine the equilibria of (8.5): given y, the other two variables may be obtained from it as in the above analysis.

[5]The now unstable equilibrium at $x = 0$ for $\mu > 1$ is not included in the counting; thus one does not speak of "trifurcation." Somewhat inconsistently, the middle branch is not ignored in the term "pitchfork."

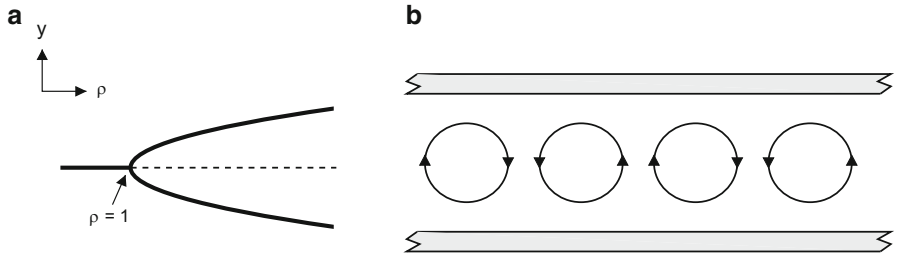

Figure 8.2: *(a) Bifurcation diagram for the Lorenz system* (8.5)*. Solid curves correspond to stable equilibria; the dashed line, an unstable equilibrium. (b) A cartoon of the physical problem from which the Lorenz equation* (8.5) *was extracted: convective rolls of fluid between two horizontal plates held at different temperatures. The plates extend to infinity in both directions; the rolls, indicated schematically, are infinite cylinders viewed end on. The figure shows only four representative cells in an infinite array.*

We now consider stability. You may either recall from Exercise 6.2(d) or calculate directly that the trivial solution $x = y = z = 0$ of (8.5) is asymptotically stable for $\rho < 1$ and unstable for $\rho > 1$. The bifurcating solutions, for which $y = \pm\sqrt{\beta(\rho - 1)}$, grow out of the trivial solution at exactly the same point—the bifurcation point—where the trivial solution loses stability. Thus, we have another instance of the central phenomenon of bifurcation theory.

As in the preceding example, the bifurcating equilibria are asymptotically stable in some neighborhood of the bifurcation point. Again this behavior follows from the principle of exchange of stability, but while you are learning the subject, it may be reassuring to verify it explicitly.

As before, (8.5) has a reflectional symmetry, specifically, with respect to the mapping $(x, y, z) \mapsto (-x, -y, z)$.

Some discussion of physical interpretations makes this example more meaningful. (And see Gleick's book [30] for some of the history of this fascinating equation.) E. Lorenz [50] studied (8.5) as a model problem to shed light on the generation of weather patterns in the atmosphere. The system arises from a massive simplification of PDEs that describe *Rayleigh–Bénard convection*. This term refers to motion of fluid confined between infinite horizontal parallel plates held at fixed temperatures. If the lower plate is sufficiently hot, then thermal expansion of the fluid induces buoyant motion as low-density expanded fluid rises and displaces fluid in the denser layer above it.

In (8.5), the variable y specifies the amplitude of a velocity field in the fluid in the form of rolls, as indicated schematically in the cartoon Figure 8.2(b). (The other two variables specify the temperature.) Rolls in adjacent cells alternate in orientation,

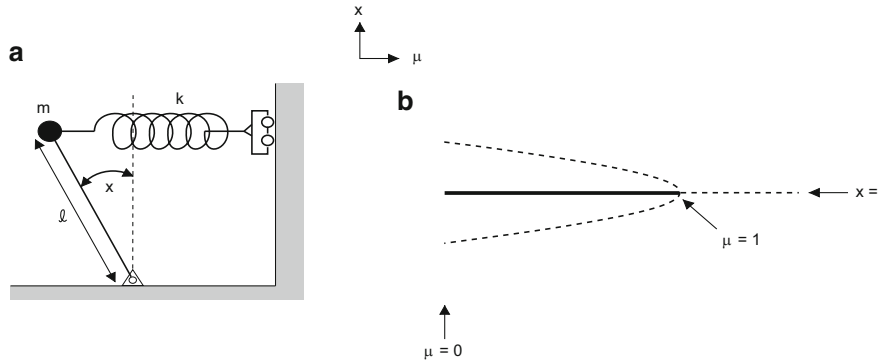

Figure 8.3: *(a) Schematic diagram of a laterally supported pendulum, the physical system described by the potential energy* (8.7)*. (b) Bifurcation diagram for this system;* μ *is the nondimensionalized mass,* $\mu = mg/k\ell$*.*

clockwise or counterclockwise. If the velocity field in the figure—*cw, ccw, cw, ccw* for the four cells shown—corresponds to $y > 0$, then $y < 0$ corresponds to the reverse orientation, *ccw, cw, ccw, cw.* This reversal of orientation gives rise to the reflectional symmetry in (8.5) mentioned above.

In (8.5), the bifurcation parameter ρ is a nondimensionalized measure of the temperature difference between the two plates, called the *Rayleigh number.* The equilibrium $x = y = z = 0$, which corresponds to a state in which the fluid is stationary, is stable when the temperature difference between the two plates is small enough, i.e., $\rho < 1$. However, if heating is increased so that $\rho > 1$, the trivial equilibrium loses stability, and motion ensues.

8.1.3 A Laterally Supported Inverted Pendulum

Both Figures 8.1(b) and 8.2(a) show the same qualitative behavior under quasistatic variation of parameters; i.e., as forcing is increased, a trivial response of the system evolves into one of two stable nontrivial steady-state responses. Our next example is a pitchfork bifurcation that exhibits different qualitative behavior.

Consider the laterally supported pendulum shown in Figure 8.3(a). The cart rolls up and down, so that the spring remains horizontal.[6] For variety let's analyze the

[6]The geometry in Figure 8.3(a) may seem a little artificial to you, but the behavior in this problem is actually representative of the collapse of a variety of structures. Such behavior is illustrated, for example, in Exercise 14. However, the more realistic geometry of the exercise is also more complicated to analyze. The artificially simple geometry in Figure 8.3(a) provides a gentler first encounter with subcritical bifurcation.

Incidentally, for the geometry in the figure, x is constrained to the interval $(-\pi/2, \pi/2)$. Although we ignore this constraint in analyzing (8.8), one might question the physical significance of the equilibrium solution $x = \pm\pi$ outside this range.

bifurcations of this system using energy considerations, rather than working directly from the ODE. In dimensional units, the potential energy of this system equals

$$\hat{V}(x) = mg\ell \cos x + k(\ell \sin x)^2/2,$$

where x measures the angle of the pendulum (with $x = 0$ being straight up). Scaling by $1/k\ell^2$, we obtain the nondimensional potential energy

$$V(x) = \mu \cos x + (\sin x)^2/2, \tag{8.7}$$

where $\mu = mg/k\ell$, which may be interpreted as the nondimensionalized mass. The equation for equilibria, $\partial V/\partial x = 0$, is

$$[-\mu + \cos x] \sin x = 0. \tag{8.8}$$

Two solutions of this equation, $x = 0$ and π, are associated with the sine factor in (8.8) vanishing. Provided $\mu < 1$, the nontrivial factor of (8.8) has two solutions located symmetrically about the inverted equilibrium $x = 0$. These solutions are graphed in the bifurcation diagram[7] of Figure 8.3(b). Stabilities of all equilibria, which may be determined from the sign of $\partial^2 V/\partial x^2$, are also indicated in the figure. (*Verify these!*) In words: if the scaled mass μ is not too great, the inverted equilibrium $x = 0$ is stable and the two nontrivial equilibria are unstable; when $\mu > 1$, the inverted equilibrium becomes unstable, and the nontrivial equilibria disappear.

Let's see how, under quasistatic variation of parameters, the behavior of this system differs from the one above. In (8.2), for example, when μ is increased (quasistatically) beyond $\mu_* = 1$, we expect the system to move away from the trivial equilibrium $x = 0$, but its response is still continuous in μ. Specifically, it will follow one of the solutions $x = \pm \arccos(1/\mu)$ of (8.4). Moreover, there is essentially no dynamics in this response; i.e., x changes only in proportion to the amount μ is increased. By contrast, the laterally supported pendulum in the straight-up position *collapses* when μ is increased ever so slightly beyond $\mu_* = 1$. In more restrained language, the system undergoes a significant transient and evolves to states far removed from $x = 0$.

The pitchfork bifurcations in Sections 8.1.1 and 8.1.2 are called *supercritical*. This term refers to the fact that the nontrivial equilibria appear for forcing that is greater than what is required to destabilize the trivial solution. By contrast, the bifurcation in the present section is called *subcritical* because the nontrivial equilibria exist for smaller forcing than required to destabilize the trivial solution. Some authors use the terms *soft* and *hard* for supercritical and subcritical bifurcations, respectively. Others refer to *continuous* and *discontinuous* bifurcations, respectively.

[7] Although $x = \pm\pi$ satisfies (8.8), we don't show this solution in the figure.

8.2 Perspectives on This Chapter

8.2.1 An Outline of the Chapter

The term "bifurcation" seems natural to describe the phenomena analyzed in Section 8.1. More generally, however, this term has come to be used to refer to any change in the qualitative behavior of solutions of an ODE as a parameter changes. Such changes include local phenomena, which are the focus of the present chapter, and global phenomena, which are the focus of Chapter 9. Particularly in the latter case, the relevant behavior may have no association whatsoever with the word bifurcation.

The context of local bifurcation theory is a one-parameter family of ODEs $\mathbf{x}' = \mathbf{F}(\mathbf{x}, \mu)$ such that for some specific parameter value μ_*, the equation $\mathbf{x}' = \mathbf{F}(\mathbf{x}, \mu_*)$ has an isolated nonhyperbolic equilibrium. The central phenomenon of the theory is that there (usually) are additional, nonobvious, solutions of the equation near the nonhyperbolic equilibrium. This behavior is illustrated by numerous examples in this chapter. In the examples of Sections 8.1, 8.3, 8.4, and 8.6, the additional solutions are equilibrium solutions, which exhibit what is called *steady-state* bifurcation. By contrast, in the examples of Section 8.7, the new solutions are (time-dependent) periodic solutions, which represent what is called *Hopf* bifurcation.

To supplement the examples-oriented sections, the remainder of Section 8.2 and Sections 8.5, and 8.8 address theory. Section 8.2.2 formulates a theorem that unifies many steady-state bifurcation phenomena. Section 8.5 introduces the Lyapunov–Schmidt reduction; this is an exceedingly long section, and it is not easy reading, but a thorough understanding of steady-state bifurcation can be gleaned with this technique. In particular, the partial classification of bifurcation problems given in Table 8.1 is based on this technique. Section 8.8 presents some of the theory concerning Hopf bifurcation.

8.2.2 A Bifurcation Theorem

The following theorem identifies the common theme in many bifurcation problems, including the three examples of Section 8.1. Consider a one-parameter family of ODEs such that for all μ near μ_*, the equation

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}, \mu) \tag{8.9}$$

has an isolated equilibrium, $\mathbf{x} = \mathbf{x}_{\text{eq}}(\mu)$, which varies smoothly with μ. We call this the *trivial solution*. Suppose that for $\mu = \mu_*$, the Jacobian $\mathbf{DF}_*$ of this equation at $(\mathbf{x}_{\text{eq}}(\mu_*), \mu_*)$ has a simple eigenvalue zero; i.e., assume

$$\lambda_1(\mathbf{DF}_*) = 0, \qquad \lambda_j(\mathbf{DF}_*) \neq 0, \ j = 2, \ldots, d. \tag{8.10}$$

Under this hypothesis, Proposition C.3.1 yields the following information about the Jacobian $\mathbf{DF}(\mathbf{x}_{\rm eq}(\mu),\mu)$ at nearby equilibria. (You need not consult Appendix C; we will derive this result below in Section 8.5.2(a) below.)

Lemma 8.2.1. *Along the trivial solution, there is a unique smoothly varying (real) eigenvalue $\lambda(\mu)$ of $\mathbf{DF}(\mathbf{x}_{\rm eq}(\mu),\mu)$ in a neighborhood of μ_* such that $\lambda(\mu_*) = \lambda_1(\mathbf{DF}_*) = 0$.*

The bifurcation theorem requires an additional hypothesis[8] on how this eigenvalue depends on μ, i.e.,

$$\frac{d\lambda}{d\mu}(\mu_*) \neq 0. \tag{8.11}$$

It may seem unclear how to differentiate $\lambda(\mu)$, but we will show below that (8.11) is equivalent to

$$\frac{d}{d\mu}\det \mathbf{DF}(\mathbf{x}_{\rm eq}(\mu),\mu)\Big|_{\mu=\mu_*} \neq 0, \tag{8.12}$$

and this is readily verifiable. The following theorem, which we prove in Section 8.5.2, characterizes the local geometry of the bifurcation diagram

$$\{(\mathbf{x},\mu)\in\mathbb{R}^d\times\mathbb{R} : \mathbf{F}(\mathbf{x},\mu)=\mathbf{0}\}$$

as two smooth curves that cross at $(\mathbf{x}_*,\mu_*)$, where $\mathbf{x}_* = \mathbf{x}_{\rm eq}(\mu_*)$. The parametrization (8.13) in the theorem, which may seem unintuitive, is discussed below.

Theorem 8.2.2. *Under the above hypotheses, there is a one-parameter family of nontrivial equilibrium solutions of (8.9) near $(\mathbf{x}_*,\mu_*)$. Assuming $\mathbf{F}\in\mathcal{C}^3$, these solutions admit a $\mathcal{C}^2$ parametrization $(\mathbf{x},\mu) = (\mathbf{X}(a),\mu(a))$ in terms of an amplitude parameter a (which lies in an interval around $a=0$),*

$$(a)\ \mathbf{X}(a) = \mathbf{x}_{\rm eq}(\mu(a)) + a\mathbf{v} + \mathcal{O}(a^2), \quad (b)\ \mu(a) = \mu_* + \mu_1 a + \mu_2 a^2 + o(a^2), \tag{8.13}$$

where $\mathbf{v}$ spans the kernel of $\mathbf{DF}_*$ and μ_1,μ_2 are constants. Every equilibrium of (8.9) in an appropriate neighborhood $\mathcal{N}\subset\mathbb{R}^d\times\mathbb{R}$ of $(\mathbf{x}_*,\mu_*)$ lies on either the trivial branch $\{(\mathbf{x}_{\rm eq}(\mu),\mu)\}$ or the bifurcating branch (8.13).

Let's interpret the theorem for the Lorenz equations (8.5), starting with hypotheses. We have $\mathbf{x}_{\rm eq}(\rho) = (0,0,0)$ and

$$\mathbf{DF}(\mathbf{0},\rho) = \begin{bmatrix} -\sigma & \sigma & 0 \\ \rho & -1 & 0 \\ 0 & 0 & -\beta \end{bmatrix}. \tag{8.14}$$

[8] If the nonzero eigenvalues of $\mathbf{DF}_*$ have negative real parts, then (8.11) implies that $\mathbf{x}_{\rm eq}(\mu)$ is stable for μ on one side of μ_* and unstable on the other. In other words, $\mathbf{x}_{\rm eq}(\mu)$ loses stability as μ crosses μ_*. This is the most interesting case.

Let $\rho_* = 1$; then $\mathbf{DF}_*$ has one zero eigenvalue and negative eigenvalues $-(\sigma+1)$ and $-\beta$. Taking the determinant of (8.14), you may easily verify (8.12).

Regarding the parametrization (8.13) of the bifurcating solutions, in Section 8.1.2 we showed that for every value of a, the point $(a, a, a^2/\beta)$ is an equilibrium solution of (8.5) with $\rho = 1 + a^2/\beta$. Thus, we have (8.13) with $\mathbf{v} = (1, 1, 0)$ as the null eigenvector of $\mathbf{DF}_*$, $\mu_1 = 0$, and $\mu_2 = 1/\beta$.

For this example, the parameter a can be eliminated from the description of the bifurcating solutions: inverting $\rho = 1 + a^2/\beta$ gives $a = \pm\sqrt{\beta(\rho-1)}$, and thus the nontrivial equilibria can be expressed directly as a function of the bifurcation parameter. However, this is not possible in general.[9] For instance, the trivial linear scalar equation

$$x' = \mu x \tag{8.15}$$

satisfies the hypotheses of the theorem, with $x_{\mathrm{eq}}(\mu) \equiv 0$. The bifurcating solutions, which lie on the line $\{\mu = 0\}$, may be parametrized by amplitude as in (8.13),

$$\mathbf{X}(a) = a \quad \text{and} \quad \mu(a) \equiv 0,$$

but x cannot be expressed as a function of μ.

Remarks: (i) Exercises 9 and 10 illustrate some of the pathology that can appear if $\mathbf{DF}_*$ has a multiple eigenvalue 0 or if (8.11) is not satisfied. *(ii)* Much of bifurcation theory, including the above theorem, extends naturally to infinite-dimensional problems. (Cf. the elastica in Chapter 10, equation (10.71).)

8.3 Examples of Transcritical Bifurcation

8.3.1 Section 1.6 Revisited: Part IV

Recall from Section 1.6.2 the Lotka–Volterra model of a predator–prey system (1.41) modified to have logistic growth for the prey:[10]

$$\begin{aligned} (a) \quad x' &= x(1 - x/K) - xy, \\ (b) \quad y' &= \rho(xy - y). \end{aligned} \tag{8.16}$$

[9] Besides (8.15), you should be aware of mischievous possibilities like $x' = \varphi(x) - \mu x$, where $\varphi(x) = e^{-1/x^2}\sin(1/x)$, defined for $x = 0$ by $\varphi(0) = 0$. The bifurcating equilibria wiggle back and forth infinitely many times above the point $\mu = 0$.

[10] For simplicity, we do not include the Allee effect. Locally near the bifurcation point, this effect would make little difference, as you can easily verify.

We regard K as the bifurcation parameter and the prey-only equilibrium $\mathbf{x}_{\rm eq}(K) = (K, 0)$ as the trivial solution. Along this solution branch,

$$\mathbf{DF}(\mathbf{x}_{\rm eq}(K), K) = \begin{bmatrix} -1 & -K \\ 0 & \rho(K-1) \end{bmatrix}.$$

With $K_* = 1$, the Jacobian $\mathbf{DF}_*$ has a simple eigenvalue zero, and (8.11) may be verified by inspection. As shown in the bifurcation diagram in Figure 8.4(a), a new branch of equilibria, coexistence equilibria, given by

$$(1, 1 - 1/K), \tag{8.17}$$

bifurcates from the prey-only equilibrium branch at precisely this value of K. This kind of bifurcation is known as a *transcritical bifurcation*, because the bifurcating solutions exist for K both below and above the bifurcation point.

Regarding Theorem 8.2.2, our calculation has shown that all equilibrium solutions of (8.16) lie on either the trivial solution branch $\{(\mathbf{x}_{\rm eq}(K), K)\}$ or the branch of bifurcating solutions, (8.17). The key to obtaining the exact parametrization (8.13) depends on choosing the function $K(a)$ in (8.13b) appropriately. As you will show in Exercise 2, $K(a) = 1 + a$ is one possible choice.

It is easily verified that as indicated in Figure 8.4(a), the bifurcating solution is stable for $K > 1$, and it is unstable for $K < 1$; moreover, in the latter case it lies in the unphysical domain $\{y < 0\}$.

Let us articulate behavior implied by Figure 8.4(a) if K is varied quasistatically: if $K > 1$, then every solution of (8.16) with a nonzero prey population at $t = 0$ will converge to the coexistence equilibrium, but if K is reduced below unity, the predators will die out. (Incidentally, at this level of detail, the behavior of the system doesn't depend on the parameter ρ, but ρ will affect the global dynamics studied in Chapter 9.)

8.3.2 The Chemostat

In Section 6.7.1 we studied the phase portrait of the chemostat,

$$\begin{aligned} x' &= \frac{y}{y+1}x - \rho x, \\ y' &= -\frac{y}{y+1}x - \rho(y - \sigma), \end{aligned} \tag{8.18}$$

under the assumption (6.72), or equivalently that

$$\rho < \frac{\sigma}{\sigma + 1}. \tag{8.19}$$

This inequality is significant, because if ρ is regarded as a bifurcation parameter, the system undergoes a transcritical bifurcation at $\rho_* = \sigma/(\sigma + 1)$. Specifically, if

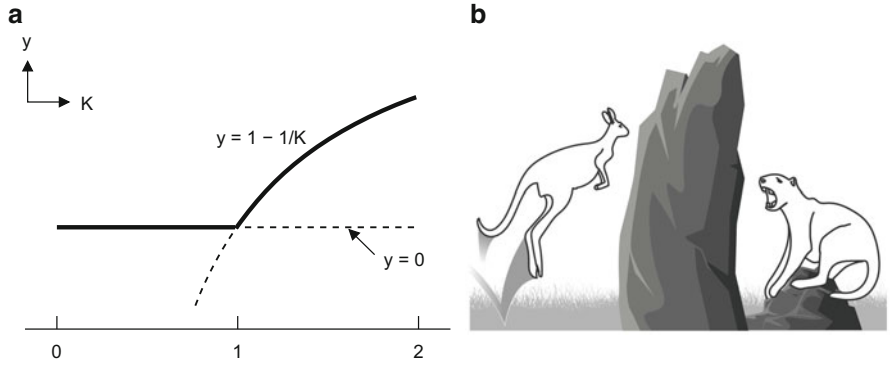

Figure 8.4: *(a) Transcritical bifurcation in an augmented Lotka–Volterra model* (8.16) *at* $K = 1$*: exchange of stability between prey-only and coexistence equilibria as the carrying capacity* K *crosses 1. (b) A million years ago, the marsupial lion in the cartoon was the top predator in Australia, but currently there are no mammalian carnivores weighing more than twenty pounds. What happened? There is no general agreement about this, but one theory is that climate changes, aggravated by low soil fertility, reduced the prey population below what will support large mammalian predators. Assuming that the extinction can be captured by a two-species mathematical model, it is suggested in Chapter 12 of [85] and the article by Flannery reprinted there that* K *in* (8.16) *has fallen below its critical value. (Cartoon drawn by Jeff Poe, 2016. Copyright 2016 by D.G. Schaeffer and J.W. Cain.)*

$\rho > \rho_*$, the only equilibrium in the physical domain is the trivial state $(0, \sigma)$ without any bacteria, which is stable. As ρ decreases across ρ_*, a new branch of stable equilibria bifurcates from the trivial equilibrium, the latter becoming unstable. We leave the more-or-less painless verification of these claims to you, urging you to draw the bifurcation diagram that summarizes them.

In physical terms, if the (nondimensionalized) flow rate ρ is too large, bacteria get flushed out of the vessel faster than they can replace themselves through growth. However, if $\rho < \rho_*$, a colony can get established.

8.4 Examples of Saddle-Node Bifurcation

In pitchfork and transcritical bifurcations, a smoothly varying equilibrium solution of an ODE loses stability as a parameter crosses a threshold. In saddle-node bifurcations, two equilibria, typically one of them stable, both disappear as a parameter crosses a threshold. Unlike previous bifurcations, saddle-node bifurcations do not fit into the rubric[11] of Theorem 8.2.2.

8.4.1 The Torqued Pendulum

The most intuitive saddle-node bifurcation is in the (scaled) torqued pendulum introduced in Section 4.3.3 (cf. Figure 8.5(a)),

$$\begin{aligned} x' &= y, \\ y' &= -\sin x - \beta y + \mu, \end{aligned} \tag{8.20}$$

where we may assume without loss of generality that the bifurcation parameter μ is nonnegative. The bifurcation diagram for this equation is shown in Figure 8.5(b). If $\mu < 1$, (8.20) has two equilibria. The equilibrium that is between 0 and $\pi/2$—the pendulum lies below the horizontal—is a sink, and at least for μ close to 1 it is a (stable) node; i.e., the eigenvalues of the Jacobian are real (and negative). By contrast, the equilibrium between $\pi/2$ and π is a saddle. (*Check these claims!*) The node disappears when it and the saddle annihilate one another as the point $\mu = 1$ is crossed. This behavior is called *saddle-node* bifurcation,[12] or *limit-point* bifurcation [31], or rather poetically *blue-sky* bifurcation [1].

Once again, the bifurcation diagram suggests a specific scenario under quasistatic increase of μ: While $\mu < 1$, the system can follow its stable equilibrium in the interval

[11] Despite this mismatch, we articulated Theorem 8.2.2 because we think it is informative for both pedagogical and historical reasons. The Lyapunov–Schmidt reduction of Section 8.5 applies to all types of steady-state bifurcation, including saddle-node bifurcations.

[12] The term "saddle-node bifurcation" is natural for two-dimensional problems like (8.20). It is not natural, but is still used, for one-dimensional problems like $x' = -x^2 + \mu$ (cf. Table 8.1), as well as for problems with more than two variables.

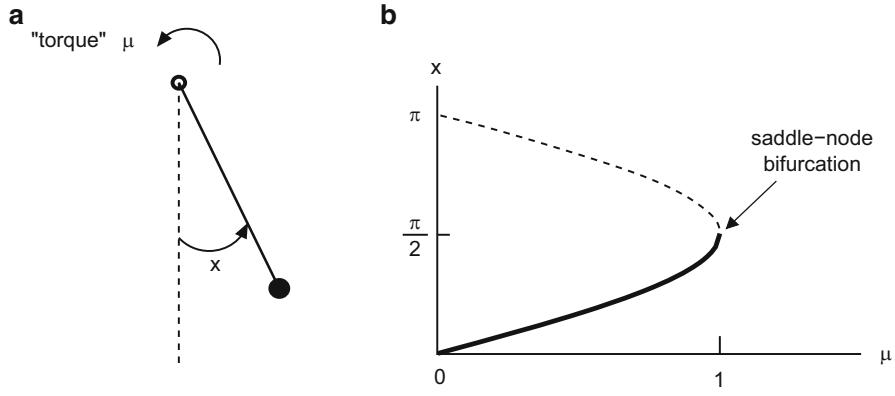

Figure 8.5: *(a) A torqued pendulum, described nondimensionally by* (8.20)*. (b) Saddle-node bifurcation in the torqued pendulum* (8.20) *at* $\mu = 1$*.*

$(0, \pi/2)$, but when μ passes 1, the system evolves to states far removed from this equilibrium. Specifically, it converges to the periodic solution that was analyzed in Section 7.4. This behavior seems obvious on physical grounds—when the torque is too large, the pendulum will rotate indefinitely.

8.4.2 Activator–Inhibitor Systems

We may use bifurcation-theory language to reinterpret our calculations in Section 6.3.1 with the activator–inhibitor system,

$$\begin{aligned} (a) \quad x' &= \sigma x^2/(1+y) - x, \\ (b) \quad y' &= \rho\,[x^2 - y]. \end{aligned} \tag{8.21}$$

As we saw there, if $\sigma > 2$, then (8.21) has two nontrivial equilibria, say $\mathbf{P}_\pm = (x_\pm, y_\pm)$, where $x_\pm$ satisfies $x^2 - \sigma x + 1 = 0$ and $y_\pm = x_\pm^2$. As indicated in the bifurcation diagram, Figure 8.6(a), when σ decreases through $\sigma_* = 2$, these two solutions undergo a saddle-node bifurcation; $\mathbf{P}_-$ is always a saddle point, and for σ near σ_*, $\mathbf{P}_+$ is a node; $\mathbf{P}_+$ may be either stable or unstable, according as $\rho > 1$ or $\rho < 1$, respectively.

If $\rho < 1$, no significant change in behavior derives from the bifurcation at $\sigma = 2$; specifically, if $\sigma < 2$, then all solutions of (8.21) in the physical domain decay to the unique equilibrium $x = y = 0$, while if $\sigma > 2$, then as illustrated in Figure 6.12(b), *virtually* all solutions of (8.21) decay to $x = y = 0$, i.e., solutions not lying on the stable manifold of $\mathbf{P}_-$. By contrast, suppose that $\rho > 1$ and that starting from $\sigma > 2$, the latter parameter is decreased quasistatically. If initially the system is in the (stable) top equilibrium $\mathbf{P}_+$, the solution follows this equilibrium branch until the

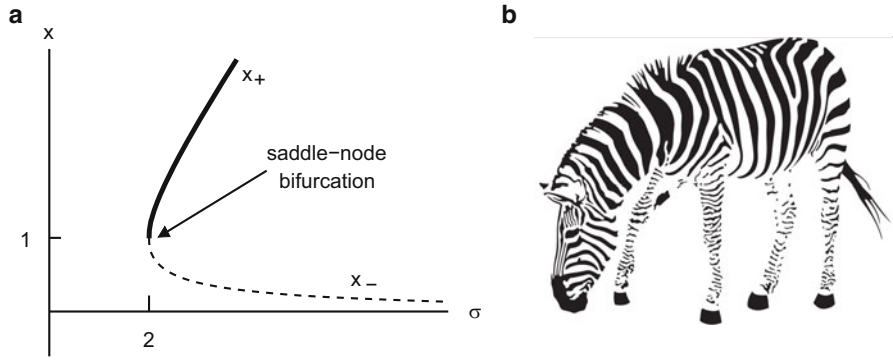

Figure 8.6: *(a) Saddle-node bifurcation in nonzero solutions of the activator–inhibitor system (8.21) at $\sigma = 2$. Although the upper branch is shown as a solid curve (indicating stability), in fact it is stable only if $\rho > 1$. (b) How did the zebra get its stripes? Some authors believe that the pattern in zebras' and other animals' coats is laid down in the developing embryo by the Turing instability in morphogen concentrations. (Cf. Chapters 2 and 3 of [58].) In the full-scale Turing instability, an activator–inhibitor system is the reaction part of a reaction–diffusion PDE. Although such problems are outside the scope of this book, in Section 8.6.1 we analyze bifurcation in a toy model of this instability, which involves just ODEs. (Cartoon drawn by Jeff Poe, 2016. Copyright 2016 by D.G. Schaeffer and J.W. Cain.)*

bifurcation point at $\sigma = 2$ is crossed, and then it "collapses" to the state $x = y = 0$.

8.5 The Lyapunov–Schmidt Reduction

8.5.1 Bare Bones of the Reduction

With the Lyapunov–Schmidt reduction, one may greatly reduce the number of variables in calculations of steady-state bifurcation. Indeed, in all the above examples, the reduced problem has only one state variable (plus of course the various parameters). For instance, recall how we analyzed the bifurcation at $\rho = 1$ in the Lorenz equations (8.5): first we solved (8.5a) to obtain $x = y$; using this, we solved (8.5c) to obtain $z = y^2/\beta$; and we finally substituted into (8.5b), yielding (8.6). This one-dimensional equation is the relation graphed in Figure 8.2(b). In general, the reduction proceeds in pretty much the same way. Regrettably, the notation needed for the general case obscures a simple idea.

To begin, we explore the hypotheses needed for the reduction. Consider a one-parameter family of ODEs

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}, \mu) \tag{8.22}$$

that for $\mu = \mu_*$ has an equilibrium solution $\mathbf{x} = \mathbf{x}_*$. If $\mathbf{x}_*$ were hyperbolic, or even if the Jacobian matrix $\mathbf{DF}_*$ merely satisfied

$$\det \mathbf{DF}_* \neq 0, \tag{8.23}$$

then by the implicit function theorem, we could solve the equilibrium equation

$$\mathbf{F}(\mathbf{x}, \mu) = \mathbf{0} \tag{8.24}$$

uniquely near $(\mathbf{x}_*, \mu_*)$ for $\mathbf{x}$ as a smooth function of μ; i.e., *no bifurcation occurs.*[13] Thus, to have bifurcation, (8.23) must fail.

The minimal failure of (8.23) occurs if zero is a simple eigenvalue of $\mathbf{DF}_*$; in symbols,[14]

$$\lambda_1(\mathbf{DF}_*) = 0, \qquad \lambda_j(\mathbf{DF}_*) \neq 0,\ j = 2, \ldots, d. \tag{8.25}$$

Under this hypothesis, the Lyapunov–Schmidt technique may be used to *reduce* (8.24) *to a single scalar equation.*[15]

To proceed with the reduction, let us choose coordinates to simplify the Jacobian $\mathbf{DF}_*$. By Exercise 4, equation (8.25) implies that $\dim \ker \mathbf{DF}_* = 1$ and $\ker \mathbf{DF}_* \cap \operatorname{range} \mathbf{DF}_* = \{\mathbf{0}\}$. Therefore, without loss of generality we may make a linear change of variables on $\mathbb{R}^d$ so that

$$\begin{array}{ll} \text{(a)} & \mathbf{e}_1 \text{ is a null eigenvector of } \mathbf{DF}_*, \\ \text{(b)} & \text{the range of } \mathbf{DF}_* \text{ is spanned by } \mathbf{e}_2, \ldots, \mathbf{e}_d, \end{array} \tag{8.26}$$

where $\mathbf{e}_j$ is the jth coordinate vector in the new coordinate system. We write the Jacobian in block notation based on these coordinates,

$$\mathbf{DF} = \begin{bmatrix} \partial F_1/\partial x_1 & \tilde{\nabla} F_1 \\ \partial \tilde{\mathbf{F}}/\partial x_1 & \tilde{\mathbf{D}}\tilde{\mathbf{F}} \end{bmatrix}, \tag{8.27}$$

where tilde denotes a vector quantity whose first component has been deleted. According to (8.26), at $(\mathbf{x}_*, \mu_*)$ this simplifies to

$$\mathbf{DF}_* = \begin{bmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \tilde{\mathbf{D}}\tilde{\mathbf{F}}_* \end{bmatrix},$$

[13] More accurately, no steady-state bifurcation occurs; (8.23) is typically satisfied at a Hopf bifurcation point.

[14] This equation repeats (8.10), but with a difference: here we make no assumption about equilibria of the ODE for $\mu \neq \mu_*$.

[15] In general, a bifurcation problem may be reduced to a system with n variables, where n equals the dimension of $\ker \mathbf{DF}_*$; see Chapter VII of [31].

where by (8.25), the submatrix $\tilde{\mathbf{D}}\tilde{\mathbf{F}}_*$ is nonsingular. We apply the implicit function theorem to conclude that the last $d-1$ equations of (8.24),

$$\begin{aligned} F_2(x_1, x_2, \ldots, x_d, \mu) &= 0, \\ \ldots \qquad & \qquad \ldots \\ F_d(x_1, x_2, \ldots, x_d, \mu) &= 0\,, \end{aligned} \tag{8.28}$$

may be solved near $(\mathbf{x}_*, \mu_*)$ for $\mathbf{y} = (x_2, \ldots, x_d)$ as functions of x_1 and μ, say

$$\mathbf{y} = \mathbf{Y}(x_1, \mu).$$

Define the scalar function $g(x_1, \mu)$ by substituting this formula for $\mathbf{y}$ into the first component of $\mathbf{F}$:

$$g(x_1, \mu) = F_1(x_1, \mathbf{Y}(x_1, \mu), \mu).$$

Then under the above hypotheses, *there is a one-to-one correspondence between solutions near* $(\mathbf{x}_*, \mu_*)$ *of the full problem* (8.24) *and the reduced problem*

$$g(x_1, \mu) = 0. \tag{8.29}$$

Specifically, if (x_1, μ) is a solution of (8.29), then $(x_1, \mathbf{Y}(x_1, \mu), \mu)$ is a solution of (8.24), and every solution of (8.24) near $(\mathbf{x}_*, \mu_*)$ arises in this way. This great reduction of the problem results from merely processing the equations in (8.24) sequentially, handling the nonsingular part first.

We call the above reduction, based on choosing coordinates to obtain (8.26), a *standard* reduction. It is useful for analytical purposes, but in specific problems it may not be the most convenient choice.[16] For example, to reduce the equilibrium equations for the chemostat (8.18) to a single equation, it is simplest to add the two equations to deduce that $\rho(x + y - \sigma) = 0$, solve this relation for x, and substitute into the first equation in (8.18) to obtain

$$(\sigma - y)\left(\frac{y}{y+1} - \rho\right) = 0. \tag{8.30}$$

In general, there is great flexibility in choosing which component of $\mathbf{x}$ to retain in the reduced equation and in choosing equations for eliminating the other $d-1$ unknowns. We don't introduce notation to formalize this procedure in all possible cases, because that would obscure more than it would clarify.

[16] Even the reduction of the Lorenz equation in Section 8.1.2, which we used to motivate the Lyapunov–Schmidt reduction, does not quite follow the standard procedure; cf. Exercise 5.

8.5.2 Proof of Theorem 8.2.2

The proof of the theorem is based on the implicit function theorem. All the effort goes into understanding the structure of the equations to see how this theorem may be applied. But first, we must address a few matters.

(a) Some unfinished business:

Proof of Lemma 8.2.1. The eigenvalues of $\mathbf{DF}(\mathbf{x}_{\rm eq}(\mu),\mu)$ satisfy the implicit equation

$$f(\lambda,\mu) = \det\big(\mathbf{DF}(\mathbf{x}_{\rm eq}(\mu),\mu) - \lambda I\big) = 0. \tag{8.31}$$

Fixing $\mu = \mu_*$, we have

$$f(\lambda,\mu_*) = (\lambda_1(\mathbf{DF}_*) - \lambda)(\lambda_2(\mathbf{DF}_*) - \lambda)\ \ldots\ (\lambda_d(\mathbf{DF}_*) - \lambda).$$

Therefore, by (8.10), we have $f(0,\mu_*) = 0$ and

$$\frac{\partial f}{\partial \lambda}(0,\mu_*) = -\lambda_2(\mathbf{DF}_*)\lambda_3(\mathbf{DF}_*)\ \ldots\ \lambda_d(\mathbf{DF}_*) \neq 0. \tag{8.32}$$

Thus, the lemma follows from the implicit function theorem. □

We also need to relate (8.11) and (8.12). By implicit differentiation of (8.31),

$$\frac{d\lambda}{d\mu}(\mu_*) = -\frac{\partial_\mu f(0,\mu_*)}{\partial_\lambda f(0,\mu_*)}, \tag{8.33}$$

where by (8.32) the denominator is nonzero. Regarding the numerator, we see from (8.31) that

$$\partial_\mu f(0,\mu_*) = \frac{\partial}{\partial\mu}\det\mathbf{DF}(\mathbf{x}_{\rm eq}(\mu),\mu)\Big|_{\mu=\mu_*}.$$

Thus, it follows from (8.33) that (8.11) and (8.12) are equivalent, as claimed.

(b) Setting up the proof:

We collect our assumptions and notation from above, together with various reductions that simplify the proof. By introducing $\mathbf{x}-\mathbf{x}_{\rm eq}(\mu)$ as a new unknown, we may assume without loss of generality that the trivial solution $\mathbf{x}_{\rm eq}(\mu)$ vanishes identically, in symbols,

$$\mathbf{F}(\mathbf{0},\mu) \equiv \mathbf{0}. \tag{8.34}$$

We translate the bifurcation point to $\mu = 0$ and rewrite hypothesis (8.11),

$$\frac{d\lambda}{d\mu}(0) \neq 0. \tag{8.35}$$

Making a linear change of coordinates as above, we may assume that at $(\mathbf{x}, \mu) = (\mathbf{0}, 0)$,

$$\mathbf{DF}(\mathbf{0}, 0) = \begin{bmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \tilde{\mathbf{D}}\tilde{\mathbf{F}}(\mathbf{0}, 0) \end{bmatrix}, \tag{8.36}$$

where $\tilde{\mathbf{D}}\tilde{\mathbf{F}}(\mathbf{0}, 0)$ is nonsingular. In the Lyapunov–Schmidt reduction of (8.9) we solve[17]

$$\tilde{\mathbf{F}}(x, \mathbf{Y}(x, \mu), \mu) = \mathbf{0} \tag{8.37}$$

for the $(d-1)$-dimensional vector $\mathbf{Y}(x, \mu)$ on an appropriate neighborhood of $(0, 0)$, and we define the reduced function

$$g(x, \mu) = F_1(x, \mathbf{Y}(x, \mu), \mu). \tag{8.38}$$

Both $\mathbf{Y}$ and g are as smooth as $\mathbf{F}$, i.e, at least $\mathcal{C}^3$.

(c) Key facts about the reduced function:

Because of (8.34), we have

$$g(0, \mu) \equiv 0, \tag{8.39}$$

which lets us define the quotient $g(x, \mu)/x$ for use below.

Lemma 8.5.1. *The function $h(x, \mu) = g(x, \mu)/x$, defined by $h(0, \mu) = \partial_x g(0, \mu)$ when $x = 0$, is $\mathcal{C}^2$ in a neighborhood of $(0, 0) \in \mathbb{R}^2$.*

Proof. The fundamental theorem of calculus gives us

$$g(x, \mu) = g(0, \mu) + \int_0^x \partial_x g(x', \mu)\, dx'.$$

By (8.39), the first term vanishes. Reparametrizing the integral with $s = x'/x$, we calculate that

$$h(x, \mu) = \int_0^1 \partial_x g(sx, \mu)\, ds,$$

which is indeed a $\mathcal{C}^2$ function, even near $x = 0$. □

Lemma 8.5.2.

$$(a)\ \frac{\partial g}{\partial x}(0, 0) = 0 \qquad \textit{and} \qquad (b)\ \frac{\partial^2 g}{\partial \mu \partial x}(0, 0) = \frac{d\lambda}{d\mu}(0). \tag{8.40}$$

Proof. For the first derivative, the chain rule applied to (8.38) yields

$$\partial_x g(x, \mu) = \partial_1 F_1(x, \mathbf{Y}(x, \mu), \mu) + \tilde{\nabla} F_1(x, \mathbf{Y}(x, \mu), \mu) \cdot \partial_x \mathbf{Y}(x, \mu), \tag{8.41}$$

and this vanishes when $(x, \mu) = (0, 0)$, because the first row of (8.36) is zero.

[17] We'll write x instead of x_1 to cut down clutter in some upcoming equations.

To derive (8.40b), we insert an intermediate quantity and show that

$$\frac{\partial^2 g}{\partial \mu \partial x}(0,0) = \frac{\partial^2 F_1}{\partial \mu \partial x_1}(\mathbf{0},0) = \frac{d\lambda}{d\mu}(0). \tag{8.42}$$

Regarding the first equality, it may seem impossible that $\partial_\mu \partial_x g(0,0)$ could reduce to something as simple as (8.42): the x-derivative (8.41) already has two terms, and taking the μ-derivative will generate several more. However, because many terms vanish at the origin, only the term $\partial_\mu \partial_1 F_1(\mathbf{0},0)$ survives, to contribute to $\partial_\mu \partial_x g(0,0)$, which we prove with the following three claims. The first claim addresses derivatives of $\mathbf{Y}$; the second two address the μ-derivatives of the two terms in (8.41). Combining (8.41) with Claims 2 and 3, we verify the desired equation for $\partial_\mu \partial_x g(0,0)$.

Claim 1:

$$\partial_x \mathbf{Y}(0,0) = \partial_\mu \mathbf{Y}(0,0) = 0. \tag{8.43}$$

Proof. Taking derivatives of (8.37) with respect to x and μ, using the chain rule, we calculate

$$\partial_1 \tilde{\mathbf{F}} + \tilde{\mathbf{D}}\tilde{\mathbf{F}} \cdot \partial_x \mathbf{Y} = 0 \quad \text{and} \quad \tilde{\mathbf{D}}\tilde{\mathbf{F}} \cdot \partial_\mu \mathbf{Y} + \partial_\mu \tilde{\mathbf{F}} = 0.$$

We see from (8.36) and (8.34) that $\partial_1 \tilde{\mathbf{F}}(\mathbf{0},0) = 0$ and $\partial_\mu \tilde{\mathbf{F}}(\mathbf{0},0) = 0$, respectively, and of course $\tilde{\mathbf{D}}\tilde{\mathbf{F}}(\mathbf{0},0)$ is invertible. This proves the claim.

Claim 2:

$$\frac{\partial}{\partial \mu} \tilde{\nabla} F_1(0, \mathbf{Y}(0,\mu), \mu) \cdot \partial_x \mathbf{Y}(0,\mu)\Big|_{\mu=0} = 0.$$

Proof. We calculate the derivative with Leibniz's product rule, using the chain rule when differentiating $\tilde{\nabla} F_1(0, \mathbf{Y}(0,\mu), \mu)$. This yields a total of three terms. However, two of these terms have a factor $\partial_x \mathbf{Y}$, and the third has a factor $\tilde{\nabla} F_1$, so all three terms vanish at the origin, which proves the claim.

Claim 3:

$$\frac{\partial}{\partial \mu} \partial_1 F_1(0, \mathbf{Y}(0,\mu), \mu)\Big|_{\mu=0} = \partial_\mu \partial_1 F_1(0, \mathbf{0}, \mu). \tag{8.44}$$

Proof: We have from the chain rule that the LHS of (8.44) equals

$$\tilde{\nabla}[\partial_1 F_1](0, \mathbf{Y}(0,0), 0) \cdot \partial_\mu \mathbf{Y}(0,0) + \partial_\mu \partial_1 F_1(0, \mathbf{Y}(0,0), 0).$$

By (8.43), the factor $\partial_\mu \mathbf{Y}(0,0)$ in the first term of this formula vanishes, yielding Claim 3.

It remains to prove the second equality in (8.42), for which we derive alternative representations for the numerator and denominator in (8.33). Regarding the denominator, in the notation of (8.36), we may rewrite equation (8.32) as

$$\partial_\lambda f(0,0) = -\det \tilde{\mathbf{D}}\tilde{\mathbf{F}}(\mathbf{0},0).$$

Regarding the numerator, consider the expansion of $\mathbf{DF}(\mathbf{0},\mu)$ in a Taylor series in μ. Using (8.36) to evaluate the zeroth-order terms in the expansion, we have

$$\mathbf{DF}(\mathbf{0},\mu) = \begin{bmatrix} \partial_\mu(\partial F_1/\partial x_1)(\mathbf{0},0)\cdot\mu + \mathcal{O}(\mu^2) & \mathcal{O}(\mu) \\ \mathcal{O}(\mu) & \tilde{\mathbf{D}}\tilde{\mathbf{F}}(\mathbf{0},0) + \mathcal{O}(\mu) \end{bmatrix},$$

where we write out first-order terms only in the $(1,1)$-entry of the matrix. Taking the determinant of this matrix, we find that

$$\det \mathbf{DF}(\mathbf{0},\mu) = \partial_\mu(\partial F_1/\partial x_1)(\mathbf{0},0)\cdot \det \tilde{\mathbf{D}}\tilde{\mathbf{F}}(\mathbf{0},0)\cdot\mu + \mathcal{O}(\mu^2).$$

Thus,

$$\partial_\mu f(0,0) = \frac{\partial}{\partial\mu}\det \mathbf{DF}(\mathbf{0},\mu)\Big|_{\mu=0} = \partial_\mu(\partial F_1/\partial x_1)(\mathbf{0},0)\cdot \det \tilde{\mathbf{D}}\tilde{\mathbf{F}}(\mathbf{0},0).$$

The second equality in (8.42) follows on substituting these formulas for $\partial_\lambda f(0,0)$ and $\partial_\mu f(0,0)$ into (8.33). □

(d) Proof of Theorem 8.2.2:

The theorem asserts that the equilibria of (8.9) near the bifurcation point lie on two crossed curves in $\mathbb{R}^d \times \mathbb{R}$. By the Lyapunov–Schmidt reduction, it suffices to show that the zero set of the reduced function

$$\{(x,\mu) \in \mathbb{R}\times\mathbb{R} : g(x,\mu) = 0\} \tag{8.45}$$

has this structure. Now, by (8.39),

$$g(x,\mu) = 0 \quad \text{iff} \quad x = 0 \ \text{ or } \ h(x,\mu) = 0,$$

where $h(x,\mu)$ is defined in Lemma 8.5.1. The line $\{x = 0\}$ is one curve contained in the set (8.45). We claim that the zero set $\{h(x,\mu) = 0\}$ is also a smooth curve through the origin. Note that $h(0,0) = \partial_x g(0,0) = 0$, but

$$\partial_\mu h(0,0) = \partial_\mu\partial_x g(0,0) = \frac{d\lambda}{d\mu}(0) \neq 0,$$

where we have used Lemma 8.5.2 and invoked hypothesis (8.35). Thus, by the implicit function theorem, the equation $h(x,\mu) = 0$ may be solved for μ as a function of x, which proves the claim and shows that (8.45) consists of two crossed curves.

Vanishing der.	Normal form for g	Name	Example
(none)	$x' = -x^2 + \mu$	saddle-node	Activator–inhibitor, (8.21)
$\partial_\mu g = 0$	$x' = -x^2 + \mu x$	transcritical	Lotka–Volterra, (8.16)
$\partial_\mu g = 0$	$x' = -x^2 - \mu^2$	isola center	CSTR with bath, (8.70)
$\partial_{xx} g = 0$	$x' = \pm x^3 + \mu$	hysteresis point	CSTR without bath, (8.68)
$\partial_{xx} g = 0, \partial_\mu g = 0$	$x' = \pm x^3 + \mu x$	pitchfork	Lorenz equations, (8.5)

Table 8.1: *Partial classification of one-dimensional bifurcation problems, based on enumeration of derivatives of the reduced function, beyond* (8.47)*, that vanish. Section I.3 of [31] gives general formulas for calculating these derivatives.*

Regarding parametrization of the bifurcating solutions, define $\mu(a)$ by solving the equation $h(a, \mu) = 0$ for μ. Then every nontrivial solution of $g(x, \mu) = 0$ has the form $(a, \mu(a))$, and

$$\mathbf{X}(a) = (a, \mathbf{Y}(a, \mu(a))) \tag{8.46}$$

parametrizes the nontrivial equilibria of (8.9). Recalling the various reductions above, i.e., $\mathbf{x}_{\mathrm{eq}}(\mu(a)) \equiv \mathbf{0}$, the kernel of $\mathbf{DF}_*$ is spanned by $\mathbf{v} = (1, \mathbf{0})$, and $\mathbf{Y}(a, \mu(a)) = \mathcal{O}(a^2)$, we see that (8.13) and (8.46) are equivalent. □

8.5.3 One-Dimensional Bifurcation Problems

With the Lyapunov–Schmidt reduction, we may introduce a partial hierarchy of problems with steady-state bifurcation from a simple eigenvalue. Without loss of generality, we can translate coordinates so that the bifurcation point is located at $\mathbf{x} = \mathbf{0}$, $\mu = 0$. The reduced function $g(x, \mu)$, defined near $(0, 0)$, then satisfies

$$g(0, 0) = \partial_x g(0, 0) = 0. \tag{8.47}$$

Bifurcation problems that satisfy the simple-eigenvalue hypothesis (8.25) can be roughly classified by the number of derivatives of the reduced function g, beyond (8.47), that vanish. A few of the simplest cases are listed in Table 8.1; see Chapter IV of [31] for more detail. The phrase *normal form*, which appears in the table, refers to a particularly simple version of a bifurcation problem that captures the essential behavior of a class of problems. (We shall use this phrase informally, shying away from a precise technical definition.)

To gain intuition, let us discuss the construction of normal forms for pitchfork bifurcations. Consider a degenerate equilibrium (i.e., satisfying (8.47)) such that in

addition $\partial_\mu g(0,0) = \partial_{xx} g(0,0) = 0$. Then modulo higher-order terms,[18] we have the reduced ODE

$$x' = Ax^3 + B\mu x. \tag{8.48}$$

Suppose the coefficients A and B are nonzero. If we rescale $\overline{x} = |A|^{1/2}x$ and $\overline{\mu} = |B|\mu$, then we may reduce this equation to either

$$\text{(a) } \overline{x}' = -\overline{x}^3 \pm \overline{\mu}\,\overline{x} \quad \text{or} \quad \text{(b) } \overline{x}' = \overline{x}^3 \pm \overline{\mu}\,\overline{x}, \tag{8.49}$$

according as $A < 0$ or $A > 0$, respectively. What difference do the $\pm$ choices in (8.49), which come from B, make? Considering, for example, (8.49a), if we choose $+\overline{\mu}\,\overline{x}$, corresponding to $B > 0$ in (8.48), the trivial solution is stable for $\overline{\mu} < 0$ and unstable for $\overline{\mu} > 0$, while if we choose $-\overline{\mu}\,\overline{x}$, it is the other way around. However, in both cases, *as the bifurcation parameter crosses zero* in one direction or the other, *the trivial solution loses stability, and simultaneously two new, stable, solutions bifurcate*, behavior that we called supercritical in Section 8.1. The only difference between the two cases is the reversal of the orientation of the bifurcation parameter, a difference that does not seem important to us. Therefore, as is conventional, we collapse these two cases into the one case $-x^3 + \mu x$ in the table by orienting the bifurcation parameter so that the trivial solution loses stability as μ is *increased*.[19]

Unlike B, the sign of A in (8.48) impacts the qualitative nature of the bifurcation. Specifically, the case $A < 0$ leads to (8.49a), for which the bifurcation is supercritical, while $A > 0$ leads to (8.49b), which is subcritical.

We have not yet encountered isola-center or hysteresis-point bifurcations, but we will discuss these in Section 8.5.6. In the meanwhile, let's compare one of them, the normal form for isola-center bifurcation, with the normal form for transcritical bifurcation, purely as mathematical objects divorced from any applications. The defining relations for both bifurcations are the same. In geometric terms, the zero set $\{-x^2 + \mu x = 0\}$ for transcritical bifurcation consists of two crossed lines, while the zero set $\{-x^2 - \mu^2 = 0\}$ for isola-center bifurcation consists of a single point. In algebraic terms, both normal forms are quadratic forms $Q(x,\mu)$, a (negative) definite form for isola-center bifurcation and an indefinite form for transcritical bifurcation.[20]

Incidentally, note that the different behaviors between the various bifurcations in Table 8.1 are determined by the higher-order terms at the bifurcation point, which reminds us of the importance of such terms at a nonhyperbolic equilibrium.

---

[18] In fact, the normal forms typically still apply even when higher-order terms are included; cf. Section II.9 of [31].

[19] This convention is natural, because in applications it is more common for instability to appear as the parameters in the problem, which usually measure forcing, are increased.

[20] Incidentally, our choice of $-x^2 + \mu x$ for transcritical bifurcation is not important; any of the choices $\pm x^2 \pm \mu x$ would work equally well. For that matter, $\pm(x^2 - \mu^2)$ are also possible, although we prefer a normal form in which $x = 0$ is a trivial solution branch.

8.5.4 Exchange of Stability

Let's state the main result straightaway and explore context later. The principle of exchange of stability deduces stability information for a higher-dimensional bifurcation problem

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}, \mu), \tag{8.50}$$

based on information about the reduced equation. For this theorem, hypothesis (8.25) must be strengthened to require that the $d-1$ nonzero eigenvalues of $\mathbf{DF}_*$ lie in the left half-plane, i.e.,

$$\lambda_1(\mathbf{DF}_*) = 0, \qquad \Re\lambda_j(\mathbf{DF}_*) < 0,\ j = 2, \dots, d. \tag{8.51}$$

For the proof, which expands on the less-than-inspiring calculations used to prove Lemma 8.5.2, we refer you to Section 1.4 of [31].

Theorem 8.5.3. *Let $g(x, \mu)$ be the reduced function from a standard Lyapunov–Schmidt reduction of (8.50) at a bifurcation point where (8.51) is satisfied. Then in an appropriately small neighborhood of $(\mathbf{x}_*, \mu_*)$, an equilibrium $(\mathbf{x}_0, \mu_0) \in \mathbb{R}^d \times \mathbb{R}$ of (8.50) is stable or unstable if at the corresponding equilibrium $(x_0, \mu_0) \in \mathbb{R} \times \mathbb{R}$ of the reduced problem, $\partial_x g(x_0, \mu_0)$ is negative or positive, respectively.*

Let us elaborate: At an equilibrium of (8.50) near the bifurcation point, $d-1$ of the eigenvalues of the Jacobian are safely in the left half-plane. Thus, the stability of the equilibrium is determined by the remaining eigenvalue, and according to the theorem, this has the same sign as $\partial_x g$.

It may be illuminating to relate the sign of $\partial_x g(x_0, \mu_0)$ to a one-dimensional made-up ODE using the reduced function,

$$x' = g(x, \mu). \tag{8.52}$$

Then (x_0, μ_0), an equilibrium of (8.52), is stable or unstable according as $\partial_x g(x_0, \mu_0)$ is negative or positive, respectively. In other words, we could rephrase the theorem to state, more intuitively, that an equilibrium $(\mathbf{x}_0, \mu_0)$ of the full problem has the same stability as the corresponding equilibrium of (8.52).

In many specific bifurcation problems we can apply the theorem without ever calculating a derivative of $g(x, \mu)$ if we think pictorially. Since (8.52) is one-dimensional, its flow direction can reverse itself only where the RHS vanishes, i.e., only on the bifurcation diagram. For example, suppose (8.52) represents a transcritical bifurcation,

$$x' = -x^2 + \mu x.$$

As shown in Figure 8.7(a), the bifurcation diagram $\{-x^2 + \mu x = 0\}$ divides the x, μ-plane into four regions. Within each region, the flow direction cannot change.

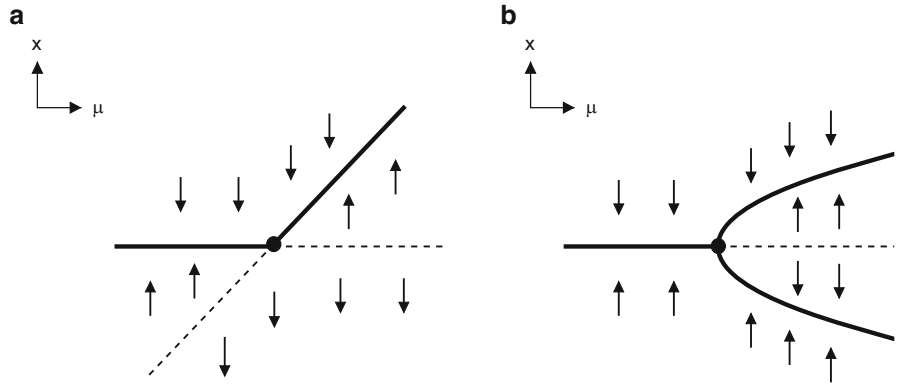

Figure 8.7: *Flow directions adjacent to (a) a transcritical bifurcation* $x' = -x^2 + \mu x$ *and (b) a supercritical pitchfork bifurcation* $x' = -x^3 + \mu x$*. In both panels, the bold dot is located at the bifurcation point,* $x = \mu = 0$*.*

Thus, without considering any derivatives, we see from flow arrows in the figure that the trivial solution $x = 0$ is stable for $\mu < 0$ and unstable for $\mu > 0$, while these stabilities are reversed for the bifurcating solutions $x = \mu$. In Exercise 6 we ask you to use this approach to determine the stabilities of the equilibria on both branches of (8.30), which was derived from the chemostat. Since this equation was obtained from (8.18) by a nonstandard reduction, you will need to be aware of the following minor pitfall.

Addendum to Theorem 8.5.3. In a *nonstandard* reduction of a bifurcation problem, it may happen that stabilities computed from the sign of $\partial_x g$ *are exactly reversed*, i.e., stable if $\partial_x g(x_0, \mu_0) > 0$ and unstable if $\partial_x g(x_0, \mu_0) < 0$.

Incidentally, the phrase *exchange of stability* is intended to describe the transfer of stability, as in Figure 8.7(a), from the trivial solution to the bifurcating solution when they cross one another. This phrase is natural for transcritical bifurcation. Even though it makes less sense in other cases, it is still used to describe the general result, Theorem 8.5.3.

8.5.5 Symmetry and the Pitchfork Bifurcation

By a *reflection* on $\mathbb{R}^d$ we mean a linear map $R : \mathbb{R}^d \to \mathbb{R}^d$, different from the identity, such that $R^2 = I$. We say that an ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$ is *symmetric*[21] under R if

$$\mathbf{F}(R\mathbf{x}) = R\mathbf{F}(\mathbf{x}) \quad \text{for all } \mathbf{x}. \tag{8.53}$$

[21]The technical term is *equivariant*.

In this section we explore a connection between pitchfork bifurcations and reflectional symmetry. To make the discussion more meaningful, for each of the three examples of pitchfork bifurcation in Section 8.1 you should identify the reflectional symmetry in the problem and consider Theorem 8.5.5 for the specific example.

Lemma 8.5.4. *If $R^2 = I$, then the only possible eigenvalues of R are ± 1, and R is diagonalizable.*

We ask you to derive this bit of linear algebra in Exercise 4.

Theorem 8.5.5. *In addition to the hypotheses of Theorem 8.2.2, assume that for all μ, $\mathbf{F}(\mathbf{x}, \mu)$ is symmetric under the reflection R and that $R\,\mathbf{x}_{\rm eq}(\mu) = \mathbf{x}_{\rm eq}(\mu)$. If the null eigenvector $\mathbf{v}$ of $\mathbf{DF}_*$ at the bifurcation point satisfies $R\,\mathbf{v} = -\mathbf{v}$, then the reduced function $g(x, \mu)$ obtained in the standard Lyapunov–Schmidt reduction is odd with respect to x:*

$$g(-x, \mu) = -g(x, \mu). \tag{8.54}$$

If g is odd, we deduce that

$$\partial_{xx} g(0, 0) = \partial_{\mu} g(0, 0) = 0.$$

Thus, a bifurcation satisfying the hypotheses of the theorem is a candidate for a pitchfork. (To show that it is actually a pitchfork we would need to check that $\partial_{xxx} g(0, 0) \neq 0$ and $\partial_{\mu x} g(0, 0) \neq 0$; i.e., it is not more singular.)

Remarks: (i) It is not difficult to prove this result. If you want guidance, you may consult the proof of Proposition 3.3 on p. 306 in [31], which characterizes the analogous behavior in a far more general context. In fact, any sensible reduction, whether standard or not, will give a reduced function $g(x, \mu)$ that is odd in x. *(ii)* In words, we describe the hypothesis that $R\,\mathbf{v} = -\mathbf{v}$ by saying that the *bifurcation breaks the symmetry*. Incidentally, if $R\,\mathbf{v} = +\mathbf{v}$, the other possibility, symmetry has no implications for the bifurcation (cf. Exercise 11).

8.5.6 Additional Parameters in Bifurcation Problems

(a) Imperfect bifurcation. ODEs represent an idealized description of some physical system, but real systems will differ from the idealized description in myriad ways that are impossible to enumerate. As an example, in the bead equation (8.2), suppose the axis of rotation of the ring is very slightly off-center, say displaced by δ as in Figure 8.8(a). In this case, the radius of the rotating motion is slightly altered, from $\ell \sin x$ to $\ell \sin x + \delta$. This perturbation changes the nondimensional equations of motion (8.2) to read

$$\begin{aligned} x' &= y, \\ y' &= -\beta y - \sin x + \mu(\sin x + \varepsilon)\cos x, \end{aligned} \tag{8.55}$$

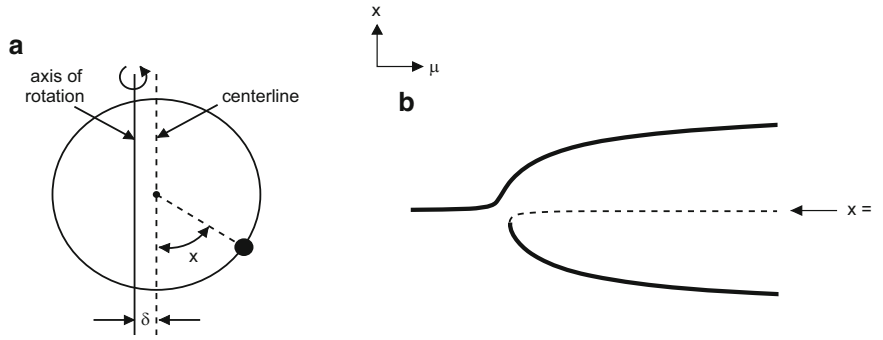

Figure 8.8: *(a) Bead on a hoop rotating about a vertical axis displaced from the centerline of the hoop: an example of imperfect bifurcation. This system is governed by (8.55). (b) Perturbed pitchfork bifurcation diagram for (8.55), with $\varepsilon > 0$. If ε is small, the saddle-node bifurcation is located near the unperturbed bifurcation point at $\mu = 1$.*

where $\varepsilon = \delta/\ell$. When $\varepsilon \neq 0$, the perturbation splits a bifurcation diagram, which is connected, into two separate pieces, as illustrated[22] in Figure 8.8(b) for $\varepsilon > 0$. To interpret Figure 8.8(b) in physical terms, if $\varepsilon > 0$ is fixed and μ is increased quasistatically from zero, then the equilibrium of the bead will evolve smoothly from an equilibrium near the bottom of the loop (i.e., $x = 0$) to equilibria with $x > 0$. In other words, making ε nonzero removes the indeterminacy of the idealized, perfectly symmetric, problem. (If $\varepsilon < 0$, the unique solution for small μ connects to the *lower* solution branch at large μ.)

Note in Figure 8.8(b) that even with $\varepsilon \neq 0$, for μ sufficiently large, both nontrivial equilibria exist and are stable. Only one of them can be reached, starting from small μ, by quasistatic increase of the bifurcation parameter. Given an appropriately large value of μ, a finite (noninfinitesimal) perturbation could kick the system from either equilibrium onto the other. However, when μ is decreased quasistatically, if the system is on the "wrong" solution branch, it will undergo a sudden jump when μ

[22]It is not hard to believe that the perturbation (8.55) splits the pitchfork into two components, but it is less clear whether for $\varepsilon > 0$ the negative μ-axis connects to the upper branch, as shown in Figure 8.8(b), or to the lower branch. The issue may be resolved with a Taylor-series expansion about the bifurcation point $x_* = 0$, $\mu_* = 1$. Specifically, to lowest order, an equilibrium of (8.55) satisfies

$$x\,(\mu - 1) + \varepsilon = 0, \tag{8.56}$$

where we have neglected terms that are $\mathcal{O}(|x|^3)$ and $\varepsilon\mathcal{O}(|x| + |\mu - 1|)$. If $\varepsilon > 0$, the branch of the hyperbola (8.56) that is asymptotic to the negative μ-axis lies in the half-plane $\{x > 0\}$, and even with the higher-order terms, the solution of (8.55) will retain this behavior, as the figure indicates.

reaches the saddle-node bifurcation point. Incidentally, it is worth recording that (i) perturbation of a pitchfork bifurcation leads to a saddle-node bifurcation, and (ii) the latter is robust under further small perturbations.

In an attempt to model deviations of a physical problem from an idealized description, one may consider subjecting the equation to an arbitrary small perturbation. This kind of analysis goes by the name of *imperfect bifurcation*. For bifurcation from a simple eigenvalue, it suffices to consider one-dimensional equations because of the Lyapunov–Schmidt reduction. For example, perturbing in this way the normal form for a supercritical pitchfork, we obtain

$$x' = -x^3 + \mu x + \varepsilon, \tag{8.57}$$

whose equilibria have the same qualitative behavior just seen in the (two-dimensional) bead example (8.55).

(b) Unfoldings: hysteresis-point and isola-center bifurcation. Adding additional parameters to a bifurcation problem, as in (8.57), is sometimes called *unfolding*. Unfoldings clarify the significance of two of the normal forms in Table 8.1, as we now explain.

(i) *Hysteresis points:* Consider the unfolding of one of the normal forms for a hysteresis point[23],

$$x' = -x^3 + \mu + \alpha x. \tag{8.58}$$

As illustrated in Figure 8.9(b), if $\alpha > 0$, there is a range of the bifurcation parameter μ for which (8.58) has three solutions, two of them stable. If μ is fixed in the bistability interval, then generically as $t \to \infty$, a solution of (8.58) will tend to one of the stable equilibria of this equation, *but which equilibrium it tends to depends on initial conditions.*[24] By contrast, if $\alpha < 0$, then for every μ, the equation has a unique equilibrium, which moreover is stable (globally attracting, in fact) and varies smoothly with μ.

The name of this singularity derives from the phenomenon of hysteresis, which describes behavior of (8.58) for $\alpha > 0$ implicit in Figure 8.9(b): If μ is varied (quasistatically) back and forth across the bistability interval, the solution of (8.58) will jump repeatedly between the lower and upper equilibria. The term *hysteresis*

[23]What might seem like the simplest unfolding, $x' = -x^3 + \mu + \alpha$, doesn't change anything; it just yields another hysteresis-point bifurcation at a slightly shifted location. Unfolding theory, the focus of Chapter III of [31], explains what perturbations in unfolding a singular bifurcation problem make an actual difference.

[24]In Section 8.6.2 we discuss an application from chemical engineering with a hysteresis point in which such bistability can have disastrous consequences. Isola centers, which pose different risks, also appear in that application.

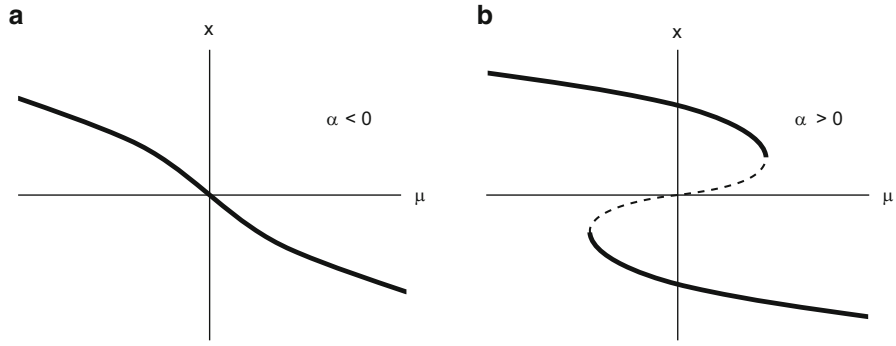

Figure 8.9: *Bifurcation diagrams for* $x' = -x^3 + \mu + \alpha x$*, an unfolding of a hysteresis-point bifurcation. Left panel: For* $\alpha < 0$*, a unique, stable equilibrium varies smoothly with* μ*. Right panel: For* $\alpha > 0$*, there is hysteresis, which includes an interval of bistability.*

refers to the fact the system jumps at different points, while μ is increasing and while it is decreasing.

(ii) *Isola center:* Consider the unfolding of the normal form for an isola center,

$$x' = -x^2 - \mu^2 + \alpha. \tag{8.59}$$

As illustrated in Figure 8.10(a), if $\alpha > 0$, then (8.59) has a "circle" of equilibria, with $x = +\sqrt{\alpha - \mu^2}$ being stable, while if $\alpha < 0$, the equation has no equilibria at all. The term *isola*, the Italian word for island, is used to describe such an isolated branch of equilibria. At the isola-center bifurcation $\alpha = 0$, the isola is reduced to a single point.

8.6 Steady-State Bifurcation in Two Applications

8.6.1 The Two-Cell Turing Instability

(a) Bifurcation from the trivial solution.

Armed with our newly acquired expertise in bifurcation theory, let us return to the Turing instability with two interacting cells, introduced in Section 6.3.2,

$$\begin{array}{lrcl} (a) & x_1' & = & \sigma x_1^2/(1+y_1) - x_1, \\ (b) & y_1' & = & \rho\,[x_1^2 - y_1] + D(y_2 - y_1), \\ (c) & x_2' & = & \sigma x_2^2/(1+y_2) - x_2, \\ (d) & y_2' & = & \rho\,[x_2^2 - y_2] + D(y_1 - y_2). \end{array} \tag{8.60}$$

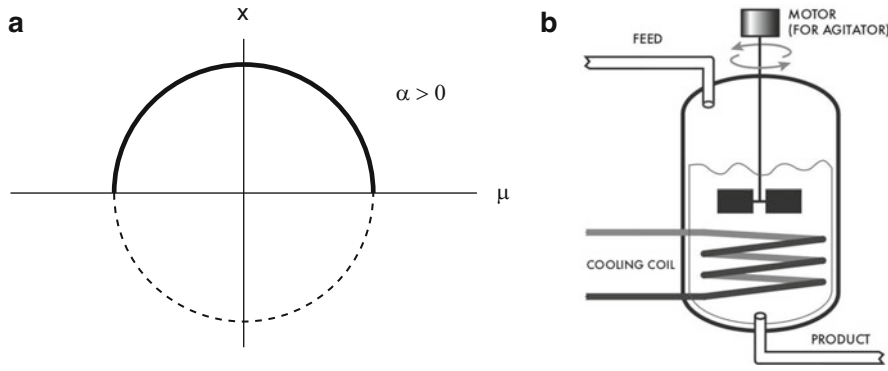

Figure 8.10: *(a) Bifurcation diagram for $x' = -x^2 - \mu^2 + \alpha$, an unfolding of an isola-center bifurcation, assuming $\alpha > 0$. (For $\alpha < 0$ there are no equilibria.) (b) Cartoon of a continuous stirred-tank reactor. In Section 8.6.2(b) we find an isola-center bifurcation in ODEs modeling the CSTR. (Cartoon drawn by Jeff Poe, 2016. Copyright 2016 by D.G. Schaeffer and J.W. Cain.)*

Suppose $\sigma > 2$, so that the single-cell ODE (8.21) has nonzero equilibria $\mathbf{P}_\pm = (x_\pm, y_\pm)$, obtained by solving

$$x^2 - \sigma x + 1 = 0, \qquad y = x^2. \tag{8.61}$$

Also suppose $\rho > 1$, so that $\mathbf{P}_+$ is a stable equilibrium of the single-cell equation. We consider (8.60) as a bifurcation problem with the diffusion coefficient D as the bifurcation parameter. We recall from Section 6.3.2 that:

- For all D, equation (8.60) has two "trivial" equilibria $(x_\pm, y_\pm, x_\pm, y_\pm)$ in which the concentrations are the same in both cells.
- The stability of the solution (x_+, y_+, x_+, y_+) changes at the threshold

$$D_* = \rho\left(\frac{x_+}{\sigma} - \frac{1}{2}\right); \tag{8.62}$$

  specifically, this solution is stable if $0 \le D < D_*$ and unstable if $D > D_*$. Moreover, at $D = D_*$ the simple-eigenvalue condition (8.51) and the nondegeneracy condition (8.12) are satisfied.
- Equation (8.60) is symmetric with respect to the interchange of concentrations in the two cells,

$$R \cdot (x_1, y_1, x_2, y_2) = (x_2, y_2, x_1, y_1), \tag{8.63}$$

  which is a reflection. The null eigenvector $\mathbf{v} \in \mathbb{R}^4$ of $\mathbf{DF}_*$ has the form $\mathbf{v} = (\mathbf{w}, -\mathbf{w})$ as in (6.31), so that $R\mathbf{v} = -\mathbf{v}$.

How do solutions of (8.60) behave when $D > D_*$? Theorem 8.5.5 strongly suggests that the trivial solution undergoes a pitchfork bifurcation. To investigate this, we turn to simulations. Both panels of Figure 8.11 show three traces of $x_1(t)$ vs. t with $\sigma = \rho = 2.5$ and the indicated value of D. Initial conditions are small perturbations of the trivial equilibrium $(2, 4, 2, 4)$,

$$\begin{bmatrix} x_1(0) \\ y_1(0) \\ x_2(0) \\ y_2(0) \end{bmatrix} = \begin{bmatrix} 2.1 \\ 4 \\ 1.9 \\ 4 \end{bmatrix}, \quad \begin{bmatrix} 1.9 \\ 4 \\ 2.1 \\ 4 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} 1.9 \\ 4 \\ 1.9 \\ 4 \end{bmatrix}.$$

With these values for σ, ρ, the critical value of diffusion is $D_* = 0.75$. In Panel (a), $D = 0.7 < D_*$, and each perturbation decays to zero. In Panel (b), $D = 0.8 > D_*$, and we expect the perturbation to grow until a new equilibrium is reached. For the first and second initial conditions, the solution does indeed tend to a nontrivial equilibrium. For the third initial condition, however, the solution decays back to the trivial solution. To understand the latter behavior, observe that in this case, $x_1(0) = x_2(0)$ and $y_1(0) = y_2(0)$; by symmetry, both cells will have the same concentrations for all time; thus, for these initial conditions, solutions of (8.60) will follow the single-cell dynamics, for which $(2, 4)$ is a stable equilibrium. In other words, the third initial condition lies on the (three-dimensional) stable manifold of the trivial equilibrium.

It is clear that some sort of bifurcation occurs for D between 0.7 and 0.8. Given that for D only slightly greater than D_*, the new equilibria are far from the trivial equilibrium, one would conjecture that the bifurcation is subcritical. In principle, one could verify this conjecture by performing a Lyapunov–Schmidt reduction and showing that $\partial_{xxx} g > 0$. Mercifully, this uninspiring calculation may be avoided. The governing ODEs (8.60) are simple enough that all equilibria may be determined analytically, which we do in the next section, and these calculations show that the bifurcation is indeed a subcritical pitchfork.

(b) Enumeration of equilibria of (8.60).

The ideas in the above bifurcation analysis are widely applicable. By contrast, the following calculations, which we merely outline, apply only to the specific equations (8.60).

Because x_1 is a factor of the equilibrium equation for (8.60a) and x_2 is such a factor for (8.60c), we may distinguish three classes of equilibria of (8.60), as follows:

1. solutions for which both x-concentrations vanish,
2. solutions for which only one of the x-concentrations vanishes,
3. solutions for which neither x-concentration vanishes.

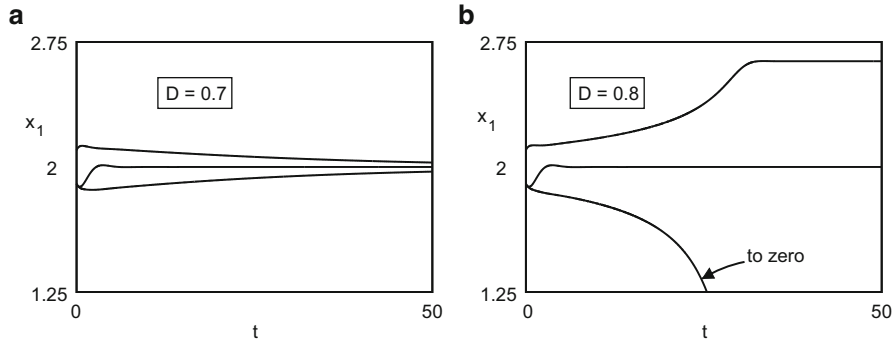

Figure 8.11: *Evidence of a pitchfork bifurcation in* (8.60)*, the Turing instability, as D is increased. Here, $\sigma = \rho = 2.5$. (a) For the diffusion coefficient $D = 0.7 < D_* = 0.75$, the trivial equilibrium (x_+, y_+, x_+, y_+) appears to be stable to perturbations in x_1. (b) For larger $D = 0.8 > D_*$, the trivial equilibrium is unstable, and most perturbations drive the system to a different stable equilibrium. (The two nondecaying trajectories in Panel (b) are images of each other under the reflection* (8.63)*, but this symmetry is not apparent in the figure.)*

In view of the interchange symmetry, the third class may be subdivided:

3a. solutions for which $x_1 = x_2$,

3b. solutions for which $x_1 \neq x_2$.

We study these equilibria in reverse order, focusing on solutions with $D \geq 0$.

Solutions of Class 3a: These are the trivial equilibria $(x_\pm, y_\pm, x_\pm, y_\pm)$ noted above. To see this, we regard the equilibrium equations of (8.60b,d) as a pair of linear equations for y_1, y_2. This 2×2 system is nonsingular, provided $D \neq -\rho/2$, in particular if $D \geq 0$. Hence, given that $x_1 = x_2$, it follows that $y_1 = y_2$. Thus, for solutions of this class, the diffusion term vanishes, so $(x_1, y_1) = (x_2, y_2)$ both satisfy the one-cell equilibrium equations.

These equilibria are plotted in Figure 8.12(a) as black dots labeled $3a_\pm$; note that their location does not vary with D. As described above, (x_+, y_+, x_+, y_+) is stable for (8.60) if $0 \leq D < D_*$ but becomes unstable when $D > D_*$. The other solution is unstable for all $D \geq 0$. (The position of these and other solutions in the bifurcation diagram Figure 8.12b will be discussed below.)

Solutions of Class 3b: These are the equilibria that bifurcate from the trivial solution (x_+, y_+, x_+, y_+) at $D = D_*$. To see this, we deduce from the equilibrium equations of (8.60a,c) that $y_k = \sigma x_k - 1$, $k = 1, 2$. Substituting into the equilibrium

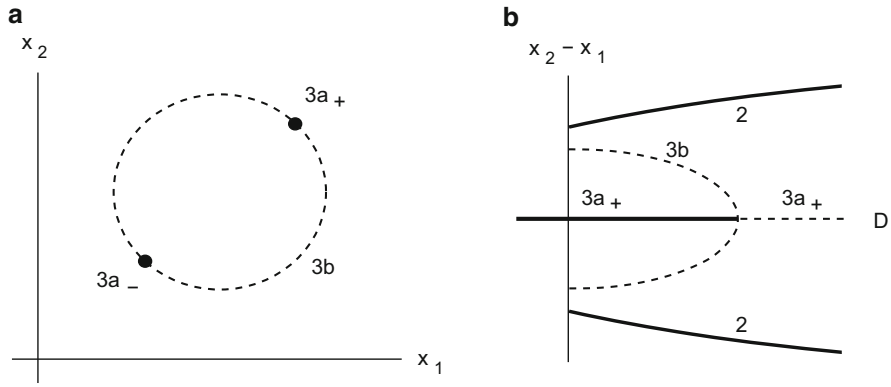

Figure 8.12: *(a) Projection into the x_1, x_2-plane of equilibria of (8.60) of Classes 3a and 3b, with $\sigma = \rho = 2.5$ and D variable. Solutions of Class 3a, which don't depend on D, project onto single dots. (b) A bifurcation diagram for (8.60), restricted to $D \geq 0$. After the bifurcation at $D = D_*$, generically solutions of (8.60) converge to one of the stable solutions of Class 2 as $t \to \infty$.*

equations of (8.60b,d), adding these equations to eliminate D, and rearranging, we find that the solutions lie on a circle

$$(x_1 - \sigma/2)^2 + (x_2 - \sigma/2)^2 = \sigma^2/2 - 2. \tag{8.64}$$

We claim that as sketched in Figure 8.12(a), this circle intersects the diagonal $\{x_1 = x_2\}$ at (the projections of) the equilibria $(x_\pm, y_\pm, x_\pm, y_\pm)$ of Class 3a. Indeed, if $x_1 = x_2$, then (8.64) reduces to $x^2 - \sigma x + 1 = 0$, the equation for the one-cell equilibria $x_\pm$. Moreover, we subtract equations (8.60b,d) and rearrange, using the fact that $x_1 \neq x_2$ for solutions of Class 3b, to obtain a formula for D along the circle:

$$D = \rho \left(\frac{x_1 + x_2}{2\sigma} - \frac{1}{2} \right). \tag{8.65}$$

At the diagonal point $x_1 = x_2 = x_+$, this expression equals $D = D_*$, the bifurcation value. Thus, we have explicitly identified the bifurcating solutions whose existence was deduced abstractly from the fact that the trivial solution (x_+, y_+, x_+, y_+) loses stability at $D = D_*$.

The bifurcation is a pitchfork; is it subcritical or supercritical? From the graph of (8.64) in Figure 8.12(a), you can see that $x_1 + x_2$ *decreases* along the circle as you move away from the bifurcation point (x_+, x_+). By (8.65), D also decreases, so the bifurcation is subcritical.

Solutions of Class 2: These are the ω-limits of the nondecaying trajectories in Figure 8.11. To explore this, suppose that $x_2 = 0$. Then the equilibrium equation of (8.60d) is linear and may be solved for y_2. Also, since $x_1 \neq 0$, the equilibrium

equation of (8.60a) implies that $y_1 = \sigma x_1 - 1$. Substituting these formulas for y_j into the equilibrium equation of (8.60b) gives

$$\frac{\rho + D}{\rho + 2D} x_1^2 - \sigma x_1 + 1 = 0, \tag{8.66}$$

which has two positive roots, provided

$$D > -\rho \frac{\sigma^2 - 4}{2\sigma^2 - 4}. \tag{8.67}$$

The larger root of (8.66) gives a stable equilibrium of (8.60); for the parameter values in Figure 8.11, stability may be inferred from the observation that this equilibrium is the ω-limit of a numerical solution, but it may be verified analytically in general. To conclude, if D satisfies (8.67), then there are four equilibria of (8.60) of Class 2: the two we just found plus two others for which $x_1 = 0$ but $x_2 \neq 0$.

Solutions of Class 1: The zero solution is the obvious equilibrium of this class. As it happens, there are some additional solutions that exist for $D < 0$. Although these have no physical significance, they may have some interest as a mathematical phenomenon, and we invite you to explore them in Exercise 24.

Figure 8.12(b) shows a partial bifurcation diagram. Equilibria of Classes $3a_+$, 3b, and 2 are plotted as functions of $D \geq 0$. To make the symmetry of the bifurcation more evident, we plot the difference $x_2 - x_1$ rather than just one variable by itself. In this projection, the trivial solutions (x_+, y_+, x_+, y_+) of Class $3a_+$ lie along the D-axis. Solutions of Class 3b bifurcate subcritically from them at $D = D_*$. To reduce clutter, only the stable solutions of Class 2 are included in the figure.

8.6.2 The CSTR

The continuous stirred-tank (chemical) reactor is a rich source of bifurcation phenomena. We study it here to illustrate hysteresis-point and isola-center bifurcations, but it also exhibits interesting time-dependent behavior; see, for example, [89].

(a) Hysteresis-point bifurcation: Recall from Exercise 5.9 that we simplified the equations of the CSTR to

$$\begin{aligned} dx/dt &= \mu(1 - x) - xe^y, \\ dy/dt &= -\mu y + \sigma x e^y. \end{aligned} \tag{8.68}$$

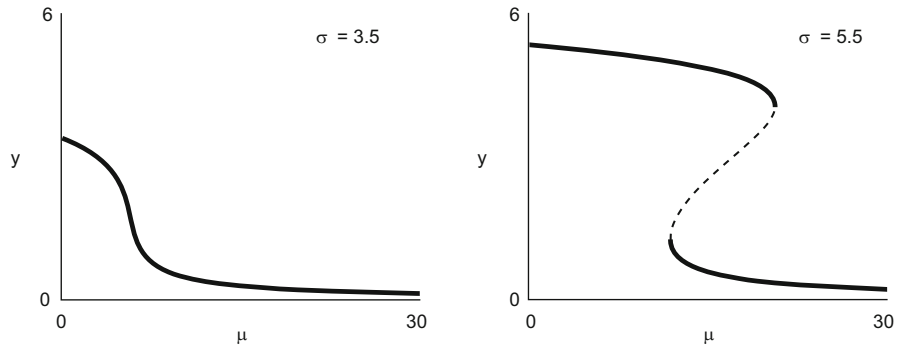

Figure 8.13: *Two bifurcation diagrams for the CSTR equations (8.68). As illustrated in Panel (b), the bifurcation diagram exhibits hysteresis if $\sigma > 4$.*

We solve the first equilibrium equation for x as a function of y and μ, substitute this into the second equation, and manipulate the result, canceling the uninteresting solution $\mu = 0$, to obtain the reduced equation

$$g(y, \mu) = (\sigma - y)e^y - \mu y = 0. \tag{8.69}$$

Note that since $\mu > 0$, this equation has solutions only if y and $\sigma - y$ have the same sign, i.e., only if $0 < y < \sigma$. Bifurcation diagrams described by (8.69) are plotted in Figure 8.13 for two different values of σ. The key point is that if $\sigma > 4$, then over a range of μ the bifurcation diagram exhibits bistability and has two saddle-node bifurcation points, while if $\sigma < 4$, there is a unique equilibrium for all μ, which varies smoothly with μ.

To derive this behavior analytically, it is convenient to study μ as a function of y,

$$\mu(y) = \frac{\sigma - y}{y} e^y,$$

where $0 < y < \sigma$. At the lower limit, $\mu(y) \to \infty$ as y tends to zero, and at the upper limit, $\mu(y) \to 0$ as y tends to σ. We ask you to apply calculus to show that if $\sigma < 4$, then $\mu(y)$ is monotonically decreasing on $0 < y < \sigma$, and if $\sigma > 4$, there is an interval of y where its derivative is positive. This information confirms the behavior shown in Figure 8.13.

At the transition point $\sigma = 4$, equation (8.69) has a hysteresis-point bifurcation, as defined in Table 8.1. We ask you to show that if $\sigma = 4$, then at $(y_*, \mu_*) = (2, e^2)$, the reduced function satisfies

$$g(y_*, \mu_*) = \partial_y g(y_*, \mu_*) = \partial_{yy} g(y_*, \mu_*) = 0.$$

If we indicate explicitly the dependence on σ in (8.69), say

$$G(y, \mu, \sigma) = (\sigma - y)e^y - \mu y = 0,$$

then $G(y, \mu, \sigma)$ is an *unfolding* of $G(y, \mu, 4)$, as defined in Section 8.5.6. Comparing Figures 8.9 and 8.13, we see that both $G(y, \mu, \sigma)$ and the unfolding of the normal form (8.58) exhibit the same qualitative behavior, i.e., in words, a transition from unique equilibria to bistability.

These phenomena have significance for industrial applications. The CSTR is typically studied as a model for some chemical process in a plant. If the model admits multiple steady states, great care is required during startup of the plant to guarantee that after transients, it winds up operating on the *right solution branch*. Mistakes are very costly. If the transients put the plant on the lower solution branch when the upper branch was anticipated, the process may fizzle and need to be restarted from scratch. Or worse, if the transients put the plant on the upper branch when the lower branch was anticipated, the whole plant could explode!

(b) Isola-center bifurcation: For this example, we consider an extension of the CSTR model from Chapter 5 that includes an additional effect: cooling of the tank through contact with a heat bath. This extension adds a term to the temperature equation; after scaling, we have

$$\begin{aligned} dx/dt &= \mu(1 - x) - xe^y, \\ dy/dt &= -\mu y + \sigma xe^y - \beta(y - \xi), \end{aligned} \tag{8.70}$$

where β and ξ are the (nondimensionalized) heat-conduction coefficient and temperature of the bath, respectively. Our interest in this model is to illustrate a bifurcation phenomenon, not to catalogue all possible behavior of a reactor, so we shall restrict our attention to a special case of (8.70). Specifically, we assume that $\xi = 0$, or in dimensional terms, that the temperature of the bath is the same as the temperature of the feed.

In Figure 8.14 we show bifurcation diagrams for the equilibria of (8.70), assuming $\sigma = 8$, for various values of β. To understand these diagrams, we start by comparing Panel (a), where the coefficient β of the new effect is small, with Figure 8.13(b), where $\beta = 0$ (and σ is large). These diagrams are similar in that both have a window in μ of bistability. They differ as the flow rate μ approaches 0, a limit that is described more suggestively by saying the "residence time" $1/\mu$ tends to infinity. This difference is not hard to understand. Over a long residence time essentially all of the reactant is consumed, and its latent heat is released. When $\beta = 0$, no heat is lost to the environment, which means that the nondimensional temperature y rises to σ, its maximum possible value. By contrast, when $\beta > 0$, over the long residence

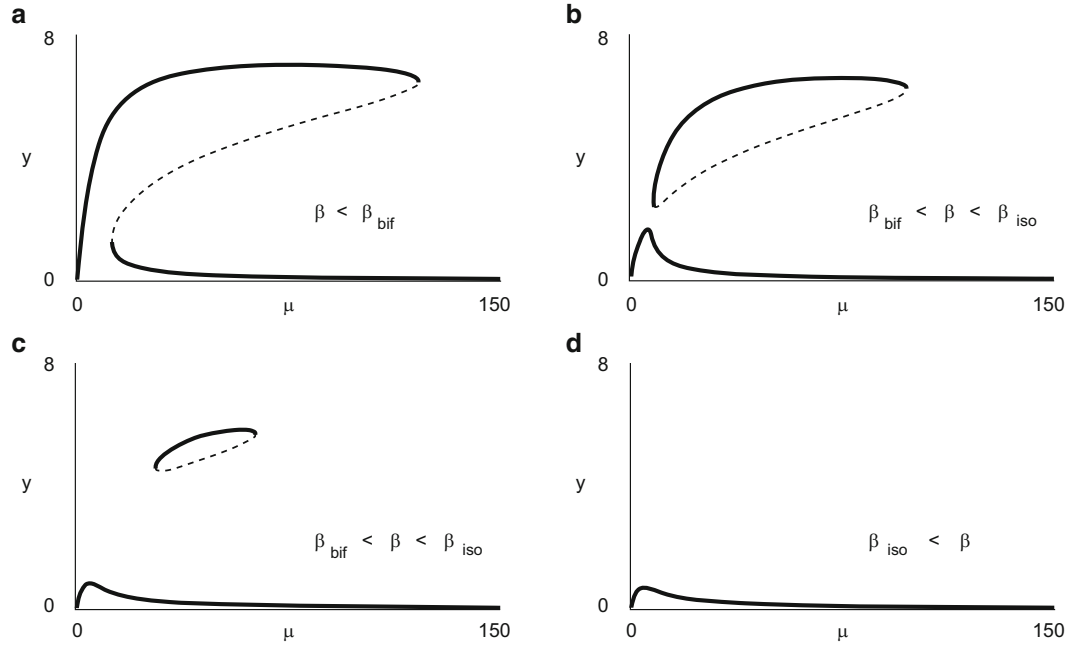

Figure 8.14: *Bifurcation diagrams for* (8.70) *assuming* $\sigma = 8$. *(a) For* $\beta = 5$, *there is hysteresis but no isola. (b) Increasing* β *to* 7.5 *causes the "neck" in Panel (a) to pinch off and form an isola. (c) For* $\beta = 10$, *the isola still exists but has shrunk. (d) At* $\beta = 11$, *the isola has disappeared.*

time, this heat is absorbed by the bath, so y equilibrates to the temperature of the bath, which we assumed has the (nondimensional) value zero.

The other panels in Figure 8.14 can be understood through bifurcations from this first case. As β is increased, the "neck" in Panel (a) pinches off at a transcritical[25] bifurcation when $\beta = \beta_{\text{bif}} \approx 7.39$ and then forms an isola, as shown in Panel (b). As β is further increased, the isola shrinks, as illustrated in Panel (c), all the way down to an isola-center bifurcation when $\beta = \beta_{\text{iso}} \approx 10.47$. Beyond β_{iso}, the system does not exhibit bistability for any μ, as in Panel (d).

Much of this behavior can be derived analytically. In Exercise 23 we offer hints for doing this.

As discussed above, when starting operations at an industrial plant, it is important to know whether the system admits multiple steady states. The designer of the plant may look for multiple steady states by continuously varying all possible parameters, either in a laboratory-scale mock-up of the plant or in a simulation. However, an isola will never be seen under quasistatic variation of parameters; a finite, possibly large, perturbation is required to "kick" the system onto a disconnected branch of equilibria. It warms our mathematicians' hearts to say so, but there is no substitute for adequate theoretical analysis.

8.7 Examples of Hopf Bifurcation[26]

8.7.1 An Academic Example

To establish the ideas regarding Hopf bifurcation, let us begin with a simple, academic, example of it in a by-now-familiar equation, written in matrix notation

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \mu & -1 \\ 1 & \mu \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} - (x^2 + y^2) \begin{bmatrix} x \\ y \end{bmatrix}. \tag{8.71}$$

For all μ, this equation has an equilibrium at $x = y = 0$, and this is the only equilibrium. *(Check this!)* At the equilibrium, the Jacobian is

$$\mathbf{DF}(\mathbf{0}, \mu) = \begin{bmatrix} \mu & -1 \\ 1 & \mu \end{bmatrix},$$

[25] Since no symmetry is present in this problem, on grounds of genericity we expect the bifurcation to be transcritical. High-resolution computations confirm this expectation, but the actual bifurcation diagram is so nearly symmetric that even magnified 50 times, it is visually indistinguishable from a pitchfork. These circumstances give a warning about the limitations of arguments based on genericity.

[26] It would be historically more accurate to refer to Andronov–Hopf bifurcation, but the shorter name has come into widespread use.

which has eigenvalues $\mu \pm i$. In particular, it is stable for $\mu < 0$ and unstable for $\mu > 0$. However, the mechanics for the loss of stability are quite different from the previous examples of bifurcation. Rather than a single real eigenvalue passing through zero, here stability is lost as a pair of complex-conjugate eigenvalues of the Jacobian cross the imaginary axis.[27]

Taking the central message of bifurcation theory to heart, we expect some change in the set of solutions of (8.71) as μ crosses zero. However, no new equilibria appear. The only equilibrium of (8.71) is $x = y = 0$, no matter what the value of μ. To see what change does occur near $\mu = 0$, let us invoke polar coordinates one more time to rewrite the equations as

$$\begin{aligned} r' &= \mu r - r^3, \\ \theta' &= 1. \end{aligned} \tag{8.72}$$

Note that r' vanishes if $r = 0$ or if $r^2 = \mu$. The first solution just represents the equilibrium $x = y = 0$. By contrast, $r^2 = \mu$ represents a new type of orbit that appears as μ crosses zero, i.e., the periodic solution $r = \sqrt{\mu}$, $\theta = t$, where we have chosen the phase arbitrarily. In Cartesian coordinates, the solution is

$$x = \sqrt{\mu}\cos t, \quad y = \sqrt{\mu}\sin t.$$

This behavior is typical: in Hopf bifurcation, periodic solutions appear, as a parameter varies, when two eigenvalues of the Jacobian at an equilibrium cross the imaginary axis.

It is instructive to consider a more general, but still academic,[28] equation that exhibits Hopf bifurcation: in polar coordinates on $\mathbb{R}^2$,

$$\begin{aligned} r' &= \mu r - \alpha r^3, \\ \theta' &= 1 + \beta r^2 + \gamma \mu, \end{aligned} \tag{8.73}$$

with parameters α, β, γ. The equilibrium $r = 0$ is stable for $\mu < 0$ and unstable for $\mu > 0$. If $\alpha > 0$, then the bifurcating periodic solutions exist for $\mu > 0$ and are stable, as indicated schematically in Figure 8.15(a); this case is called *supercritical*. In this case, if μ is increased (quasistatically) beyond zero, solutions of (8.73) follow the bifurcating periodic orbits. On the other hand, if $\alpha < 0$, then the periodic solutions exist for $\mu < 0$ and are unstable, which is called *subcritical*. For $\mu < 0$, the basin of attraction of the equilibrium is bounded by the periodic orbit, and the basin shrinks to a point as μ tends to zero. If μ increases beyond zero, the solution of (8.73) blows up.

The parameters β and γ in (8.73) have only a limited effect on the behavior of solutions: they cause the period of the oscillatory solutions to vary with their amplitude.

[27] In connection with this behavior, we recommend that you revisit Exercise 2.15.

[28] In fact, (8.73) has greater generality than may be apparent; cf. Section 8.10.1.

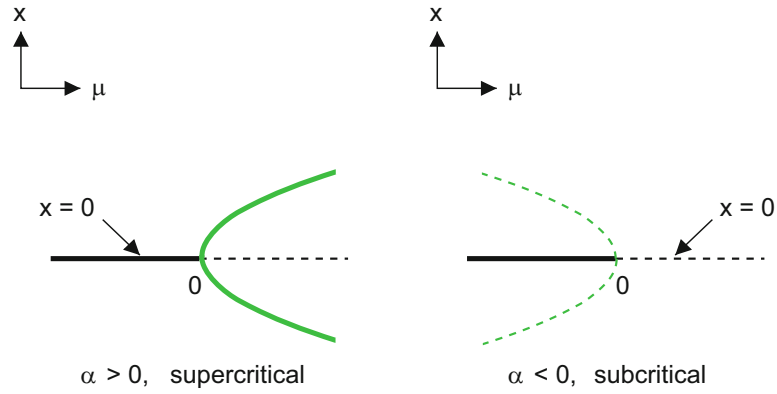

Figure 8.15: *Bifurcation diagrams with Hopf bifurcation in the system* (8.73) *for the cases* $\alpha > 0$ *and* $\alpha < 0$. *For a fixed value of* μ, *a periodic orbit is represented by the maximum and minimum of the Cartesian coordinate* x *along the orbit.*

8.7.2 The "Repressilator"

Hopf bifurcations may occur in any dimension. To illustrate this point, let us consider a gene network called the *repressilator*[21]. The name tries to capture the idea that oscillations occur through the mutual repression, or inhibition, of three genes, as indicated schematically in Figure 8.16(a). It is believed that some oscillations in biological systems—biological clocks in informal parlance—are based on such networks.[29]

Mathematically, after scaling, the repressilator is described by the system

$$\begin{aligned} x' &= \frac{\mu}{1+y^4} - x, \\ y' &= \frac{\mu}{1+z^4} - y, \\ z' &= \frac{\mu}{1+x^4} - z, \end{aligned} \tag{8.74}$$

where to simplify the analysis, we assume that the interactions are completely symmetric; this is not necessary for bifurcation. These equations have a steady-state solution with $x = y = z = x_{\text{eq}}(\mu)$, where x_{eq} satisfies the equation

$$x_{\text{eq}}(1 + x_{\text{eq}}^4) = \mu. \tag{8.75}$$

[29]The cell cycle is a prime example of such a biological clock. Although the cell cycle is vastly more complicated than the simple system (8.74), nevertheless, this model is considered to be a useful point of departure for studying the cell cycle (S. Haase, private communication). The model is mentioned in passing on pp. 564–565 of Winfree's encyclopedic work [97].

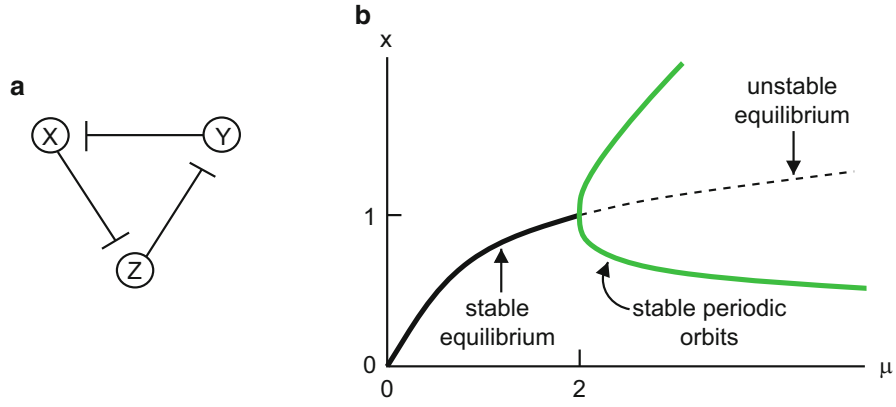

Figure 8.16: *(a) Schematic diagram of the repressilator system, (8.74); y inhibits the production of x, etc. (b) Bifurcation diagram for the repressilator (8.74). For each $\mu > 2$, there is a stable periodic orbit.*

It is easily seen that for all $\mu > 0$, (8.75) has a unique solution with $x_{\text{eq}} > 0$. Moreover, this solution depends smoothly and monotonically on μ, and it tends to infinity as $\mu \to \infty$. The graph of this solution is the "backbone" of the bifurcation diagram of Figure 8.16(b).

To determine how the stability of $\mathbf{x}_{\text{eq}}(\mu)$ depends on μ, we compute that the Jacobian of the system at the equilibrium equals

$$\mathbf{DF}(\mathbf{x}_{\text{eq}}(\mu), \mu) = \begin{bmatrix} -1 & -\rho(\mu) & 0 \\ 0 & -1 & -\rho(\mu) \\ -\rho(\mu) & 0 & -1 \end{bmatrix}, \tag{8.76}$$

where

$$\rho(\mu) = \frac{4\mu x_{\text{eq}}^3(\mu)}{(1 + x_{\text{eq}}^4(\mu))^2} = \frac{4x_{\text{eq}}^4(\mu)}{1 + x_{\text{eq}}^4(\mu)} \,. \tag{8.77}$$

This matrix has the form $\mathbf{DF}(\mathbf{x}_{\text{eq}}(\mu), \mu) = -\rho(\mu)B - I$, where

$$B = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

The eigenvalues of B are cube roots of unity, so the eigenvalues of $\mathbf{DF}(\mathbf{x}_{\text{eq}}(\mu), \mu)$ are

$$-\rho(\mu) - 1 \quad \text{and} \quad -\rho(\mu)\left(-\frac{1}{2} \pm i\frac{\sqrt{3}}{2}\right) - 1.$$

As μ increases, so does $\rho(\mu)$. The real eigenvalue of $\mathbf{DF}(\mathbf{x}_{\text{eq}}(\mu), \mu)$ just becomes more negative. By contrast, the complex-conjugate eigenvalues, which lie in the

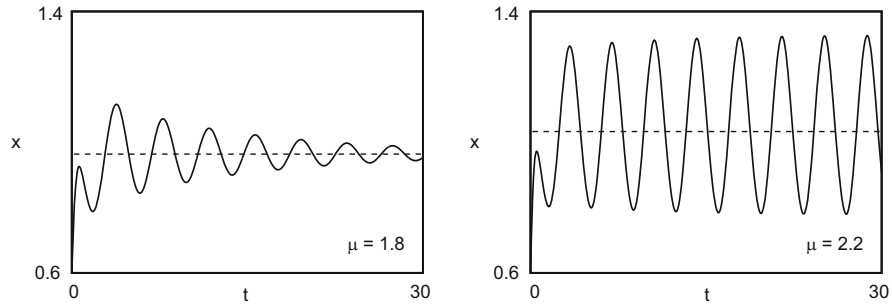

Figure 8.17: *Sample traces of x versus t obtained by numerical solution of the repressilator (8.74) with initial conditions $x(0) = 0.6$, $y(0) = 0.4$, $z(0) = 0.2$. Horizontal lines locate the equilibrium solutions. Left panel: With $\mu = 1.8$, transient oscillations occur before the system settles to equilibrium. Right panel: With $\mu = 2.2$, the solution trajectory approaches a limit cycle. (If you solve the equations yourself, you may view the orbit of this periodic solution in, say, the x, y-plane.)*

left half-plane for small μ, cross the imaginary axis when $\rho(\mu) = 2$. By (8.77), this happens when $x_{\rm eq}^4(\mu) = 1$, which according to (8.75) implies that $\mu = 2$. In summary, the equilibrium (8.75) is stable if $\mu < 2$ and unstable if $\mu > 2$.

Calling on simulations, we see in Figure 8.17 that stable small-amplitude periodic solutions of (8.74) appear after μ crosses 2. As shown in the bifurcation diagram of Figure 8.16(b), they grow continuously as μ increases. Thus, (8.74) exhibits a *supercritical* Hopf bifurcation. This behavior could be determined analytically, but it would take more patience than your authors have to carry this through.

8.7.3 Section 1.6 Revisited: Part V

Another example of a Hopf bifurcation occurs in the Lotka–Volterra equations augmented to include logistic growth and the Allee effect for the prey,

$$\begin{aligned} (a) \quad x' &= x\left(\frac{x-\varepsilon}{x+\varepsilon}\right)(1-\frac{x}{K}) - xy, \\ (b) \quad y' &= \rho(xy - y). \end{aligned} \tag{8.78}$$

In Section 6.2.3, we saw that if $K > 1$, then (i) the coexistence equilibrium $(x_{\rm eq}, y_{\rm eq})$, where $x_{\rm eq} = 1$ and $y_{\rm eq} = (1 - 1/K)(1-\varepsilon)/(1+\varepsilon)$, is in the physical domain $\{x \geq 0, y \geq 0\}$ and (ii) at this equilibrium, $\det \mathbf{DF}(\mathbf{x}_{\rm eq}(K), K) > 0$ and

$$\operatorname{tr} \mathbf{DF}(\mathbf{x}_{\rm eq}(K), K) = \frac{2\varepsilon - (1 + 2\varepsilon - \varepsilon^2)/K}{(1+\varepsilon)^2}.$$

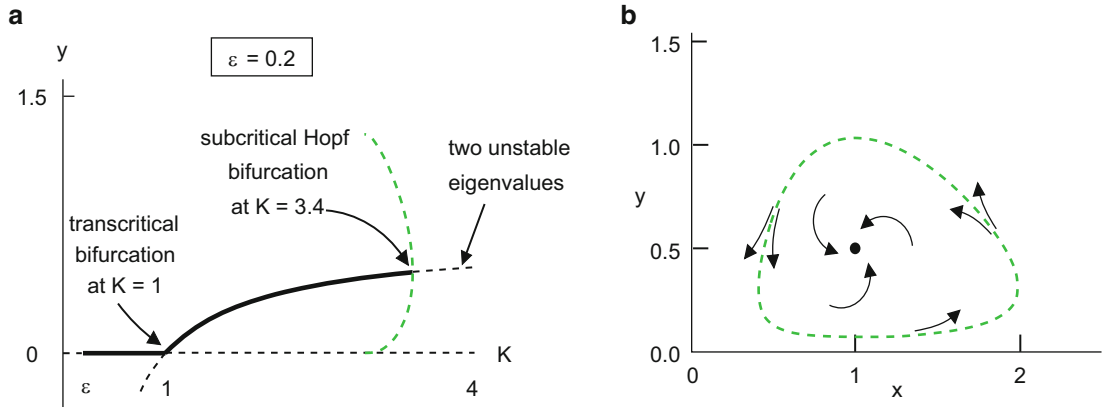

Figure 8.18: *(a) Bifurcation diagram for the augmented Lotka–Volterra equations* (8.78). *(b) A phase-plane plot for* $\varepsilon = 0.2$, $K = 3.2$, *and* $\rho = 1$. *The green dashed curve is an unstable periodic orbit. For initial conditions inside this orbit, trajectories spiral in to the coexistence equilibrium; for initial conditions outside, they spiral out and (generically) eventually approach the extinction equilibrium* $(0, 0)$.

Thus, the equilibrium is asymptotically stable if $1 < K < (1 + 2\varepsilon - \varepsilon^2)/2\varepsilon$, but it is unstable if $K > (1 + 2\varepsilon - \varepsilon^2)/2\varepsilon$. If $K = (1 + 2\varepsilon - \varepsilon^2)/2\varepsilon$, the eigenvalues of $\mathbf{DF}(\mathbf{x}_{\text{eq}}(K), K)$ lie on the imaginary axis; in other words, the coexistence equilibrium undergoes a Hopf bifurcation as K crosses this value.

With numerics we find that equation (8.78) has periodic solutions for values of K near the bifurcation point, but as indicated in Figure 8.18(a), they exist when $K < (1+2\varepsilon-\varepsilon^2)/2\varepsilon$, where the coexistence equilibrium is asymptotically stable; thus, the bifurcation is *subcritical*. These periodic solutions are unstable, so they are not readily observable in a straightforward simulation. As illustrated in Figure 8.18(b), they may be observed indirectly,[30] since the basin of attraction of the coexistence equilibrium is bounded by these periodic solutions.

Preview of coming attractions: As K decreases, the amplitude of these periodic solutions grows, but they cease to exist abruptly at $K \approx 2.952$, when the minimum of y along the orbit vanishes. What happens to them? Stay tuned. We shall take up this issue in Section 9.1.3.

8.7.4 The "Denatured" Morris–Lecar System

The Morris–Lecar system, an experimentally based model for the electrical behavior of neurons, has a rich bifurcation structure. The system is introduced and analyzed

[30] Alternatively, they may be observed directly by solving the equations with "time running backward." We invite you to do this computation.

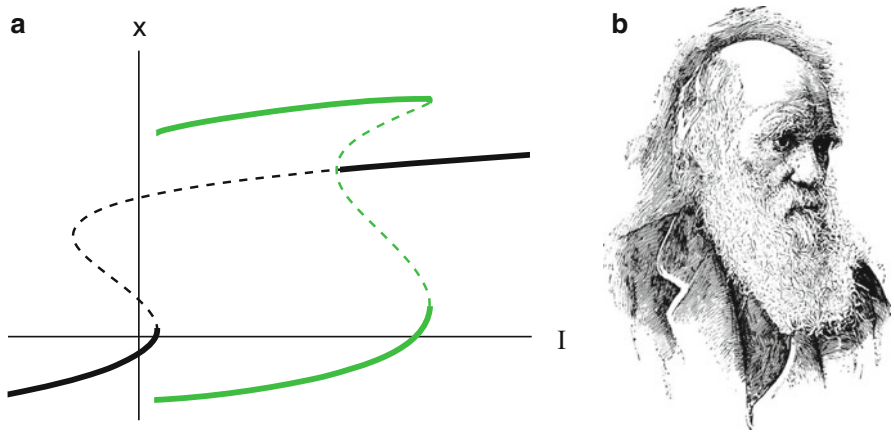

Figure 8.19: *(a) A bifurcation diagram for the denatured Morris–Lecar equation* (8.79) *with parameters* (8.82)*. A subcritical Hopf bifurcation occurs when* $I \approx 0.07573$*. (b) The Morris–Lecar model was derived from experiments on muscle fibers of the giant acorn barnacle. Recognizing this origin, we include a [public domain] reprinted sketch of Charles Darwin, who was passionate about barnacles [78]. He kept a first draft of* On the Origin of Species *in a desk drawer for eight years while he worked out a detailed taxonomy of barnacles! Source: Charles Darwin [Public domain], via Wikimedia Commons.*

in Section 3.2 of [23]. (Cf. also Section 9.7 of [96].) Here we study a simplification of it, which we call the *denatured Morris–Lecar* equation, the system[31]

$$\begin{aligned} x' &= x^2(1-x) - y + I, \\ y' &= Ae^{\alpha x} - \gamma y, \end{aligned} \tag{8.79}$$

where I, A, α, γ are constants; the bifurcation parameter I may have either sign, while the other constants are all positive. Figure 8.19 shows a bifurcation diagram, x vs. I, for (8.79) for one choice of the parameters.[32] An S-shaped curve of equilibria is plotted in black. Both stable and unstable periodic solutions are plotted in green, displaying the maximum and minimum values of x along the orbit. Our goals in this section are (i) to understand why the graph of equilibria is S-shaped and (ii) to locate the Hopf bifurcation on the upper branch of the figure.

The S-shaped steady-state bifurcation diagram can be understood from the nullclines for (8.79), such as drawn in Figure 8.20. As I varies, the x-nullcline shifts up

[31]It may be informative to compare (8.79) with the FitzHugh–Nagumo equation (Exercise 5.11), which is a simplification of the Hodgkin–Huxley model. Both systems have the same x-nullclines, apart from scaling. While the y-nullclines of the FitzHugh–Nagumo equations are straight lines, the y-nullclines of (8.79) curve upward (see Figure 8.20), because a linear term is replaced by the exponential.

[32]There's a long story that we're not telling you about how we chose these parameters. Let us remind you of the old adage, "A good teacher doesn't tell you everything he/she knows."

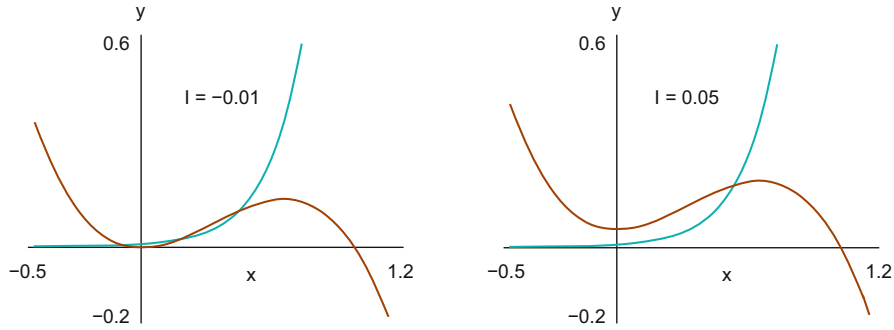

Figure 8.20: *Nullclines for* (8.79) *with parameters* (8.82). *Left panel: For* $I = -0.01$, *there are three equilibria. Right panel: For* $I = 0.05$, *there is just one equilibrium.*

or down. When I is sufficiently large, as in the right panel of the figure, the cubic x-nullcline is so high that the nullclines intersect only once, at a large value of x. When I is somewhat smaller, as in the left panel, the cubic may cross the exponential three times. When I is sufficiently negative (not shown in figure), the x-nullcline is so low that again there is only one intersection, at a negative value of x. Hence, the S-shape.

Examining the slopes of the nullclines and applying the information in Table 6.2, we see that $\det \mathbf{DF} < 0$ along the middle branch—these are saddle points—and $\det \mathbf{DF} > 0$ along the upper and lower branches. Equilibria on the upper and lower branches are mostly stable, but some equilibria on the upper branch may be destabilized by a Hopf bifurcation.

Let's look for a Hopf bifurcation. For (x_*, y_*) to be a Hopf-bifurcation point of (8.79) when $I = I_*$, this data must satisfy

$$\begin{array}{ll} \text{(a)} & x_*^2(1 - x_*) - y_* + I_* = 0, \\ \text{(b)} & Ae^{\alpha x_*} - \gamma y_* = 0, \\ \text{(c)} & 2x_* - 3x_*^2 - \gamma = 0. \end{array} \tag{8.80}$$

The first two of these equations are the equilibrium equations of (8.79); the third requires that the Jacobian

$$\mathbf{DF}_* = \begin{bmatrix} 2x_* - 3x_*^2 & -1 \\ \alpha A e^{\alpha x_*} & -\gamma \end{bmatrix} \tag{8.81}$$

have trace zero. Equations (8.80) are necessary for Hopf bifurcation but not sufficient; we need also to show that $\det \mathbf{DF}_* > 0$ to rule out the possibility that $\mathbf{DF}_*$ has real eigenvalues that happen to add up to zero.

Provided $\gamma < 1/3$, equations (8.80) may solved sequentially in reverse order. For the parameters

$$A = 0.001, \quad \alpha = 6, \quad \text{and} \quad \gamma = 0.15, \tag{8.82}$$

we obtain[33]

$$\begin{array}{llcll} (a) & x_* & = & (1+\sqrt{1-3\gamma})/3 & \approx & 0.58054, \\ (b) & y_* & = & Ae^{\alpha x_*}/\gamma & \approx & 0.21710, \\ (c) & I_* & = & y_* - x_*^2(1-x_*) & \approx & 0.075731. \end{array} \tag{8.83}$$

For these parameters[34]

$$\det \mathbf{DF}_* = \alpha A e^{\alpha x_*} - \gamma^2 \approx 0.17289 > 0.$$

Thus, the eigenvalues of $\mathbf{DF}_*$ are imaginary, specifically

$$\lambda(\mathbf{DF}_*) = \sqrt{\det \mathbf{DF}_*}\, i \approx \pm 0.41580\, i;$$

i.e., we've located the desired Hopf bifurcation.

Incidentally, the calculations outlined in Exercise 26 show that the bifurcation is subcritical.

Another preview of coming attractions: The periodic solutions in Figure 8.19 cease to exist abruptly at $I \approx 0.007137$, where the curve in the figure ends. What happens to them? Is it an accident that the end of the periodic orbits appears to line up with one of the saddle-node bifurcations? Answers in Section 9.2.2.

8.8 Theoretical Description of Hopf Bifurcation

8.8.1 A Bifurcation Theorem

The Lyapunov–Schmidt reduction provides a simple, effective tool for understanding steady-state bifurcation. By contrast, the tools for studying Hopf bifurcation are more complicated. Even the statement of the bifurcation theorem below is irritatingly long, and we do not prove it. (But see the Pearls for discussion and references.)

Consider a one-parameter family of ODEs

$$\mathbf{x}' = \mathbf{F}(\mathbf{x}, \mu), \tag{8.84}$$

[33]Equation (8.80c) has two roots $x_* = (1 \pm \sqrt{1-3\gamma})/3$. In (8.83a) we consider only the root with the plus sign, because the other case gives real eigenvalues for the Jacobian.

[34]For other parameters, e.g., keeping $\alpha = 6$ and $\gamma = 0.15$ but reducing A to 0.0001152, this determinant may vanish. In this case, we have what's called a Takens–Bogdanov bifurcation (see Exercise 9).

where we assume that $\mathbf{F} \in \mathcal{C}^3$. As with steady-state bifurcation, we suppose that the equation has a nonhyperbolic equilibrium $(\mathbf{x}_*, \mu_*)$, but the hypothesis (8.10) on the Jacobian $\mathbf{DF}_*$ is altered as follows: we assume that

$$\lambda_{1,2}(\mathbf{DF}_*) = \pm i\omega_* \neq 0, \qquad \Re\,\lambda_j(\mathbf{DF}_*) \neq 0,\ j = 3, \dots, d. \tag{8.85}$$

In words, $\mathbf{DF}_*$ has nonzero simple eigenvalues $\pm i\omega_*$ and no other eigenvalues on the imaginary axis. In particular, zero is not an eigenvalue of $\mathbf{DF}_*$, so it follows from the implicit function theorem that for μ near μ_*, there is a smooth branch of equilibria $\mathbf{x}_{\text{eq}}(\mu)$ passing through $(\mathbf{x}_*, \mu_*)$. (In Theorem 8.2.2, the steady-state counterpart of Theorem 8.8.1, we needed a separate hypothesis to this effect.)

By (8.85), the linearized ODE

$$\mathbf{w}' = \mathbf{DF}_* \cdot \mathbf{w} \tag{8.86}$$

has (real) nonzero $2\pi/\omega_*$-periodic solutions, the components of $\mathbf{w}(t)$ being trig functions. Such solutions are not unique: if $\mathbf{w}(t)$ is one such solution, so is the scaled phase-shifted function $c\mathbf{w}(t - t_0)$. Nevertheless, in Theorem 8.8.1 we choose one specific solution $\mathbf{w}(t)$ of (8.86) for parametrizing the bifurcating solutions. Let us define vectors in $\mathbb{R}^d$

$$\mathbf{v}_1 = \mathbf{w}(0), \qquad \mathbf{v}_2 = \mathbf{w}'(0). \tag{8.87}$$

As in Lemma 8.2.1, we may deduce that the Jacobian $\mathbf{DF}(\mathbf{x}_{\text{eq}}(\mu), \mu)$ has a smoothly varying (complex) eigenvalue $\lambda(\mu)$ such that $\lambda(\mu_*) = i\omega_*$. In analogy with (8.11), we require that

$$\frac{d}{d\mu}\,\Re\,\lambda(\mu_*) \neq 0. \tag{8.88}$$

Despite potential difficulties in verifying (8.88) in higher dimensions, in two dimensions the formula

$$\Re\,\lambda(\mu) = \frac{1}{2}\,\operatorname{tr}\mathbf{DF}(\mathbf{x}_{\text{eq}}(\mu), \mu),$$

which holds for μ near μ_*, makes this task trivial.

Theorem 8.8.1. *Under the above hypotheses, a one-parameter family of periodic solutions of (8.84), say $\boldsymbol{\gamma}(t, a)$, where $0 \leq a < \varepsilon$, bifurcates from $(\mathbf{x}_*, \mu_*)$; the function $\boldsymbol{\gamma}(t, a)$ satisfies (8.84) for a specific parameter value $\mu = \mu(a)$. If $\mathbf{w}(t)$ is a nonzero periodic solution of the linearized equation (8.86), this family admits a $\mathcal{C}^2$-parametrization*

$$(a)\ \boldsymbol{\gamma}(t, a) = \mathbf{x}_{\text{eq}}(\mu(a)) + a\mathbf{w}(\omega(a)t) + \mathcal{O}(a^2), \qquad (b)\ \mu(a) = \mu_* + \mu_2 a^2 + o(a^2), \tag{8.89}$$

where the frequency factor satisfies $\omega(a) = 1 + \mathcal{O}(a^2)$ and μ_2 is a constant; in particular, the period of $\boldsymbol{\gamma}(t, a)$ equals $2\pi/\omega_*$ plus an $\mathcal{O}(a^2)$-correction. The union of the orbits

$$\bigcup_{t\in\mathbb{R}} \bigcup_{0\le a<\varepsilon} \{(\boldsymbol{\gamma}(t,a),\mu(a))\} \tag{8.90}$$

is a two-dimensional surface in $\mathbb{R}^d\times\mathbb{R}$ *of class* $\mathcal{C}^2$ *through* $(\mathbf{x}_*,\mu_*)$*, which is tangent to the subspace spanned by* $(\mathbf{v}_1,0)$ *and* $(\mathbf{v}_2,0)$*, where* $\mathbf{v}_1,\mathbf{v}_2$ *are defined by* (8.87)*. Every periodic solution of* (8.9) *that is contained in an appropriate neighborhood* $\mathcal{N}\subset\mathbb{R}^d\times\mathbb{R}$ *of* $(\mathbf{x}_*,\mu_*)$ *belongs to the family* $\boldsymbol{\gamma}(t,a)$*.*

Remarks: (i) We suggest that you interpret the theorem for the normal form (8.73). In particular, doing so shows the need for the frequency factor $\omega(a)$. *(ii)* Nothing would be gained by allowing the amplitude parameter a to take negative values. Since $\mathbf{w}(t)$ has trigonometric dependence, $\mathbf{w}(t+\pi/\omega_*)=-\mathbf{w}(t)$. Thus, including $a<0$ in the union (8.90) would simply give a double enumeration of points in this surface. *(iii)* If $\mu_2\neq 0$ in (8.89b), the amplitude of the periodic solution $\boldsymbol{\gamma}(t,a)$ is approximately proportional to $\sqrt{|\mu-\mu_*|}$. In such a case, the periodic orbits could be parametrized by μ, and the amplitude parameter a could be eliminated from the formulation of the theorem. However, this is not possible in general. The trivial linear equation

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \mu & -1 \\ 1 & \mu \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix},$$

whose periodic solutions all lie in the plane $\{\mu=0\}$, is the simplest counterexample. In Section 8.8.2 we will encounter a more serious example with the same behavior.

Stability information is available if $\mu_2\neq 0$ in (8.89b) and if the nonimaginary eigenvalues in (8.85) all have negative real parts: in symbols,

$$\lambda_{1,2}(\mathbf{DF}_*)=\pm i\omega_*\neq 0, \qquad \Re\,\lambda_j(\mathbf{DF}_*)<0,\ j=3,\dots,d. \tag{8.91}$$

Theorem 8.8.2. *If* (8.91) *holds, then near a supercritical Hopf bifurcation the bifurcating periodic solutions are stable, while they are unstable near a subcritical one.*

It follows from Theorem 8.8.1 that a Hopf bifurcation is supercritical if the two quantities μ_2 and $\Re\, d\lambda/d\mu(\mu_*)$ have the same sign. (*Check this!*) Conversely, the bifurcation is subcritical if μ_2 and $\Re\, d\lambda/d\mu(\mu_*)$ have opposite signs.

8.8.2 The Activator–Inhibitor: Extreme Nongeneric Behavior

Recall once again from Section 6.3.1 the following two facts about equilibria of the activator–inhibitor equations (8.21): (i) If $\sigma>2$, then (8.21) has two nonzero equilibria $\mathbf{P}_\pm=(x_\pm,y_\pm)$, where $x_\pm$ satisfies $x^2-\sigma x+1=0$ and $y_\pm=x_\pm^2$. (ii) If $\rho>1$, then the top equilibrium $\mathbf{P}_+$ is a sink, and if $\rho<1$, it is a source.

Fix $\sigma>2$ and consider these equations as a bifurcation problem with ρ as the bifurcation parameter. We ask you to show that the equilibrium $\mathbf{P}_+$ undergoes a Hopf

bifurcation at $\rho_* = 1$, in which the hypotheses (8.85) and (8.88) of Theorem 8.8.1 are satisfied. By Theorem 8.8.1, a one-parameter family of periodic solutions of (8.21) bifurcates from $\mathbf{P}_+$. However, the bifurcation is neither supercritical nor subcritical. In fact, *all the bifurcating solutions lie in the plane*

$$\{(x, y, \rho) \in \mathbb{R}^2 \times \mathbb{R} \,:\, \rho = 1\}.$$

Indeed, these are the periodic solutions shown in Figure 7.2, which we discussed in Example 2 of Section 7.1.2. Equation (8.21) does not have periodic solutions for any $\rho \neq 1$. (Incidentally, phase portraits for (8.21) with $\rho = 0.9$ and 1.1 are shown in Figure 6.12.)

This behavior is very nongeneric, and virtually any perturbation will disrupt it. For example, recall that (8.21) was obtained as the limit $\kappa \to \infty$ of a more general activator–inhibitor model (5.26). Taking κ finite is a natural perturbation to investigate, and in Exercise 25, we ask you to show that the Hopf bifurcation of (5.26) at $\rho \approx 1$ is subcritical.

8.8.3 Sub/Supercriticality in Two Dimensions

In most applications, it is painful to determine analytically whether a Hopf bifurcation is subcritical or supercritical. Usually, we prefer to resort to numerics. For two-dimensional systems there is actually a "plug-and-chug" formula for a quantity whose sign determines this behavior. Unfortunately, a lot of calculation is required to evaluate it, which limits its usefulness in applications; but let's give it to you anyway.

Let $\mathbf{x}' = \mathbf{F}(\mathbf{x}, \mu)$ be a family of two-dimensional ODEs with an equilibrium that undergoes a Hopf bifurcation at $(\mathbf{x}, \mu) = (\mathbf{x}_*, \mu_*)$, where (8.85) and (8.88) are satisfied. Performing a linear change of coordinates, we may assume that

$$\mathbf{F}(\mathbf{x}, \mu_*) = \Omega_*(\mathbf{x} - \mathbf{x}_*) + \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}, \tag{8.92}$$

where

$$\Omega_* = \begin{bmatrix} 0 & -\omega_* \\ \omega_* & 0 \end{bmatrix}, \tag{8.93}$$

and all higher-order terms are put into $f(\mathbf{x}), g(\mathbf{x})$. The stability of bifurcating solutions depends on the following combination of quadratic and cubic terms in $f(\mathbf{x}), g(\mathbf{x})$:

$$\Gamma = f_{xxx}+f_{xyy}+g_{xxy}+g_{yyy}+\frac{1}{\omega_*}[f_{xy}(f_{xx}+f_{yy})-g_{xy}(g_{xx}+g_{yy})-f_{xx}g_{xx}+f_{yy}g_{yy}], \tag{8.94}$$

where we use subscripts to indicate partial derivatives of f and g with respect to the components x and y of $\mathbf{x}$, evaluated at the bifurcation point $(\mathbf{x}_*, \mu_*)$.

Theorem 8.8.3. *The Hopf bifurcation of* (8.92) *is supercritical if* $\Gamma < 0$ *and subcritical if* $\Gamma > 0$.

In Exercise 30 we outline a proof of this result in the special case in which all second-order derivatives in (8.94) vanish, and we give a reference for the general proof. In Exercise 26 we ask you to apply the theorem to the denatured Morris–Lecar model and confirm that the Hopf bifurcation found in Section 8.7.4 is subcritical. The tedious part of the latter exercise is performing the initial linear transformation to put $\mathbf{DF}_*$ into real normal form as in (8.92).

8.9 Exercises

After the core exercises, many specific applications of bifurcation theory are proposed in Section 8.9.2.

8.9.1 Core Exercises

The core exercises address the following issues:

Unfinished business	1, 3, 4, 12
Parametrization of bifurcating solutions	2, 13
Applying the bifurcation theorems	5, 6
More analysis of normal forms	7, 8
Counterexamples	9–11

1. (a) For the bifurcations of (8.5), (8.8), and (8.16), if you haven't already done so, verify that the predictions of Theorem 8.5.3 regarding exchange of stability are consistent with the relevant bifurcation diagrams in the text.

   (b) For the bifurcations of (8.2), (8.5), and (8.60), if you haven't already done so, verify that the hypotheses of Theorem 8.5.5 for symmetric bifurcations are satisfied; i.e., check (8.10) and (8.12).

2. Show that the bifurcating solutions of the Lotka–Volterra model (8.16) may be parametrized as in (8.13).

   Hint: If you choose $\mathbf{v} = (-1, 1)$ to span $\ker \mathbf{DF}_*$, then (8.13a) reads

   $$\mathbf{X}(a) = (K(a), 0) + a(-1, 1) + \mathcal{O}(a^2).$$

   On the other hand, from (8.17), the coexistence equilibrium is given by

   $$\mathbf{X}(a) = (1, 1 - 1/K(a)).$$

   Check that these two representations are consistent with the definition $K(a) = 1 + a$.

3. Rederive (8.47) from scratch.

   Remark: Focusing on this one small part of the Lyapunov–Schmidt reduction may help ease your way to a better understanding of the general theory.

4. (a) Show that both of the following conditions are equivalent to the hypothesis that the (square) matrix A has a simple eigenvalue zero:
   - dim ker $A = 1$ and ker $A \cap$ range $A = \{\mathbf{0}\}$.
   - dim ker $A^2 = 1$.

   (b) Prove Lemma 8.5.4; i.e., show that if $R^2 = I$, then the only possible eigenvalues of R are ± 1, and R is diagonalizable.

5. Write the Lorenz system (8.5) in vector notation $\mathbf{x}' = \mathbf{G}(\mathbf{x}, \rho)$, where $\mathbf{x} = (x, y, z)$. Given a square matrix S, consider a transformed unknown $\overline{\mathbf{x}} = S^{-1}\mathbf{x}$, which satisfies the ODE $\overline{\mathbf{x}}' = \mathbf{F}(\overline{\mathbf{x}}, \rho)$ with $\mathbf{F}(\overline{\mathbf{x}}, \rho) = S^{-1}\mathbf{G}(S\overline{\mathbf{x}}, \rho)$. Determine S so that at the bifurcation point, $\mathbf{DF}_*$ satisfies (8.26).

6. (a) Draw the graph, y vs. σ, defined by solving the reduced equation (8.30) for the chemostat, $g(y, \sigma) = 0$, where

$$g(y, \sigma) = (\sigma - y)\left(\frac{y}{y+1} - \rho\right).$$

   (b) Label each branch of your graph according as $\partial_y g$ is positive or negative there.

   (c) Along the equilibrium branch $x = 0$, $y = \sigma$ of (8.18) with $\sigma < \rho/(\rho + 1)$, determine whether the equilibrium is stable or unstable.

   (d) Recalling the addendum to Theorem 8.5.3, complete a bifurcation diagram for the chemostat by identifying stable and unstable branches on your graph.

7. Draw the analogues of Figure 8.7 for a subcritical pitchfork ($x' = x^3 + \mu x$), a saddle-node bifurcation ($x' = -x^2 + \mu$), and an unfolded hysteresis point ($x' = -x^3 + \mu + \alpha x$, where $\alpha > 0$).

   Remark: With this exercise we hope to make the ideas of exchange of stability so familiar that you are slightly bored with them.

8. (a) Draw the imperfect bifurcation diagrams for a subcritical pitchfork ($x' = x^3 + \mu x + \varepsilon$), transcritical bifurcation ($x' = -x^2 + \mu x + \varepsilon$), and saddle-node bifurcation ($x' = -x^2 + \mu + \varepsilon$). Articulate differences (if any) in the qualitative behavior under quasistatic variation of μ between the cases $\varepsilon > 0$, $\varepsilon = 0$, and $\varepsilon < 0$.

(b) Show that the branch of stable solutions for an imperfect subcritical pitchfork bifurcation terminates at

$$\mu = -3\left(\frac{\varepsilon}{2}\right)^{2/3}. \tag{8.95}$$

Hint: Implicit differentiation of $x^3 + \mu x + \varepsilon = 0$ with respect to x gives

$$3x^2 + \mu + x\frac{d\mu}{dx} = 0.$$

At the termination of the solution branch, $d\mu/dx = 0$, so this equation may be solved for x as a function of μ and substituted into the original equation.

Discussion: The fractional power in (8.95) is significant for engineering design. For example, the laterally supported pendulum in Figure 8.3 has a stable equilibrium for every mass less than $m_* = k\ell/g$. However, suppose the supporting spring is slightly misaligned so that the potential energy equals

$$\hat{V}(x) = mg\ell\cos x + k(\ell\sin x + \delta)^2/2.$$

Then, to lowest order in δ, the collapse load will be reduced by a quantity proportional to $\varepsilon^{2/3}$, where $\varepsilon = \delta/\ell$, and if ε is small, then $\varepsilon^{2/3} \gg \varepsilon$. In other words, a small imperfection can greatly reduce the predicted failure load of a structure. See [87] for a thorough analysis of such issues.

9. *Introduction:* This problem illustrates what can go wrong in the bifurcation result, Theorem 8.2.2, if the zero eigenvalue of $\mathbf{DF}_*$ is not simple, even if $\ker \mathbf{DF}_*$ is one-dimensional.

(a) Consider bifurcation of the trivial solution $x = y = 0$ of

$$\begin{aligned} x' &= -y, \\ y' &= (x^2 + \mu^2)x + \mu y, \end{aligned}$$

as μ crosses zero. Specifically, show that (i) zero is a repeated eigenvalue of $\mathbf{DF}_*$, (ii) the trivial solution loses stability as μ crosses zero, and (iii) no new equilibrium solutions bifurcate near $\mu = 0$.

(b) Consider bifurcation of the trivial solution $x = y = 0$ of

$$\begin{aligned} x' &= \mu x - y, \\ y' &= x^2 - \mu x, \end{aligned}$$

as μ crosses zero. Specifically, show that (i) zero is a repeated eigenvalue of $\mathbf{DF}_*$, (ii) the bifurcation diagram has two branches like a typical transcritical bifurcation, and (iii) the stabilities of these branches are not described by the principle of exchange of stability.

Discussion: These two examples, which were constructed to illustrate a mathematical point, are a little artificial. However, in real applications with many

parameters, a bifurcation in which zero is not a simple eigenvalue may be hidden in the parameter space. Indeed, such a nongeneric bifurcation occurs in the denatured Morris–Lecar equation (8.79); see footnote number 34. A problem that exhibits this degeneracy is called a *Takens–Bogdanov* bifurcation. Classification of all the phenomena that may appear in such bifurcations is beyond the scope of this book; see Section 20.6 of [95] for more information.

10. Draw the bifurcation diagram for

$$x' = x^3 - \mu^2 x$$

and determine the stabilities of all solution branches.

Discussion: This problem illustrates the need for hypothesis (8.11) in Theorem 8.2.2. The same issue is hidden in Part (a) of the preceding exercise: even though the eigenvalue at the bifurcation point in that problem is not simple, the equilibrium equations for this system may still be reduced to the analogous single equation $x^3 + \mu^2 x = 0$, which does not satisfy hypothesis (8.11).

Incidentally, a simple mechanical system that exhibits a bifurcation of the above form is presented and analyzed in Section VI.1 of [31].

11. *Introduction:* Applied naively, Theorem 8.5.5 might lead you to expect pitchfork bifurcation in any problem that has reflectional symmetry. However, the system

$$\begin{aligned} x' &= -x^2 - x + \mu y, \\ y' &= -y^2 - y + \mu x, \end{aligned}$$

which is symmetric under the reflection $R \cdot (x, y) = (y, x)$, warns that you must examine such problems more carefully before jumping to this conclusion.

(a) Verify that the hypotheses (8.10) and (8.11) are satisfied at the bifurcation of this system from the trivial solution $x = y = 0$ when $\mu = 1$.

Remark: Even the stronger hypothesis (8.51) is satisfied.

(b) Show that a branch of solutions with $x = y$ bifurcates *transcritically* from the zero solution at $\mu = 1$.

(c) Deduce from Theorem 8.2.2 that the zero solution and your solution in Part (b) are the only equilibria near $(x, y, \mu) = (0, 0, 0)$.

(d) Show that the null eigenvector $\mathbf{v}$ of $\mathbf{DF}_*$ satisfies $R\mathbf{v} = +\mathbf{v}$. (Cf. Condition (iv) of Theorem 8.5.5.)

Remark: You might find it amusing to check that this system has another bifurcation for $\mu < 0$ at which the hypotheses of Theorem 8.5.5 are satisfied.

12. For the Hopf bifurcations of (8.71), (8.74), and (8.78), if you have not already done so, verify that conditions (8.85) and (8.88) hold.

13. *Introduction:* The following exercise is intended to help you understand (8.89), the parametrization of bifurcating solutions at a Hopf bifurcation, in the context of the repressilator (8.74). Recall that periodic solutions bifurcate from the equilibrium when $\mu = \mu_* = 2$.

    (a) For $\mu = \mu_* + 10^{-k/2}$, $k = 2, 3, 4, 5$, compute the periodic solution of (8.74), say $\boldsymbol{\gamma}_\mu(t)$.

    (b) Find a basis $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ of $\mathbb{R}^3$ such that the similarity matrix $S = \mathrm{Col}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ transforms the Jacobian $\mathbf{DF}_*$ at the bifurcation point to the real canonical form

$$\begin{bmatrix} 0 & -\omega_* & 0 \\ \omega_* & 0 & 0 \\ 0 & 0 & -3 \end{bmatrix}.$$

    (c) For each value of μ as above, adjust the phase of $\boldsymbol{\gamma}(t)$ and find the best choices for a, ω to minimize

$$\sup_t \left| \boldsymbol{\gamma}_\mu(t) - \mathbf{x}_{\mathrm{eq}}(\mu) - a[(\cos \omega t)\mathbf{v}_1 + (\sin \omega t)\mathbf{v}_2] \right|. \tag{8.96}$$

    (d) Process your data in Part (c) to (i) show that the error (8.96) is $\mathcal{O}(a^2)$ and $\omega(a) - 1 = \mathcal{O}(a^2)$ and (ii) estimate the coefficient μ_2 in the Taylor-series expansion of $\mu(a)$.

8.9.2 Applications of Bifurcation Theory

Exercises 14–17 involve steady-state bifurcation; 18–21 involve Hopf bifurcation; and both types appear in Exercise 22. Let your interests guide which of these many problems you do.

14. *Introduction:* By the term *guyed tower* we mean an inverted pendulum supported by springs on either side, as illustrated in Figure 8.21. Suppose m, the mass of the system, is concentrated at the end of the pendulum. Define a, ℓ, α as in Panel (a), so $\tan \alpha = \ell/a$. Suppose both springs are linear with spring constant k, and let them have unstretched length L, where $L < \sqrt{\ell^2 + a^2}$. Thus, if the tower makes an angle x with the vertical, the potential energy in the spring on the *left* is

$$\frac{k}{2}\left(\sqrt{\ell^2 + a^2 + 2\ell a \sin x} - L\right)^2.$$

    The length of the spring, the expression under the radical, is derived from the law of cosines; the angle opposite the spring is $\pi/2 + x$, and $\cos(\pi/2 + x) = -\sin x$.

    (a) Write the total potential energy of the system as the sum of three terms, the stretching energy of the two springs plus the gravitational potential energy of the mass m.

    (b) Choose an appropriate nondimensionalization of your expression for the energy.

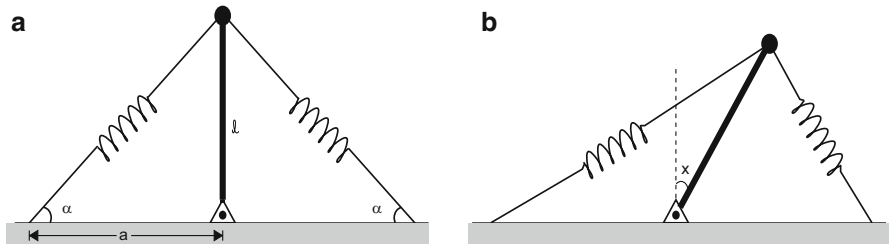

Figure 8.21: *(a) Schematic diagram of a guyed tower for Exercise 14, in its equilibrium configuration. (b) A deformed configuration, illustrating the definition of the angle x.*

(c) Examine the quadratic term in the energy to find the value of the mass at which bifurcation occurs.

(d) Examine the quartic term in the energy to show that the bifurcation is subcritical if $\sin 2\alpha < 2/\sqrt{5}$, i.e., if $\alpha < 31.7^\circ$ or $\alpha > 58.3^\circ$.

15. *Introduction:* The equations

$$\begin{aligned} x' &= \frac{\mu}{1+y^4} - x, \\ y' &= \frac{\mu}{1+x^4} - y, \end{aligned} \tag{8.97}$$

are a two-variable analogue of the repressilator (8.74); i.e., Y inhibits the production of X, and X inhibits the production of Y. See also [29].

(a) Determine the value of μ where the equilibrium with $x = y$ loses stability and verify hypotheses (8.10) and (8.12).

Discussion: Recall that in Exercise 6.2(a) we asked whether there were any equilibria of these equations off the diagonal $\{x = y\}$. Without our having to fiddle with the equations, the application of Theorem 8.2.2 guarantees the existence of such solutions. These equations are symmetric under the interchange of x and y, and because of this symmetry, the bifurcation is a pitchfork.

(b) Use the computer to show that the bifurcation is supercritical.

(c) Sketch the bifurcation diagram for the above system. Plot both x vs. μ, which hides the symmetry in the problem, and $x - y$ vs. μ, in which the symmetry is evident.

Discussion: To make a reasonable sketch, it suffices to know that the bifurcation is supercritical and to determine analytically the behavior of solutions as $\mu \to \infty$.

(d) Recalling (8.57) as a model for imperfect bifurcation, make a prediction about how your bifurcation diagram will change if you break the symmetry between the equations, say changing one of the production rates

$$\begin{aligned} x' &= \frac{(1+\varepsilon)\mu}{1+y^4} - x, \\ y' &= \frac{\mu}{1+x^4} - y. \end{aligned}$$

(e) Test your prediction numerically for some nonzero choice of ε.

Discussion: The above system is sometimes called a *bioswitch* (cf. Section 12.2 of [70]). This term arises in the following context. Virtually every cell in a living organism contains the entire genome. Differences between cells arise when different combinations of genes are turned on or off. If X and Y describe the level of activity of two genes, a gene circuit as above provides a mechanism for choosing which gene is turned on and which turned off, i.e., adjusting initial conditions and/or adjusting the balance between production rates.

Despite the similarity of (8.97) to the repressilator, here is a surprising contrast: while breaking symmetry in (8.97) changes the qualitative nature of the bifurcation, breaking symmetry in (8.74) does not—stable periodic solutions bifurcate from an equilibrium even if the production coefficients μ_k and the decay rates differ in the three equations.

16. *Introduction:* Recall from Exercise 6.5 the two-predator system

$$\begin{aligned} x' &= x(1-x/K) - xy - xz, \\ y' &= \rho_1 xy - \delta_1 y, \\ z' &= \rho_2 xz - \delta_2 z. \end{aligned} \tag{8.98}$$

In Chapter 6 we ducked the task of listing the stabilities of all the equilibria as functions of the parameters. There were just too many cases, and it seemed little understanding would be gained from all the calculations. Bifurcation theory gives us an effective way to organize and understand the phenomena.

Assuming $\delta_1/\rho_1 < \delta_2/\rho_2$, analyze the two bifurcations—i.e., locate and determine their types—from the prey-only equilibrium $(K, 0, 0)$ of (8.98) as K increases.

Discussion: Note that at the second bifurcation, where $K = \delta_2/\rho_2$, exchange of stability takes a slightly different form. At this bifurcation the prey-only equilibrium and one of the coexistence equilibria cross each other. However, *neither branch is stable on either side of the bifurcation point.* Below the bifurcation point, the prey-only equilibrium has one unstable eigenvalue and the (unphysical) coexistence equilibrium has two, while above the bifurcation point these numbers are reversed.

17. *Introduction:* In Section 9.6.5 we shall study dynamics in Rössler's equations,

$$\begin{aligned} x' &= -y - z, \\ y' &= x + ay, \\ z' &= b + z(x-c), \end{aligned} \tag{8.99}$$

where a, b, c are positive parameters, as a bifurcation problem with c as the bifurcation parameter. Here we consider only steady-state bifurcation.

(a) Find a critical value c_* such that for $c < c_*$, equation (8.99) has no equilibria, but two equilibria appear through a saddle-node bifurcation as c crosses c_*.

Remark: If $c < c_*$, solutions of (8.99) grow without bound as $t \to \infty$, and even if $c > c_*$, some solutions are still unbounded. This is most easily shown numerically.

(b) Show that if $a = b$, then at the bifurcation point, $\mathbf{DF}_*$ has eigenvalues 0 and $\pm i\sqrt{2-a^2}$. (In particular, if $a < \sqrt{2}$, then hypothesis (8.10) is not satisfied.)

Hint: At the bifurcation point, one eigenvalue vanishes, say $\lambda_1(\mathbf{DF}_*) = 0$. Calculate that $\mathrm{tr}\mathbf{DF}_* = 0$, which implies that $\lambda_2 + \lambda_3 = 0$. Determine $\lambda_{2,3}$ by calculating $\mathrm{tr}\mathbf{DF}_*^2$.

Remark: The dedicated reader may show that for c just beyond the saddle-node bifurcation, if $a < b$, then one equilibrium branch is stable and the other unstable, but if $b < a$, then both branches are unstable.

18. *Advice:* Here is a gentle introduction to Hopf-bifurcation calculations. Feel free to skip it if you are ready to handle stronger fare.

Locate any Hopf bifurcations from the equilibrium solution with $x = y = z$ in (8.74) with an arbitrary Hill exponent n:

$$\begin{aligned} x' &= \frac{\mu}{1+y^n} - x, \\ y' &= \frac{\mu}{1+z^n} - y, \\ z' &= \frac{\mu}{1+x^n} - z. \end{aligned}$$

19. *Introduction:* Recall Sel'kov's model for glycolysis,

$$\begin{aligned} x' &= \rho - \sigma x - xy^2, \\ y' &= -y + \sigma x + xy^2. \end{aligned}$$

In Exercise 6.2(b) you found that (i) for all positive σ, ρ, the system has a unique equilibrium in the first quadrant and (ii) for the specific parameter value $\sigma = 0.1$, this equilibrium is unstable for ρ in an interval $\rho_1 \le \rho \le \rho_2$, and stable otherwise.

(a) Show that this equilibrium undergoes a Hopf bifurcation as ρ crosses either ρ_1 or ρ_2; in particular, verify (8.85) and (8.88) at both bifurcation points.

(b) Use the computer to make a bifurcation diagram of the periodic solutions that exist between these two values of ρ.

20. *Introduction:* In this exercise you consider the van der Pol equation as an example of Hopf bifurcation.

    (a) Show that the trivial solution $x = y = 0$ of

    $$\begin{aligned} x' &= y, \\ y' &= -x - (x^2 - \mu)y, \end{aligned}$$

    undergoes a Hopf bifurcation as μ crosses zero.

    (b) For $\mu > 0$, rescale variables to put this equation into the usual form

    $$\begin{aligned} x' &= y, \\ y' &= -x - \beta(x^2 - 1)y. \end{aligned}$$

    Challenge: Even before doing the rescaling, can you anticipate how β depends on μ as $\mu \to 0$? I.e., does β tend to infinity, remain bounded away from both zero and infinity, or tend to zero?

21. *Introduction:* Recall that if $\rho > 1$ in the Lorenz equations (8.5), then the equations have two nontrivial equilibria, as discussed in Section 8.1.

    (a) Show that if $\sigma - \beta - 1 < 0$, then these equilibria are asymptotically stable for all $\rho > 1$.

    (b) Show that if $\sigma - \beta - 1 > 0$, the equilibria are asymptotically stable if $1 < \rho < \rho_*$ and unstable if $\rho > \rho_*$, where

    $$\rho_* = \frac{\sigma(\sigma + \beta + 3)}{\sigma - \beta - 1}.$$

    Specifically, show that the equilibria undergo Hopf bifurcation at $\rho = \rho_*$, with hypotheses (8.91) and (8.88) being satisfied.

    Hint: You may find it helpful to refer back to Proposition 2.4.6.

22. *Introduction:* The Rosenzweig–MacArthur equations describe another predator–prey model. After scaling, the equations are

    $$\begin{aligned} x' &= x(1 - x) - \frac{xy}{1 + Sx}, \\ y' &= \frac{Exy}{1 + Sx} - \mu y, \end{aligned} \tag{8.100}$$

    where S, E, μ are positive parameters. The model has a finite carrying capacity for the prey, which has been scaled to unity; thus, for all parameter values, $(1, 0)$ is an equilibrium. The primary new effect in this model is that because of the factor $1 + Sx$ in the denominators, predation saturates if the prey population is large; indeed, the letter "S" is a mnemonic for "saturation." (The letter "E" is a mnemonic for the "efficiency" with which the predators convert the biomass of the prey into their own biomass.)

(a) Considering this system as a bifurcation problem with the predator death rate μ as bifurcation parameter, show that the prey-only solution $(1, 0)$ is stable for large μ, say $\mu > \mu_0$, but loses stability when μ is decreased below μ_0.

(b) Find the coexistence equilibria in these equations. Show that this equilibrium bifurcates from the prey-only equilibrium in a transcritical bifurcation at $\mu = \mu_0$.

(c) Use the principle of exchange of stability to predict, for μ near μ_0, where the coexistence equilibrium is stable and where it is unstable.

(d) Verify that your prediction in Part (c) is correct.

Remark: In the remainder of the exercise, you will see that although these stability predictions are accurate locally near the bifurcation point, they may not hold globally.

(e) Choose a value of $S > 1$ in (8.100) and use the computer to study the stability of the coexistence equilibrium as μ decreases. You will find a Hopf bifurcation at some point $\mu = \mu_* > 0$.

Remark: If $S < 1$, the coexistence equilibrium is stable, no matter how small μ gets.

(f) Calculate (analytically) the bifurcation point μ_* and verify hypotheses (8.91) and (8.88) there.

8.9.3 PHD Exercises

We hope you will be stimulated by at least some of the following problems. Although they are long, they offer opportunities to push much further into the subject. But please don't let yourself feel oppressed by them—they are not mandatory. Exercises 23–27 follow up on material in the chapter; 28–30 introduce theoretical results from dynamical systems; and 31, 32 introduce more advanced ideas from bifurcation theory.

23. *Introduction:* In this exercise we outline calculations to support the bifurcation diagrams of the CSTR plotted in Figure 8.14. In analyzing the equations, it is most effective to consider μ as a function of y.

(a) Show that if $\xi = 0$, the equilibrium equation for (8.70) can be reduced to the form

$$A(y)\mu^2 - B(y)\mu + C(y) = 0, \tag{8.101}$$

where

$$A(y) = ye^{-y}, \qquad B(y) = \sigma - y - \beta ye^{-y}, \qquad C(y) = \beta y.$$

Hint: Solve the first equilibrium equation for x as a function of y and μ, substitute this into the second equation, and rearrange.

(b) In the equation $B^2(y) - 4A(y)\,C(y) = 0$ for the discriminant of (8.101) to vanish, solve for β to obtain

$$\beta = \left[\sqrt{\frac{\sigma}{y}} \pm 1\right]^2 e^y. \tag{8.102}$$

(c) Verify that as shown in Figure 8.22, if $\sigma = 8$, the graph of (8.102) with the minus sign has a rising portion, $y_{\text{bif}} < y < y_{\text{iso}}$, sandwiched between two falling portions.

Hint: Draw the graph of $(\sqrt{\sigma/y} - 1)^2$. Although it decreases over the entire interval $0 < y < \sigma$, the exponential factor in (8.102) can create a range of y where this function increases. To test for this, differentiate (8.102). You will get a positive factor times a cubic in $1/\sqrt{y}$. The cubic is negative as $y \to 0$ and $y \to \sigma$, but you can show that if $\sigma = 8$, it is positive on an interval between these extremes.

(d) Correlate the information in Figure 8.22 with the bifurcation diagrams in Figure 8.14, as follows:

$$\begin{array}{cc} \beta < \beta_{\text{bif}}, & \text{Panel (a)} \\ \beta_{\text{bif}} < \beta < \beta_{\text{iso}}, & \text{Panels (b) and (c)} \\ \beta > \beta_{\text{iso}}, & \text{Panel (d).} \end{array}$$

Hint: If y, β lie below the curve in Figure 8.22, the quadratic formula provides two positive real roots of (8.101). If y, β lie somewhat above it, (8.101) has no real roots. (Still higher in the y, β-plane, (8.101) again has real roots, but these are negative, so they have no physical significance.) Holding β fixed in each of the three ranges, ask how the number of roots of (8.101) varies as y increases from 0 to σ.

24. *Introduction:* Equation (8.60) for two-cell Turing instability has several bifurcations for which $D < 0$. Although these have no physical significance, exploring them gives you an opportunity to practice applying the theory. In this problem assume $\sigma > 2$ and $\rho > 1$.

(a) Class 1 equilibria of (8.60) (i.e., solutions for which $x_1 = x_2 = 0$) undergo a pitchfork bifurcation for which $D < 0$. Find it and determine whether it is supercritical, subcritical, or degenerate.

(b) The trivial equilibria (x_+, y_+, x_+, y_+) of (8.60) (i.e., solutions of Class $3a_+$) are stable for $0 \le D < D_*$ and for a range with $D < 0$; they lose stability through a Hopf bifurcation for some negative value of D. Find the bifurcation point.

(c) Solutions of (8.60) of Class 3b exist only in the range $-D_* < D < D_*$. We saw in Section 8.6.1 that at the upper limit of this range, these solutions bifurcate from the trivial solution, Class $3a_+$. Show that at the lower limit they bifurcate from the "other" trivial solution, Class $3a_-$.

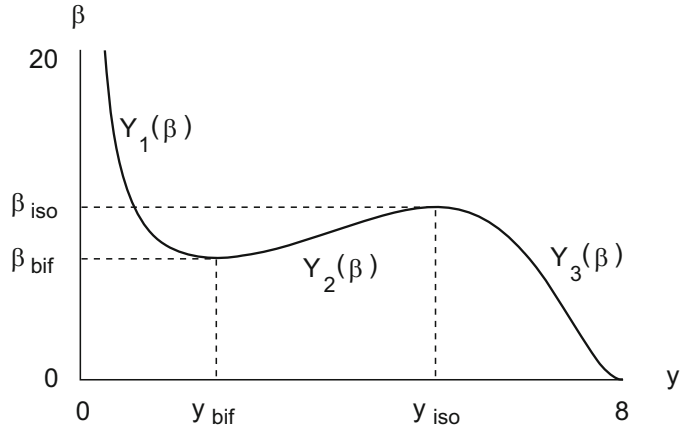

Figure 8.22: *The curve* (8.102) *with the minus sign, along which the discriminant of* (8.101) *vanishes. The graph is calculated assuming* $\sigma = 8$.

Discussion: Figure 8.23 shows a bifurcation diagram that summarizes information from Parts (b) and (c) of this exercise. Note that the number of unstable eigenvalues of the Jacobian changes along the curve of solutions of Class 3b; this indicates that a Hopf bifurcation occurs somewhere along this solution branch, meaning that the equations have some (unstable) periodic solutions we have not investigated.

25. Consider (5.26), the activator–inhibitor model with finite κ, as a bifurcation problem with bifurcation parameter ρ. Assume $\sigma = 2.1$, $\kappa = 4$. Show numerically that (i) for ρ sufficiently large, the origin and the "top equilibrium" are stable, and (ii) as ρ decreases, the top equilibrium undergoes a *subcritical* Hopf bifurcation near $\rho = 1$.

Remark: The point of this exercise is to see that a natural perturbation of (8.21) removes the nongeneric behavior in the bifurcation of this system found in Section 8.8.2.

26. *Introduction:* In this exercise, Theorem 8.8.3 is used to show that the Hopf bifurcation of the denatured Morris–Lecar model is subcritical. We rewrite (8.79) with x, y near the bifurcation point (8.83) and with $I = I_*$ in the abstract form

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \mathbf{DF}_* \begin{bmatrix} x - x_* \\ y - y_* \end{bmatrix} + \begin{bmatrix} f(x) \\ g(x) \end{bmatrix}. \tag{8.103}$$

Note that the higher-order terms $f(x), g(x)$ depend only on x, reflecting the fact that both equations in (8.79) are linear in y. A simple calculation shows that at the bifurcation point,

$$f_{xx} = 2 - 6x_* \approx -1.4832, \qquad f_{xxx} = -6, \qquad g_{xx} = \gamma y_* \alpha^2 \approx 1.1718. \tag{8.104}$$

(a) Show that the hypothesis (8.88) is satisfied for (8.79) at the point (8.83).

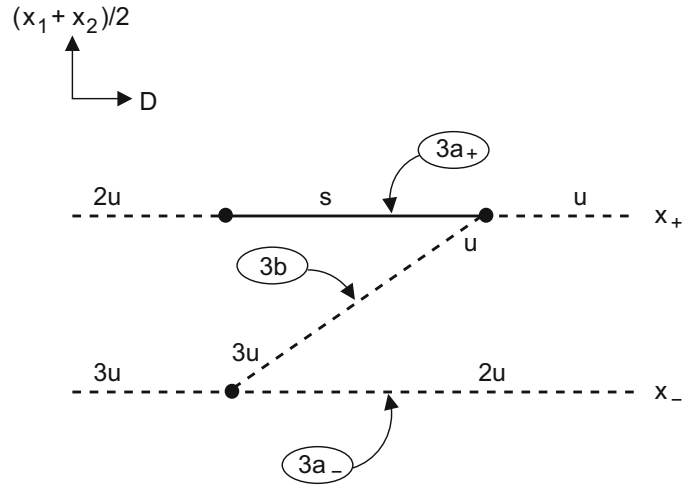

Figure 8.23: *Diagram showing one Hopf and two steady-state bifurcations for equilibria of* (8.60) *of Classes 3a and 3b. (Cf. Exercise 24.) Despite appearances, both steady-state bifurcations are pitchfork, not transcritical. When an equilibrium is unstable, the notation* u, 2u, *or* 3u *is used to indicate that the Jacobian has one, two, or three eigenvalues in the right half-plane, respectively. For solutions of Class 3a, this information is obtained from* (6.30). *For solutions of Class 3b, it is obtained (locally near a bifurcation point) from the principle of exchange of stability.*

(b) *Introduction:* To put the equation in the form (8.92), make the linear transformation defined by

$$S\begin{bmatrix} \overline{x} \\ \overline{y} \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} \qquad \text{where} \qquad S = \begin{bmatrix} 1 & 0 \\ \gamma & \omega_* \end{bmatrix};$$

i.e., the columns of S are the real and imaginary parts of the eigenvector of $\mathbf{DF}_*$ with eigenvalue $-i\omega_*$. Thus $\overline{x} = x$ and $\overline{y} = (-\gamma x + y)/\omega_*$.

Show that the new variables[35] $(x, \overline{y})$ satisfy the ODE

$$\begin{bmatrix} x' \\ \overline{y}' \end{bmatrix} = \Omega_* \begin{bmatrix} x - x_* \\ \overline{y} - \overline{y_*} \end{bmatrix} + \begin{bmatrix} f(x) \\ \overline{g}(x) \end{bmatrix}, \tag{8.105}$$

where $\overline{y_*} = (-\gamma x_* + y_*)/\omega_*$ and $\overline{g}(x) = (-\gamma f(x) + g(x))/\omega_*$.

(c) *Introduction:* Since the nonlinear terms in (8.105) are independent of y, many terms in (8.94) are zero. Thus, for (8.105), equation (8.94) reduces to

$$\Gamma = f_{xxx} - \frac{1}{\omega_*}[f_{xx}\overline{g}_{xx}].$$

Recalling (8.104), show that $\Gamma \approx 2.1442 > 0$, so by Theorem 8.8.3 the bifurcation is subcritical.

27. *Introduction:* In Section 8.7.2 we found a Hopf bifurcation in the repressilator (8.74), which is three-dimensional, and in Exercise 15 you found a steady-state bifurcation in a two-dimensional analogue of the equations.

Analyze what kind of bifurcation, steady-state or Hopf, occurs in four- and five-dimensional analogues of the repressilator.

28. *Introduction:* The following exercise shows you part of the calculation in reducing a general two-dimensional Hopf bifurcation to the normal form (8.73). It doesn't allow for a bifurcation parameter, and it addresses only quadratic terms, but it introduces the clever changes of coordinate that are the basis of the general proof. Sections 3.3 and 3.4 of [33] cover the general reduction.

In the two-dimensional ODE

$$\mathbf{x}' = L_*\mathbf{x} + \mathbf{P}(\mathbf{x}) + \mathcal{O}(|\mathbf{x}|^3),$$

suppose that the matrix L_* has eigenvalues $\pm i\omega_*$ and that $\mathbf{P}$ is a homogeneous quadratic term. Show that there is a change of coordinates such that the transformed ODE has the form

$$\mathbf{y}' = \Omega_*\mathbf{y} + \mathcal{O}(|\mathbf{y}|^3),$$

[35] Since $\overline{x} = x$, we may omit the bar over x.

where Ω_* is given by (8.93).

Hint: By performing a preliminary *linear* change of coordinates, you may assume without loss of generality that $L_* = \Omega_*$. Using the slick technique of Section 7.9.4, write the coordinate transformation in the form $\mathbf{x} = \mathbf{y} + \mathbf{q}(\mathbf{y})$, where $\mathbf{q}$ is a homogeneous quadratic vector-valued function. Since $\mathbf{q} = \mathcal{O}(|\mathbf{y}|^2)$, this change of coordinates is (locally) invertible. (Indeed, $\mathbf{y} = \mathbf{x} - \mathbf{q}(\mathbf{x}) + \mathcal{O}(|\mathbf{x}|^3)$, but this information is not needed for our calculation.) This substitution leads to the transformed ODE (cf. (7.95))

$$\mathbf{y}' = (I + \mathbf{D}\mathbf{q}(\mathbf{y}))^{-1}[\Omega_*(\mathbf{y} + \mathbf{q}(\mathbf{y})) + \mathbf{P}(\mathbf{y} + \mathbf{q}(\mathbf{y}))]. \tag{8.106}$$

In preparation for neglecting higher-order terms, note that

$$(I + \mathbf{D}\mathbf{q}(\mathbf{y}))^{-1} = I - \mathbf{D}\mathbf{q}(\mathbf{y}) + \mathcal{O}(|\mathbf{y}|^2), \qquad \mathbf{P}(\mathbf{y} + \mathbf{q}(\mathbf{y})) = \mathbf{P}(\mathbf{y}) + \mathcal{O}(|\mathbf{y}|^3).$$

Using this information, (8.106) may be simplified to

$$\mathbf{y}' = \Omega_*\mathbf{y} \ - \mathbf{D}\mathbf{q}(\mathbf{y})\Omega_*\mathbf{y} + \Omega_*\mathbf{q}(\mathbf{y}) + \mathbf{P}(\mathbf{y}) \ + \mathcal{O}(|\mathbf{y}|^3).$$

Thus, given $\mathbf{P}$, you need to choose $\mathbf{q}$ such that

$$\mathbf{D}\mathbf{q}(\mathbf{y})\Omega_*\mathbf{y} - \Omega_*\mathbf{q}(\mathbf{y}) = \mathbf{P}(\mathbf{y}).$$

Note that both $\mathbf{P}$ and $\mathbf{q}$ belong to the six-dimensional space—call it V—of quadratic maps from $\mathbb{R}^2$ into itself. Thus such a $\mathbf{q}$ is guaranteed to exist, provided the linear operator on V given by

$$\mathbf{q} \mapsto \mathbf{D}\mathbf{q}(\mathbf{y})\Omega_*\mathbf{y} - \Omega_*\mathbf{q}(\mathbf{y}) \tag{8.107}$$

is invertible.

Write out components $\mathbf{y} = (y, z)$. Your first task is to show that with respect to the basis $y^2\mathbf{e}_1, yz\mathbf{e}_1, z^2\mathbf{e}_1, y^2\mathbf{e}_2, yz\mathbf{e}_2, z^2\mathbf{e}_2$ for V, the linear transformation (8.107) is represented by the matrix in block form

$$\omega_* \begin{bmatrix} M & I \\ -I & M \end{bmatrix}, \tag{8.108}$$

where

$$M = \begin{bmatrix} 0 & -2 & 0 \\ 1 & 0 & -1 \\ 0 & 2 & 0 \end{bmatrix}.$$

To get started, observe that

$$\mathbf{D}(y^2\mathbf{e}_1)\Omega_*\mathbf{y} = \begin{bmatrix} 2y & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & -\omega_* \\ \omega_* & 0 \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = \omega_*(-2yz\mathbf{e}_1)$$

and

$$-\Omega_*(y^2\mathbf{e}_1) = \begin{bmatrix} 0 & \omega_* \\ -\omega_* & 0 \end{bmatrix} \begin{bmatrix} y^2 \\ 0 \end{bmatrix} = \omega_*(-y^2\mathbf{e}_2).$$

Do similar calculations for each element of the basis and compile your results.

Having verified (8.108), show that this 6×6-matrix is invertible, and you're home free.

Remark: As shown in Sections 3.3 and 3.4 of [33], this construction can be continued at higher orders. However, already at the cubic level it is not possible to transform away all cubic terms. The normal form (8.73) represents one choice, perhaps the most natural choice, of which cubic terms to live with.

29. *Remark:* The result of this otherwise uninteresting exercise will be needed to complete Exercise 30.

Let Ω_* be the matrix (8.93). Show that if $k \neq 1$, the system

$$\left(\frac{d}{d\tau} - \Omega_*\right)\mathbf{x} = \begin{bmatrix} a\cos k\omega_*\tau + b\sin k\omega_*\tau \\ c\cos k\omega_*\tau + d\sin k\omega_*\tau \end{bmatrix} \tag{8.109}$$

has a periodic solution for every choice of coefficients a, b, c, d, and that if $k = 1$, it has a periodic solution iff

$$d = -a \quad \text{and} \quad c = b. \tag{8.110}$$

Hint: Consider the operator $d/d\tau - \Omega_*$ mapping the four-dimensional space of functions of the form

$$\begin{bmatrix} A\cos k\omega_*\tau + B\sin k\omega_*\tau \\ C\cos k\omega_*\tau + D\sin k\omega_*\tau \end{bmatrix}$$

into itself. If $k \neq 1$, this operator is invertible. If $k = 1$, it has a two-dimensional kernel; hence (8.109) will have a periodic solution only if the RHS of the equation belongs to the (two-dimensional) range of this operator, because otherwise, secular terms will arise. Observe that

$$\left(\frac{d}{d\tau} - \Omega_*\right)\begin{bmatrix} 0 \\ \alpha\cos\omega_*\tau + \beta\sin\omega_*\tau \end{bmatrix} = \omega_*\begin{bmatrix} \alpha\cos\omega_*\tau + \beta\sin\omega_*\tau \\ \beta\cos\omega_*\tau - \alpha\sin\omega_*\tau \end{bmatrix}.$$

Combining these ideas, deduce that for $k = 1$, the range of $d/d\tau - \Omega_*$ is defined by the two equations of (8.110).

30. Check the steps in the following outline for a proof of Theorem 8.8.3 in case all second-order terms in (8.94) vanish.

Remarks: We propose a calculational proof based on the Poincaré–Lindstedt technique from Section 7.5. Not having quadratic terms simplifies (a little) the calculations; but if you want to try the general case, it's not that much more difficult. The full result is proved in the appendix to Section 3.4 of [33] using a Taylor series expansion.

Outline: Assume without loss of generality that $\mathbf{x}_* = \mathbf{0}$ and $\mu_* = 0$. Introduce notation that makes more low-order terms in (8.92) explicit:

$$\mathbf{x}' = [\Omega_* + \mu L]\mathbf{x} + \mathbf{C}(\mathbf{x}) + \mathcal{O}(|\mathbf{x}|^4, \mu|\mathbf{x}|^2, \mu^2|\mathbf{x}|), \tag{8.111}$$

where L is a 2×2 matrix with $\mathrm{tr}L \neq 0$ *(Why? Recall* (8.88)*.)* and

$$\mathbf{C}(\mathbf{x}) = \frac{1}{6}\begin{bmatrix} f_{xxx}x^3 + 3f_{xxy}x^2y + 3f_{xyy}xy^2 + f_{yyy}y^3 \\ g_{xxx}x^3 + 3g_{xxy}x^2y + 3g_{xyy}xy^2 + g_{yyy}y^3 \end{bmatrix}.$$

Following the notation in (8.89), look for an expansion[36] of a solution with period $2\pi/\omega_*$,

$$\boldsymbol{\gamma}(t,a) = a\mathbf{x}_1(\tau) + a^3\mathbf{x}_3(\tau) + \dots, \qquad \mu(a) = \mu_2 a^2 + \dots,$$

where $\tau(t) = (1 + \omega_2 a^2 + \dots)t$. Note that the trivial solution $\mathbf{x} = \mathbf{0}$ of (8.111) is unstable if $\mu\,\mathrm{tr}L > 0$. On the other hand, the bifurcating solutions exist when $\mu = \mu_2 a^2 + \dots$, i.e., when μ has the same sign as the constant μ_2. Thus, the bifurcation is supercritical, i.e., the bifurcating solutions exist when the trivial solution is unstable, if $\mathrm{tr}L$ and μ_2 have the same sign, and subcritical if they have opposite signs. Below, a calculation is described that shows that

$$\mu_2\,\mathrm{tr}L = -\Gamma/8, \tag{8.112}$$

where $\Gamma = f_{xxx} + f_{xyy} + g_{xyy} + g_{yyy}$. Thus, the bifurcation is super- or subcritical according as Γ is negative or positive, respectively.

Recognizing that $d\boldsymbol{\gamma}/dt = (1 + \omega_2 a^2 + \dots)d\boldsymbol{\gamma}/d\tau$, substitute the ansatz into (8.111) and extract the leading-order equations:

$$\begin{aligned} &\mathcal{O}(a): && (d/d\tau - \Omega_*)\mathbf{x}_1 = \mathbf{0} \\ &\mathcal{O}(a^3): && (d/d\tau - \Omega_*)\mathbf{x}_3 = -\omega_2 d\mathbf{x}_1/d\tau + \mu_2 L\mathbf{x}_1 + C(\mathbf{x}_1). \end{aligned}$$

The solution of the $\mathcal{O}(a)$-equation may be taken to be $\mathbf{x}_1(\tau) = (\cos(\omega_*\tau), \sin(\omega_*\tau))$; amplitude and phase do not matter. The crux of the calculation is to require that the $\mathcal{O}(a^3)$-equation not generate secular terms. As follows from Exercise 29, this requires that appropriate nonresonance conditions hold, which may be derived as follows. The first two terms on the RHS of the $\mathbf{x}_3$-equation are given by

$$d\mathbf{x}_1/d\tau = \omega_* \begin{bmatrix} -\sin(\omega_*\tau) \\ \cos(\omega_*\tau) \end{bmatrix}, \qquad L\mathbf{x}_1 = \begin{bmatrix} L_{11}\cos(\omega_*\tau) + L_{12}\sin(\omega_*\tau) \\ L_{21}\cos(\omega_*\tau) + L_{22}\sin(\omega_*\tau) \end{bmatrix},$$

where L_{jk} denotes the elements of L. The third term, $\mathbf{C}(\mathbf{x}_1)$, equals

$$\frac{1}{6}\begin{bmatrix} f_{xxx}\cos^3(\omega_*\tau) + 3f_{xxy}\cos^2(\omega_*\tau)^2\sin(\omega_*\tau) + 3f_{xyy}\cos(\omega_*\tau)\sin^2(\omega_*\tau) + f_{yyy}\sin^3(\omega_*\tau) \\ g_{xxx}\cos^3(\omega_*\tau) + 3g_{xxy}\cos^2(\omega_*\tau)^2\sin(\omega_*\tau) + 3g_{xyy}\cos(\omega_*\tau)\sin^2(\omega_*\tau) + g_{yyy}\sin^3(\omega_*\tau) \end{bmatrix}.$$

With the trig identities $\cos^3 z = (3/4)\cos z + (1/4)\cos 3z$ and similar expressions for the other trig functions, reduce $\mathbf{C}(\mathbf{x}_1)$ to the form

$$\frac{1}{8}\begin{bmatrix} (f_{xxx} + f_{xyy})\cos(\omega_*\tau) + (f_{xxy} + f_{yyy})\sin(\omega_*\tau) \\ (g_{xxx} + g_{xyy})\cos(\omega_*\tau) + (g_{xxy} + g_{yyy})\sin(\omega_*\tau) \end{bmatrix}$$

plus various terms proportional to $\cos(3\omega_*\tau)$ or $\sin(3\omega_*\tau)$, the latter causing no secular terms. Combining the above expressions, write the problematic terms on RHS of the $\mathbf{x}_3$-equation in the form (8.109), where

$$\begin{aligned} a &= \mu_2 L_{11} + (f_{xxx} + f_{xyy})/8, & b &= \omega_2\omega_* + \mu_2 L_{12} + (f_{xxy} + f_{yyy})/8 \\ c &= -\omega_2\omega_* + \mu_2 L_{21} + (g_{xxx} + g_{xxy})/8, & d &= \mu_2 L_{22} + (g_{xyy} + g_{yyy})/8. \end{aligned}$$

Thus, (8.112) emerges as the first solvability condition in (8.110). (The second solvability condition provides a formula for ω_2.)

Honestly, we'd hate the Poincaré–Lindstedt method if it weren't so useful.

[36] It will appear below that there is no need for a term proportional to a^2 in the expansion for $\mathbf{x}$.

31. *Introduction:* In this problem, the effects of symmetry on a bifurcation problem are illustrated in one specific example, the three-cell Turing instability, i.e., equation (8.116) in the Pearls with $n = 3$. This problem is symmetric under the dihedral group D_3, which is better known as the symmetric group S_3. (No group theory is required for the problem.) Two ways in which symmetry changes the usual rules of bifurcation are these: (i) At the bifurcation, $\ker \mathbf{DF}_*$ has dimension greater than one. (ii) The bifurcation is "transcritical" in the literal sense of the term, but the usual exchange of stability at a one-dimensional transcritical bifurcation does not apply—the bifurcating solutions are unstable both below and above the bifurcation point. This behavior is generic within the class of problems having the symmetry.

(a) Assuming that $\sigma > 2$ and $\rho > 1$, show that the equal-concentration equilibrium, $(x_k, y_k) = (x_+, y_+)$, $k = 1, 2, 3$, of (8.116) with $n = 3$ is stable if $0 \le D < D_*$ and unstable if $D > D_*$, where

$$D_* = \frac{\rho}{3}\left(\frac{2x_+}{\sigma} - 1\right).$$

Verify that $\ker \mathbf{DF}_*$ is two-dimensional.

Hint: The calculation closely follows Section 6.3.2; $\mathbf{DF}$ has the block form

$$\mathbf{DF} = \begin{bmatrix} A - 2B & B & B \\ B & A - 2B & B \\ B & B & A - 2B \end{bmatrix}, \tag{8.113}$$

where A is the Jacobian of the two-dimensional system and B accounts for diffusion, i.e.,

$$A = \begin{bmatrix} 1 & -1/\sigma \\ 2\rho x_+ & -\rho \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 \\ 0 & D \end{bmatrix}.$$

Apply a similarity transformation to (8.113) with the 6×6 matrix

$$S = \begin{bmatrix} (1/\sqrt{3})I & (1/\sqrt{2})I & (1/\sqrt{6})I \\ (1/\sqrt{3})I & -(1/\sqrt{2})I & (1/\sqrt{6})I \\ (1/\sqrt{3})I & 0 & -(2/\sqrt{6})I \end{bmatrix},$$

where I is the 2×2 identity matrix. (This matrix is orthogonal, so no work is required to calculate its inverse.) You will find that, in block notation,

$$S^{-1}\mathbf{DF}_*S = \begin{bmatrix} A & 0 & 0 \\ 0 & A - 3B & 0 \\ 0 & 0 & A - 3B \end{bmatrix},$$

from which you can deduce the desired conclusions.

(b) By finding explicit solutions of the equations, show that nontrivial equilibria with two equal concentrations bifurcate at D_*, the bifurcating solutions existing both above and below the bifurcation point.

Hint: Assume $x_1 \ne x_2 = x_3$, all three variables being nonzero. (Of course, this is only one of three possibilities that are related by symmetry.) As

in Section 8.6.1(b), process the equilibrium equations to derive a quadratic equation

$$Q(x_1, x_2) = 0 \tag{8.114}$$

analogous to (8.64) and an expression $D = \phi(x_1, x_2)$ analogous to (8.65). First show that (x_+, x_+) lies on the curve (8.114) and $\phi(x_+, x_+) = D_*$; then show that along the curve (8.114), D assumes values greater than D_* on one side of (x_+, x_+) and less than D_* on the other side.

Discussion: In contrast to exchange of stability at a transcritical bifurcation from a simple eigenvalue, these bifurcating equilibria of (8.116) are unstable on both sides of the bifurcation point. It is messy to derive this fact by working directly with the equations. A better proof is based on analysis of the Lyapunov–Schmidt reduction in the presence of symmetry. The relevant theory is contained in Sections XIV.4 and XIV.5 of [32], but substantial background must be mastered to read that material profitably.

32. *Introduction:* Mode jumping in the buckling of a plate is discussed in Section 8.10.2. In this problem, we illustrate the phenomenon by studying a two-dimensional system of ODEs analogous to what the Lyapunov–Schmidt reduction extracts from the PDE problem, i.e.,

$$\begin{aligned} x' &= -x^3 - 3xy^2 + \mu x + \varepsilon, \\ y' &= -x^2 y/2 - y^3 + (\mu - 1)y + \varepsilon, \end{aligned} \tag{8.115}$$

where μ is the bifurcation parameter and ε is a small imperfection parameter.

(a) Assuming $\varepsilon = 0$, show that when $\mu < 0$, the origin is a stable equilibrium of this system, but that as μ crosses zero, it loses stability through a supercritical pitchfork bifurcation (from a simple eigenvalue).

(b) Still assuming $\varepsilon = 0$, articulate what behavior you would expect if, starting from negative values, μ is increased quasistatically.

(c) Check your expectations numerically. That is, starting from the trivial equilibrium with $\mu = -0.5$, increase μ to 2.5 in small increments; after each increment, solve the ODEs until the solution has reconverged to equilibrium. (We suggest taking ε in (8.115) positive but small, say $\varepsilon = 10^{-5}$, because otherwise, the simulation might remain stuck on a solution branch even after it becomes unstable.)

Discussion: What to look for? For $0 < \mu < 2$, your calculated solution will have $x \neq 0$ but $y \approx 0$; as μ crosses 2, it will jump to a new equilibrium with $x \approx 0$ but $y \neq 0$. This discontinuous behavior is a manifestation of competition between the x- and y-modes. Let us elaborate. If $\mu > 1$, the trivial solution is unstable with respect to both modes; both x and y are driven away from zero. (Of course, growth saturates as nonlinear terms come into play.) Near the trivial solution, the x-mode is "more unstable," i.e., the x-mode has the larger eigenvalue in the linearization at the origin, and this explains what is seen initially in the simulation. However, as μ grows and the equilibrium solution grows with it, the nonlinear terms shift the balance of forces and come to favor the y-mode in the competition.

(d) *Introduction:* It is possible to anticipate this behavior from explicit solutions of (8.115). If $\varepsilon = 0$, then both equilibrium equations may be factored, giving rise to four classes of equilibria:

1. $x = y = 0, \quad \mu$ arbitrary : (the trivial solution)
2. $x^2 = \mu,\ y = 0, \quad \mu > 0$: (the x-mode)
3. $x = 0,\ y^2 = \mu - 1, \quad \mu > 1$: (the y-mode)
4. $\begin{cases} x^2 + 3y^2 = \mu \\ x^2/2 + y^2 = \mu - 1 \end{cases}$ (the mixed-mode solution).

These four classes of solutions are represented schematically in the bifurcation diagram in Figure 8.24. Classes 1, 2, and 3 are readily recognizable in the figure, and the next part of the exercise verifies the behavior shown for the Class 4 solutions.

Solve the two coupled linear equations for x^2 and y^2 that characterize Class 4 solutions, and then express x and y as functions of μ. Argue that (real) solutions of Class 4 exist only if $3/2 \le \mu \le 2$. Show that at $\mu = 3/2$, Class 4 solutions meet the y-mode solution branch and that at $\mu = 2$, they meet the x-mode solution branch.

(e) Verify the stabilities listed in Figure 8.24 along the trivial solution and the x- and y-mode solutions. (Stabilities along the mixed-mode solutions follow from exchange of stability.)

Remark: Mode competition holds some other surprises. For example, see Section X.4 of [31] for an overview of the phenomena that can arise from bifurcation problems of the general form (8.115).

8.10 Pearls of Wisdom

8.10.1 Comments on Proving the Hopf Bifurcation Theorem

In two dimensions, it is not very difficult to prove the Hopf bifurcation theorem, *provided* Γ, the cubic coefficient given by (8.94), is nonzero. With a clever change of coordinates, the ODE can be transformed to (8.73), modulo higher-order terms (cf. Exercise 28), and the higher-order terms have minimal effect if $\Gamma \neq 0$. In dimensions higher than two, center manifolds can be used to reduce the problem to two dimensions. These proofs are appealing in that they use only standard ODE methods. However, note that Theorem 8.8.1 makes no hypothesis about nonlinear terms, and without such a hypothesis, it's awkward to find the periodic orbits with these methods.

Our theorem on Hopf bifurcation combines different versions from several sources, e.g., [14, 16, 31, 33, 38, 52]. For example, some authors:

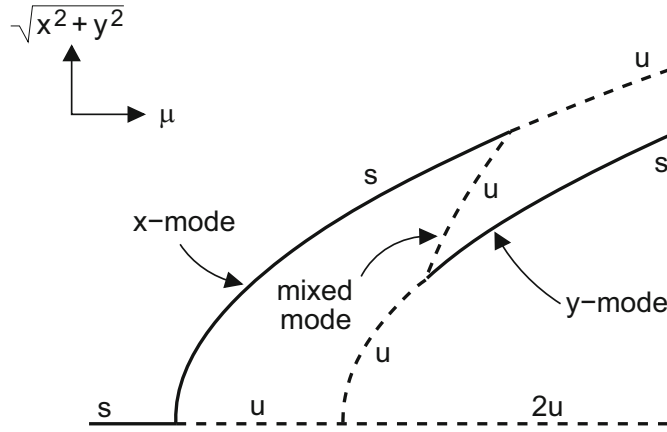

Figure 8.24: *Schematic bifurcation diagram for* (8.115)*, the norm of the solution* $\sqrt{x^2+y^2}$ *vs.* μ*. Each point on the x-mode solution branch represents two equilibria* $(\pm x, 0)$ *of* (8.115)*; on the y-mode branch, two equilibria* $(0, \pm y)$*; and on the mixed–mode branch, four equilibria* $(\pm x, \pm y)$*. Despite appearances, the mixed-mode branch has a* pitchfork *bifurcation at both ends of its interval of existence.*

- Assume throughout that at the bifurcation point, $\mathbf{DF}_*$ has no eigenvalues in the right half-plane. This simplifies the analysis, and of course it is the most interesting case.

- Allow additional eigenvalues on the imaginary axis besides the basic pair $\lambda_{1,2} = \pm i\omega_*$, with the restriction that $in\omega_*$ is not an eigenvalue for any integer n. Under this more general hypothesis, the uniqueness result is weaker; it guarantees only that every periodic solution whose *period is close to* $2\pi/\omega_*$ will lie on the surface (8.90).

- Formulate a result that also applies to PDEs. This raises a number of technical issues.

To peruse all the references above to verify every part of Theorem 8.8.1 is a truly thankless task. If you want to see a proof of Hopf bifurcation, be prepared to settle for a lesser result.

For the record, we are partial to the method of Cesari and Hale, which is used in the proof of Section VII.1 of [31]. This method, based on a Lyapunov–Schmidt reduction in an infinite-dimensional Banach space, does not need restrictive hypotheses on the higher-order terms. On the other hand, the method is extravagant with derivatives and does not easily yield the strongest uniqueness statement.

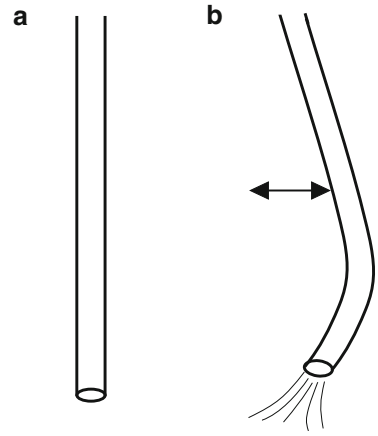

Figure 8.25: *Fluid flowing through a flexible hose with a loose end can induce oscillatory motion. (a) A stationary hose in the absence of fluid flow. (b) Flutter of the loose end of the hose due to fluid flow.*

8.10.2 High-Dimensional Bifurcation: Symmetry and Mode Competition

For either steady-state or Hopf bifurcation, symmetry often leads to bifurcation from a high-dimensional eigenvalue. In such problems, not only are there many different bifurcating solutions, but there may be many different *classes* of bifurcating solutions.[37] For example, consider fluid flowing through a thin flexible hose and emerging from an unsupported end (cf. Figure 8.25). As you probably have experienced, rapid flow causes the hose to flutter. Let's suppose, as in [5], that the flow is vertical, so that the problem is symmetric with respect to rotation about the vertical axis. The governing equations of this system have a steady-state solution in which the hose is at rest; if the flow speed is slow, this equilibrium is asymptotically stable, but it loses stability as the flow speed increases. Periodic solutions, in which the hose oscillates in a vertical plane, bifurcate from the trivial solution at the critical flow velocity; such motion is possible in every vertical plane. At the same critical speed, other solutions, in which the loose end moves in a circle, also bifurcate. Although both bifurcations are supercritical, only one of these solution classes is stable; which class is stable depends on other parameters in the problem. And rather surprisingly, there may be solutions near the bifurcation point in which the loose end traces out an elliptical path whose axes slowly precess. Although Bajaj and Sethna [5] derive this behavior from analyzing the equations of motion, much of it can be anticipated from symmetry considerations alone; cf. Figure 7.2 in Chapter XVII of [32].

[37] A time-honored example is Rayleigh–Bénard convection, the physical problem underlying the Lorenz equations in Section 8.1.2. Case Study 4 of [32] discusses information that can be derived using just the symmetries of the problem.

Figure 8.26: *Schematic of buckles in a long rectangular plate under compression. The sign of the out-of-plane deformation alternates between shaded and unshaded regions, say upward in the shaded regions and downward in the unshaded regions.*

Although ODE examples of bifurcation with symmetry are far less striking than the PDE examples, let's consider one anyway: in the n-cell Turing instability, symmetry forces a bifurcation with a multidimensional kernel. Assuming that the cells are arranged cyclically in a ring,[38] we generalize (8.60) to obtain a system of $2n$ ODEs,

$$\begin{array}{lrcl} (a) & x_k' & = & \sigma x_k^2/(1+y_k) - x_k, \\ (b) & y_k' & = & \rho\,[x_k^2 - y_k] + D(y_{k+1} - 2y_k + y_{k-1}), \end{array} \tag{8.116}$$

where $k = 1, 2, \ldots, n$. (Here, y_0 is identified with y_n, and y_{n+1} is identified with y_1.) This system is symmetric under the dihedral group D_n, the rotations and reflections that leave a regular n-sided polygon invariant. (Cf. [3], Chapter 4.) The effects of symmetry are felt already when $n = 3$, a case that we ask you to explore in Exercise 31. (No group theory is required for the exercise.) More generally, the consequences of D_n symmetry on bifurcation are studied in Section XIII.5 of [32].

Bifurcations from a multiple eigenvalue also arise in another context. A "trivial" equilibrium of a system may become unstable, as a driving parameter μ is increased, to two or more different perturbations, or modes, at close to the same value of μ. Competition between two modes may dictate key features of the response of the system. The buckling of a plate under compression provides a nice illustration of possible consequences of such competition.[39]

Some background: When a flat rectangular plate (that is supported on all four sides) is subjected to a sufficiently large longitudinal compression, it buckles in a wavelike pattern, as suggested in Figure 8.26. The buckles form most readily if they are nearly square. Thus, if the aspect ratio (length to width) of the plate is nearly an integer, say k, then k buckles form at the bifurcation point, and they are very nearly square. However, if this ratio lies between two integers, say k and $k+1$, then modes with k buckles and $k+1$ buckles compete.

[38] This geometry is more interesting than allowing diffusion between every pair of cells, because it approximates the geometry of the actual Turing instability. (Cf. [58], Section 2.2 of Volume II.)

[39] Mode competition in a fluids problem, with more complicated phenomena than for plates, is analyzed in Case Study 6 of [32].

This competition may cause a strong discontinuity in how a plate responds under compression. In the classic experiment of [77], in which the aspect ratio equaled 5.38, at the bifurcation point a pattern with five buckles appeared. As the load was increased, the amplitude of the buckles increased continuously while maintaining the same pattern, up to approximately 1.7 times the original buckling load. When the load was further increased, the plate jumped suddenly and violently to a pattern with six buckles.[40] Since the experiment involved a metal plate squeezed in an apparatus capable of generating 1,200,000 pounds of force,[41] this jump was a dramatic event; you would not want to be inside a spacecraft in which one of the panels behaved in this way.

In Case Study 3 in [31], mode-competition ideas are used to analyze such mode jumping in a mathematical model for the buckling of plates, without recourse to the computer. The Lyapunov–Schmidt technique is invoked to reduce the PDEs to a two-dimensional system, and in Exercise 32 we invite you to study a two-dimensional ODE of the same form that exhibits the exact same jump behavior.[42]

8.10.3 Homeostasis, or "Antibifurcation"

The focus of bifurcation theory is to describe how the behavior of solutions of an ODE changes as a parameter varies. As the tongue-in-cheek name "antibifurcation" is intended to suggest, the central phenomenon of homeostasis is exactly the opposite: at least one component of a (steady-state) solution changes very little as a parameter varies. This behavior is essential for life: conditions affecting critical internal processes (e.g., body temperature in mammals) need to remain close to a desired operating point, despite variations in the external environment.

The phenomenon of homeostasis is illustrated by the bathtub model[43] that is indicated in the schematic diagram, Figure 8.27(a). This model is described by the (nondimensionalized) ODEs

[40] And subsequently to seven buckles, to eight buckles, and to complete collapse, but these jumps are beyond our focus.

[41] In Stein's experiment, compression was generated mechanically, but compression generated by thermal expansion, such as on a spacecraft during reentry, is also of interest.

[42] Incidentally, Figure 8.24, a bifurcation diagram for equation (8.115) in the exercise, applies equally well to the buckling plate if the x- and y-modes in the figure are identified with the five-buckle and six-buckle modes of the plate, respectively.

[43] This network is an example of what's called a *feed-forward* loop (cf. Chapter 4 of [2]); i.e., the concentration of X is "fed forward" to influence a reaction rate of one of its products. Such feed-forward loops are embedded in many complex biological networks; the metabolic networks of [60] provide one example.

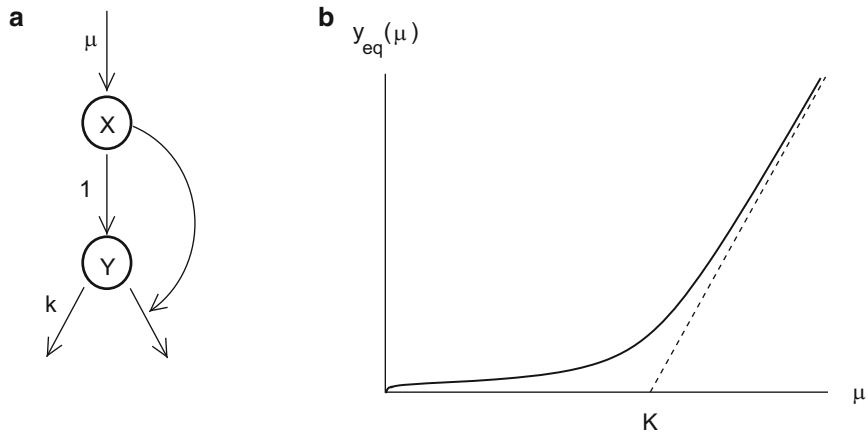

Figure 8.27: *(a) Schematic diagram for the homeostasis model equations* (8.117). *(b) Equilibrium value of y as a function of μ for $k = 0.25$ and $K = 15$.*

$$\begin{aligned} (a) \quad x' &= \mu - x, \\ (b) \quad y' &= x - ky - \frac{xy}{1 + xy/K}, \end{aligned} \tag{8.117}$$

where μ, k, and K are positive constants. Thus, X is supplied from the environment at a rate μ, and X reacts to form Y. The goal of homeostasis is to maintain the concentration y at a constant level, regardless of how the input parameter μ may change. To achieve this, Y is depleted by two processes, one linear—this could represent either consumption or a baseline decay—and one nonlinear, the mechanism behind homeostasis. Specifically, besides producing Y, X acts as an enzyme that actively promotes the decay of Y. Thus, if μ gets large, making X large, an excess of Y is avoided because of enhanced decay. However, this enzymatic promotion saturates at large concentrations, which is modeled by the term xy/K in the denominator in (8.117b).

For all $\mu > 0$, (8.117) has a unique equilibrium. In Figure 8.27(b), the equilibrium value of y is shown as a function of μ for one choice of reaction parameters. Note how little y changes over a wide range of the input parameter μ. Some authors call such a graph a "chair."

The steady-state solution $y_{\text{eq}}(\mu)$ of (8.117) can be found by the quadratic formula, but the result is unilluminating, to say the least. Far more insight is gained by considering two limiting cases: (i) If K is large and μ is not too large, we may neglect the term xy/K in (8.117a) compared to unity, which leads to the approximation

$$y_{\text{eq}}(\mu) \approx \frac{\mu}{k + \mu}.$$

(ii) If μ is very large, making x and y large, then xy/K dominates the denominator in (8.117a); neglecting 1 compared to this term gives

$$y_{\text{eq}}(\mu) \approx \frac{\mu - K}{k}.$$

Both approximations can be seen in Figure 8.27(b).

Homeostasis also insulates against the effects of a time-varying environment. We invite you to experiment with this by solving (8.117) with the parameter μ replaced by a function of time, such as a square wave.

Chapter 9

Examples of Global Bifurcation

The main goal of this chapter is to give examples of five different types of global bifurcation. While the local bifurcations of the previous chapter were associated with stability changes in an equilibrium, the bifurcations in this chapter are associated with stability changes in a periodic solution. We make no pretense of completeness. It is not remotely possible to classify all possible global bifurcations. We introduce each type of bifurcation with an academic example that may be handled analytically, but for the more interesting examples that follow, we will rely heavily on computations (which we invite you to check) and on intuitive arguments. Even in cases for which proofs are available, we may omit them because we find them unrewarding. Consequently, this chapter is much less dense than the preceding one.

Sections 9.1, 9.2, and 9.4–9.6 present the examples. Section 9.3 discusses a theoretical issue that unifies the following three sections. Two additional sections describe scientifically interesting problems in which global bifurcation plays a central role: the onset of chaos in the Lorenz system (Section 9.7) and bursting in neurons (Section 9.8).

9.1 Homoclinic Bifurcation

9.1.1 An Academic Example

We base our construction on modified-Hamiltonian systems, as we considered in Exercise 7.22,

$$\begin{aligned} x' &= \partial_y H - \partial_x (H-\mu)^2, \\ y' &= -\partial_x H - \partial_y (H-\mu)^2, \end{aligned} \tag{9.1}$$

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8_9

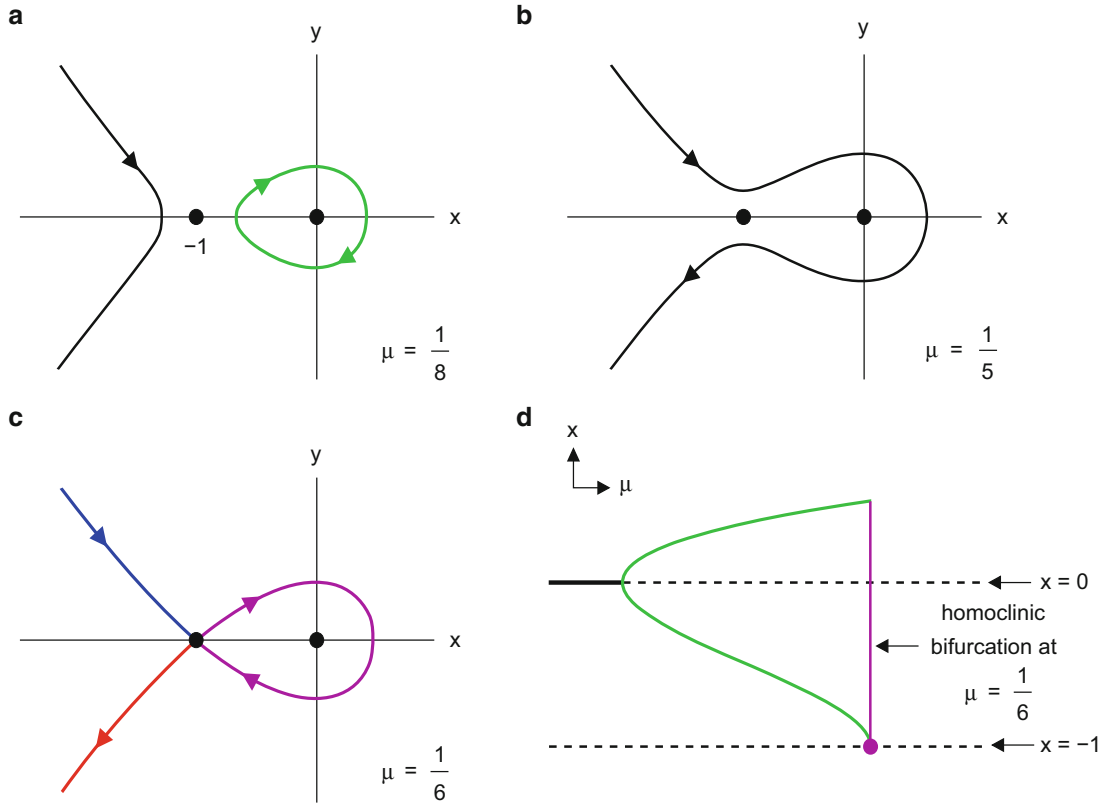

Figure 9.1: *Panels (a–c) show various orbits of* (9.1)*, with the Hamiltonian* (9.2)*, for different values of* μ*. Panels (a) and (b) are representative of orbits in the ranges* $0 < \mu < 1/6$ *and* $\mu > 1/6$*, respectively, while* $\mu = 1/6$ *exactly in Panel (c). The orbits, which are shown with the usual color conventions, are obtained from level sets* $\{H = \mu\}$ *of the Hamiltonian. (d) Bifurcation diagram indicating the range of* μ *over which* (9.1) *has a periodic orbit and the termination of the periodic orbits in a homoclinic bifurcation.*

where $H(x, y)$ is some smooth function. A simple chain-rule calculation shows that along a trajectory,

$$\frac{dH}{dt} = -2(H - \mu)\,|\nabla H|^2.$$

Thus, trajectories are driven toward the minimum of $(H - \mu)^2$, the specific level set $\{H(x, y) = \mu\}$, which is a union of orbits (or possibly a single orbit). Consider the one-parameter family of ODEs obtained by substituting into (9.1) the Hamiltonian

$$H(x, y) = y^2/2 + x^2/2 + x^3/3, \tag{9.2}$$

a quadratic kinetic energy plus a cubic potential energy with a local minimum at $x = 0$. As illustrated in Figure 9.1, the orbit structure of this problem changes as μ

passes through 1/6: if $0 < \mu < 1/6$, the level set $\{H = \mu\}$ contains a loop around the origin, which is a periodic orbit, while for $\mu > 1/6$, the ODE (9.1) has no periodic solutions. The transition at $\mu = 1/6$ includes a homoclinic orbit.

9.1.2 The van der Pol Equation with a Nonlinear Restoring Force

A virtually identical bifurcation occurs in a perturbation of the van der Pol equation by a nonlinear force εx^2,

$$\begin{aligned} x' &= y, \\ y' &= -\beta(x^2-1)y - x - \varepsilon x^2, \end{aligned} \tag{9.3}$$

a problem you analyzed in Exercise 7.9. If ε is small, the periodic orbit of the van der Pol equation continues to exist, being only slightly distorted. However, if ε is sufficiently large ($\varepsilon \geq 1$ is sufficient), the periodic orbit is destroyed by the perturbation. This homoclinic bifurcation is illustrated in Figure 9.2, which shows some computed trajectories and the stable and unstable manifolds of the saddle point $(-1/\varepsilon, 0)$. In Figure 9.2(a), for ε small, the right half of the unstable manifold $\mathcal{M}_u$ (in red) quickly approaches the periodic orbit (in green) asymptotically, while both halves of $\mathcal{M}_s$ (in blue) come in from infinity. On the other hand, for ε large (Figure 9.2(b)), half of the *stable* manifold $\mathcal{M}_s$ connects the *origin* to the saddle point, while both halves of $\mathcal{M}_u$ march off to infinity. As you would expect from continuity, there is a value $\varepsilon_{\text{homo}}$ such that $\mathcal{M}_u$ intersects $\mathcal{M}_s$, which means that the two manifolds coincide, as shown (in purple) in Figure 9.2(c). As is usual for homoclinic bifurcation, it is not possible to locate the bifurcation point analytically; numerically, if $\beta = 1/2$, we find that $\varepsilon_{\text{homo}} \approx 0.399$.

Regarding this intuitive description of the bifurcation, let us acknowledge that we have not shown that the stable and unstable manifolds through a saddle point depend continuously on parameters. Moreover, the continuity argument shows only that there is *at least* one value of ε for which $\mathcal{M}_u$ and $\mathcal{M}_s$ intersect; although it can be proved that there is only one such value, we prefer to rely on computation to conclude this.

9.1.3 Section 1.6 Revisited: Part VI

A variation of homoclinic bifurcation occurs in the augmented Lotka–Volterra equations

$$\begin{aligned} (a)\quad x' &= x\left(\frac{x-\varepsilon}{x+\varepsilon}\right)(1-\frac{x}{K}) - xy, \\ (b)\quad y' &= \rho(xy - y). \end{aligned} \tag{9.4}$$

In Section 8.7.3 we saw that a subcritical Hopf bifurcation creates a family of periodic orbits of (9.4) at $K = K_{\text{hopf}} = (1+2\varepsilon-\varepsilon^2)/2\varepsilon$. Thus, as illustrated in Figure 9.3(a), for K slightly less than K_{hopf}, equation (9.4) has an unstable periodic orbit. As

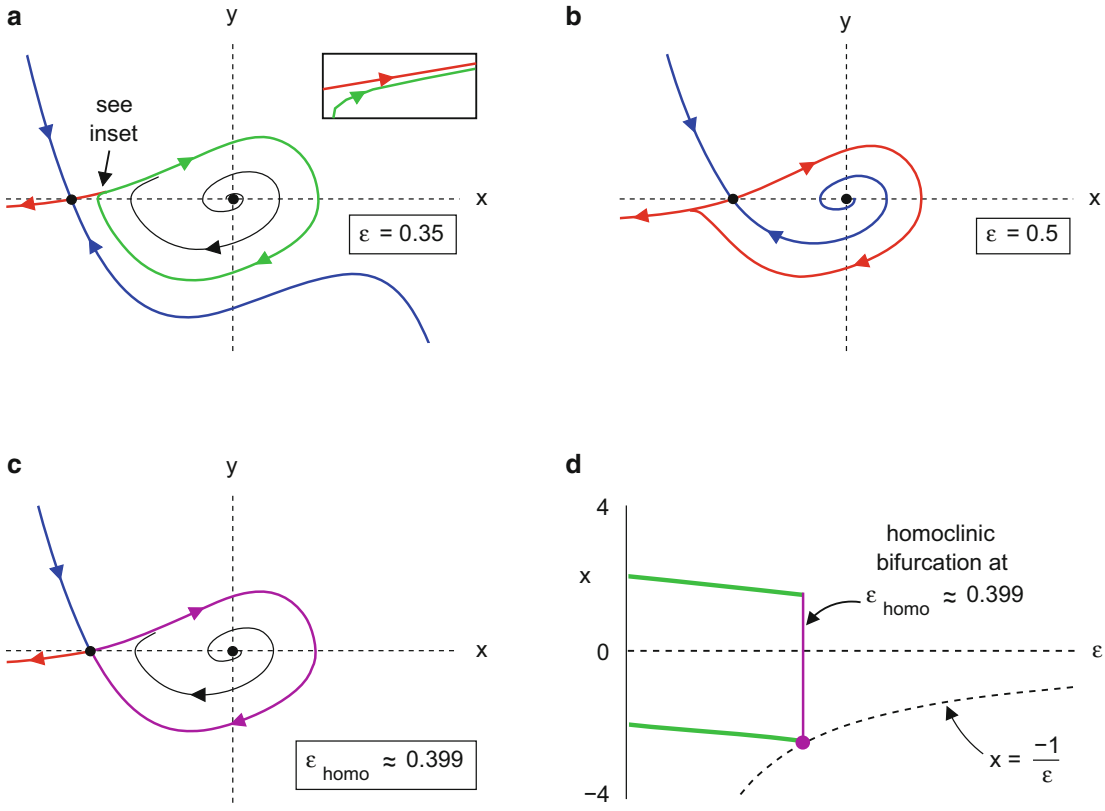

Figure 9.2: *Panels (a–c) show phase-plane plots, including stable and unstable manifolds, for the perturbed van der Pol system (9.3) for various values of ε, assuming $\beta = 1/2$. If $\varepsilon < \varepsilon_{\text{homo}}$, (Panel (a)) there is a stable periodic orbit, but for $\varepsilon > \varepsilon_{\text{homo}}$ (Panel (b)) this orbit has disappeared through the homoclinic bifurcation shown in Panel (c). Note that the asymptotic behaviors of $\mathcal{M}_s$ and $\mathcal{M}_u$ are changed by the bifurcation. Panel (d) shows a bifurcation diagram for the system.*

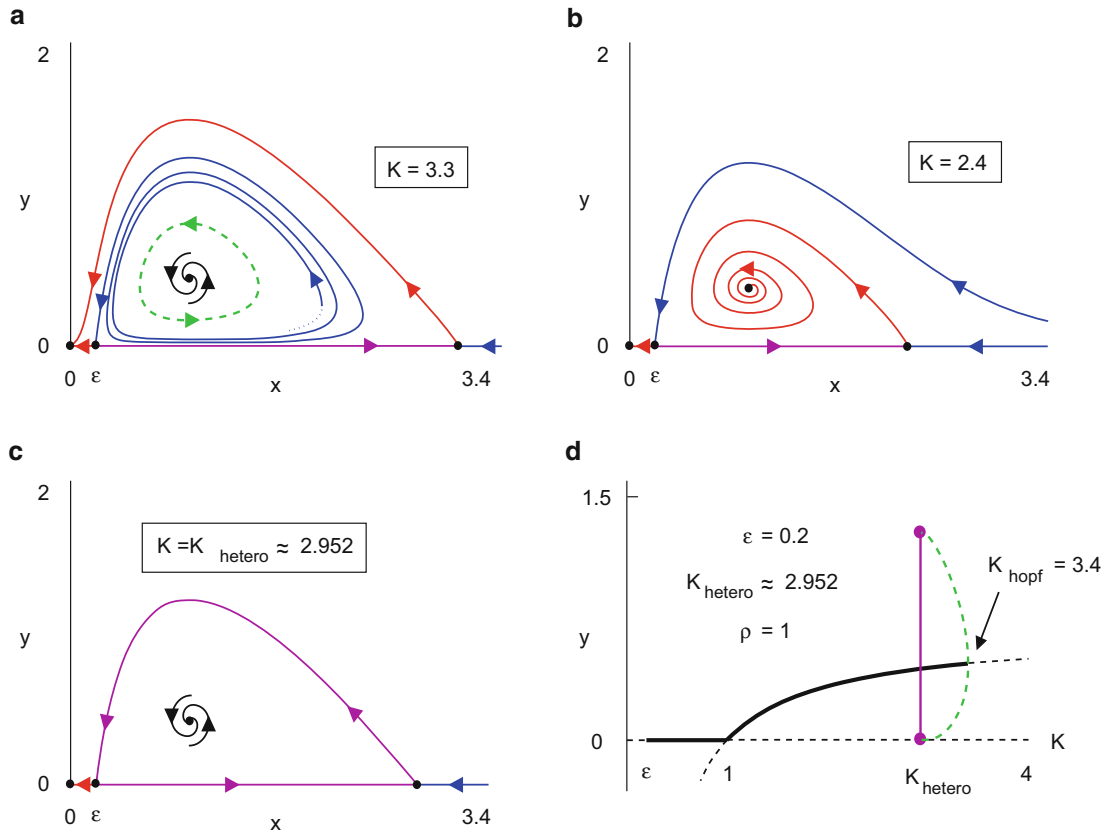

Figure 9.3: *The heteroclinic bifurcation in the augmented Lotka–Volterra equations* (9.4) *as K is varied, assuming $\varepsilon = 0.2$ and $\rho = 1$. Panels (a) and (b) show phase portraits on either side of the bifurcation, while Panel (c) shows the phase portrait at the bifurcation point. Panel (d) gives a bifurcation diagram for the system.*

K decreases from K_{hopf}, the orbit expands until at $K = K_{\text{hetero}}$, it coincides with a *heteroclinic cycle* that links the two saddle points $(\varepsilon, 0)$ and $(K, 0)$, the phase portrait shown in Figure 9.3(c). When $K < \mathcal{K}_{\text{hetero}}$, the equation no longer has a periodic orbit, as in Figure 9.3(b).

The overall behavior is summarized by the bifurcation diagram in Figure 9.3(d). The heteroclinic bifurcation point K_{hetero}, the lower limit of the range in which the periodic solution exists, depends on ε and ρ; if $\varepsilon = 0.2$ and $\rho = 1$ as in the figures, we compute that $K_{\text{hetero}} \approx 2.952$.

As with homoclinic bifurcation, the present bifurcation is accompanied by changes in the α- and ω-limits of the stable and unstable manifolds through the saddle points. Specifically, you should identify the ω-limit of the unstable manifold through $(K, 0)$ in each case, $1 < K < K_{\text{hetero}}$, $K = K_{\text{hetero}}$, and $K_{\text{hetero}} < K < K_{\text{hopf}}$, and the α-limit of the stable manifold through $(\varepsilon, 0)$ in the three cases. By contrast, the α-limits of the stable manifold through $(K, 0)$ and the ω-limits of the unstable manifold through $(\varepsilon, 0)$ are the same for all $K > 1$, as you should also check.

This bifurcation provides the final update to Section 1.6.2. The parameters used in both Panels (a) and (b) of Figure 9.3 belong to Region II of Figure 1.9. The phase portrait in Figure 9.3(b) is topologically equivalent to the plot in Figure 1.9 for Region II, shown on the upper left; but the phase portrait in Figure 9.3(a) does not appear in Figure 1.9. The dedicated reader could include this extra information in Figure 1.9 by dividing Region II into two subregions with different phase portraits as in Figure 9.3(a,b). The subregions are separated by the curve, K vs. ε, that locates the heteroclinic bifurcation. The position of this curve depends on the parameter ρ in (9.4), and the computer is needed to locate it.

9.1.4 Other Examples of Homoclinic and Heteroclinic Bifurcations

In Exercise 4 you will see that a homoclinic bifurcation occurs in the torqued pendulum equation, *provided friction is sufficiently small.* (If friction is large, a different bifurcation occurs, which we study in Section 9.2.3.)

If you reread Section 8.8.2 with the ideas of the present section in mind, you will discover that homoclinic bifurcation occurs in the activator–inhibitor model (8.21) when ρ passes through the value $\rho_{\text{homo}} = 1$. In this nongeneric example, all the bifurcation phenomena are crammed into the plane $\{\rho = 1\}$, i.e., the Hopf bifurcation that "creates" the periodic solutions, the one-parameter family of periodic solutions themselves, and the homoclinic bifurcation that "destroys" them. (Incidentally, the perturbation of (8.21) considered in Exercise 8.25 removes this nongeneric behavior.)

Homoclinic and heteroclinic bifurcations play a central role in the applications studied in Sections 9.7 and 9.8. The heteroclinic bifurcation in the former section is interesting in that an unstable manifold connects a saddle point to an unstable *periodic orbit*, rather than to another saddle point as in Section 9.1.3.

9.2 Saddle-Node Bifurcation of Limit Cycles

9.2.1 An Academic Example

The bifurcation is illustrated by the one-parameter family of ODE on $\mathbb{R}^2 \sim \{0\}$ given in polar coordinates by

$$\begin{aligned} r' &= 1 - r, \\ \theta' &= r^{-1} - \cos\theta + \mu. \end{aligned} \tag{9.5}$$

As μ crosses $\mu_{\text{snlc}} = 0$, two related phenomena occur: (i) (μ increasing) The equilibria

$$r = 1, \quad \theta = \pm \arccos(1 + \mu), \tag{9.6}$$

which exist for $-2 < \mu < 0$, disappear through a saddle-node bifurcation at $\mu = 0$. (ii) (μ decreasing) If $\mu > 0$, the unit circle is a periodic orbit of (9.5), but this periodic orbit disappears as μ becomes negative, because two equilibria appear and block the flow. Phase-plane plots on either side of the bifurcation point $\mu = 0$ are shown in Figure 9.4(a,b), while the phase-plane plot at the bifurcation point is shown in Figure 9.4(c). Overall behavior is summarized in the bifurcation diagram Figure 9.4(d). Somewhat unimaginatively, we say that (9.5) undergoes a *saddle-node bifurcation of limit cycles*, or more compactly an *SN-limit-cycle* bifurcation.

The use of purple in Figure 9.4(c) extends our color conventions. In Figures 9.1(c) and 9.2(c), the purple orbit is both the (global) stable and unstable manifold of a hyperbolic equilibrium, which coincide. By contrast, in Figure 9.4(c) the purple orbit begins and ends at the *nonhyperbolic* equilibrium $r_* = 1$, $\theta_* = 0$, where the Jacobian has eigenvalues -1 and 0; this orbit is a *center* manifold.[1]

9.2.2 The Denatured Morris–Lecar Equation

We can now answer a question raised at the end of Section 8.7.4 concerning the denatured Morris–Lecar equations (8.79),

$$\begin{aligned} x' &= x^2(1 - x) - y + I, \\ y' &= Ae^{\alpha x} - \gamma y, \end{aligned} \tag{9.7}$$

which, as before, we consider with the parameter values $A = 0.001$, $\alpha = 6$, and $\gamma = 0.15$. As indicated in the bifurcation diagram for (9.7), which we repeat in

[1]It may be instructive to compare Figures 9.3(c) and 9.4(a). In both figures, two different colored trajectories connect two equilibria. In Figure 9.3(c), both equilibria are saddle points; the trajectories are shown in purple because they are simultaneously (part of) the stable manifold of one equilibrium and the unstable manifold of the other. By contrast, in Figure 9.4(a), one of the equilibria is a sink; the trajectories are shown in red because they are the unstable manifold of the saddle point $r = 1$, $\theta = \arccos(1 + \mu)$, but they are in no way distinguished as regards the sink; they are only two of many orbits that converge to that equilibrium.

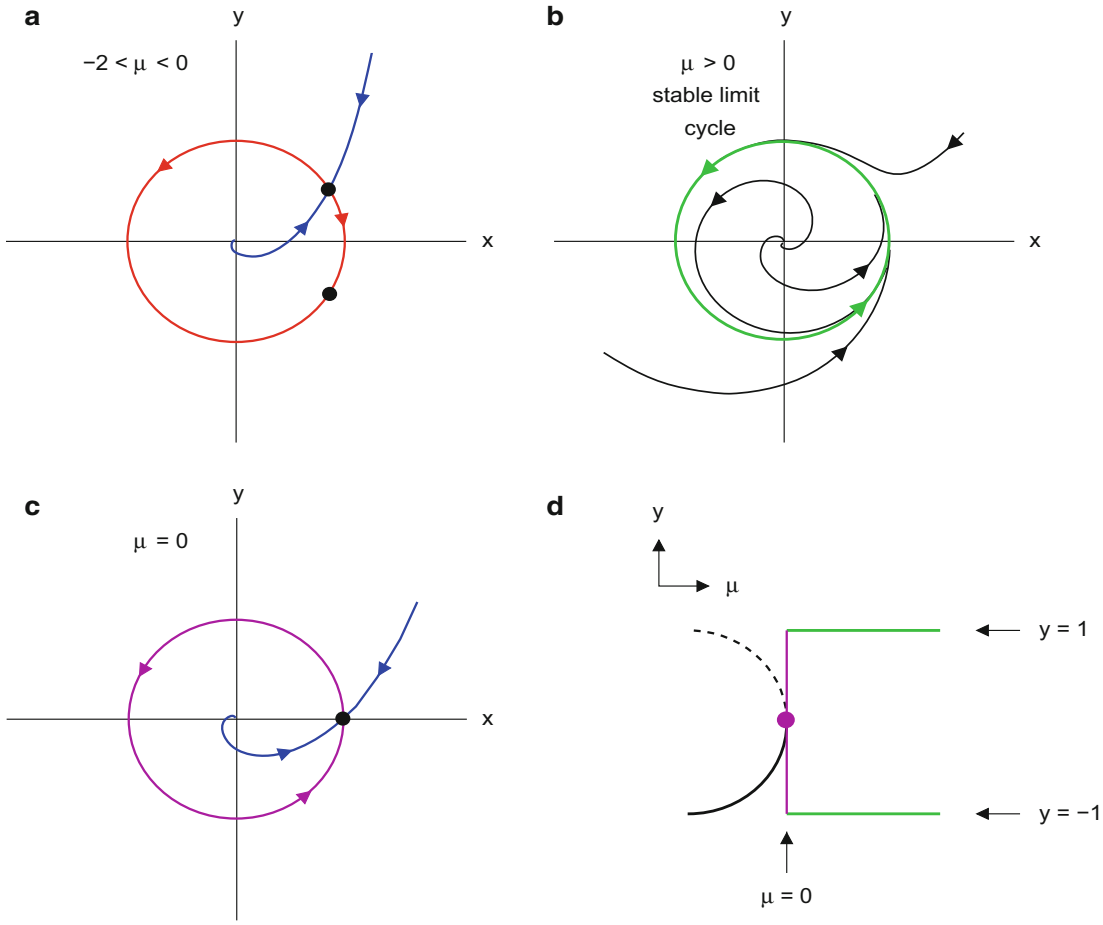

Figure 9.4: *Representative phase-plane plots and bifurcation diagram for the saddle-node bifurcation of limit cycles in* (9.5). *In Panel (c), the orbit in purple represents a center manifold, not stable and unstable manifolds that coincide.*

Figure 9.5(d), the equations have a periodic solution for a range of currents. (The periodic solution for $I = 0.02$ is shown in Figure 9.5(b), superimposed on the nullclines of (9.7).) These periodic solutions disappear if I is decreased below the threshold labeled I_{snlc} in the bifurcation diagram. *What happens to them?* Well, observe in Figure 9.5(a) that below the threshold, a pair of new equilibria appear and block the would-be periodic solution, exactly the same behavior that occurs in Figure 9.4. That is, the periodic solutions disappear through an SN-limit-cycle bifurcation.

Given this information, we can locate the bifurcation point very accurately, more accurately than trying to read it off the bifurcation diagram, by solving the equations

$$\begin{aligned} x^2(1-x) - y + I &= 0, \\ Ae^{\alpha x} - \gamma y &= 0, \\ \alpha A e^{\alpha x} - \gamma(2x - 3x^2) &= 0. \end{aligned}$$

The first two equations are the equilibrium equations for (9.7), and the third is the condition that the Jacobian is singular at a saddle-node bifurcation point. Solving these equations (numerically) for the appropriate values of A, α, γ, we obtain $I_{\text{snlc}} \approx 0.007137$.

Incidentally, two other bifurcation points are labeled in Figure 9.5(d): I_{hopf} locates the Hopf bifurcation that was discussed in Section 8.7.4, and I_{mutan} locates what is called a "mutual annihilation" bifurcation, which will be discussed in Section 9.4.2.

9.2.3 The Overdamped Torqued Pendulum

Another instance of an SN-limit-cycle bifurcation occurs in the torqued-pendulum equation

$$\begin{aligned} x' &= y, \\ y' &= -\sin x - \beta y + \mu, \end{aligned} \tag{9.8}$$

provided the friction coefficient β is large enough. We saw in Section 8.4.1 that the *equilibria* of this system undergo a saddle-node bifurcation at the critical torque $\mu = 1$.

Representative phase portraits and a bifurcation diagram for this system, assuming $\beta = 1.3$, are shown in Figure 9.6. It is instructive to compare Figure 9.6(c), which is global, with Figure 6.14, which shows the flow only locally near the degenerate equilibrium $(\pi/2, 0)$. (Note the difference in scales between the two figures, especially in the y-scales.) As indicated in Figure 6.14, to the right of the equilibrium the center manifold $\mathcal{M}_c^{(r)}$ is unique and flows away from the equilibrium, while there are many candidates for $\mathcal{M}_c^{(l)}$ to the left, all of which converge to the equilibrium. Figure 9.6(c) shows yet another candidate for $\mathcal{M}_c^{(l)}$, one not represented in Figure 6.14, i.e., the extension of $\mathcal{M}_c^{(r)}$ to a global center manifold, which converges back to the equilibrium. This global center manifold describes a trajectory in which

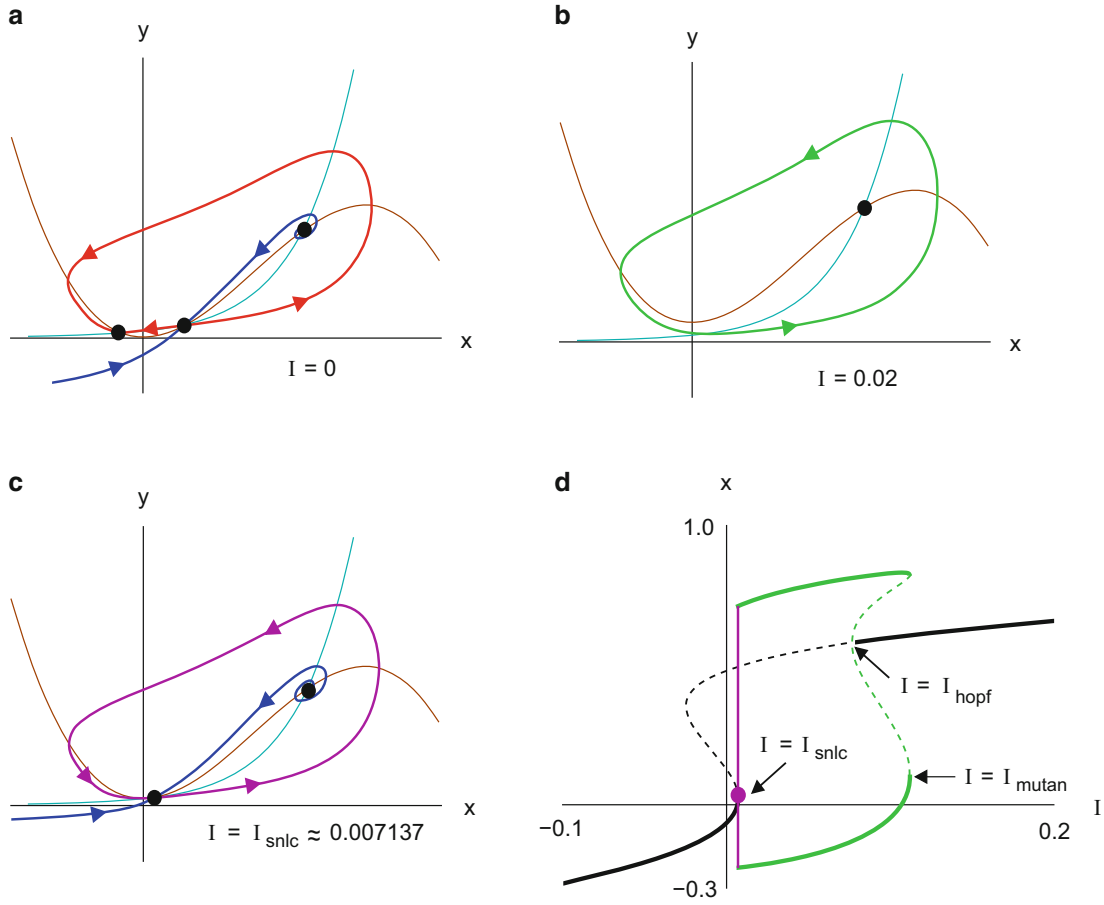

Figure 9.5: *Phase-plane plots, with nullclines superimposed, and bifurcation diagram for the saddle-node bifurcation of limit cycles in the denatured Morris–Lecar equations* (9.7)*, assuming* $A = 0.001$*,* $\alpha = 6$*, and* $\gamma = 0.15$*. In Panel (c), purple represents a center manifold.*

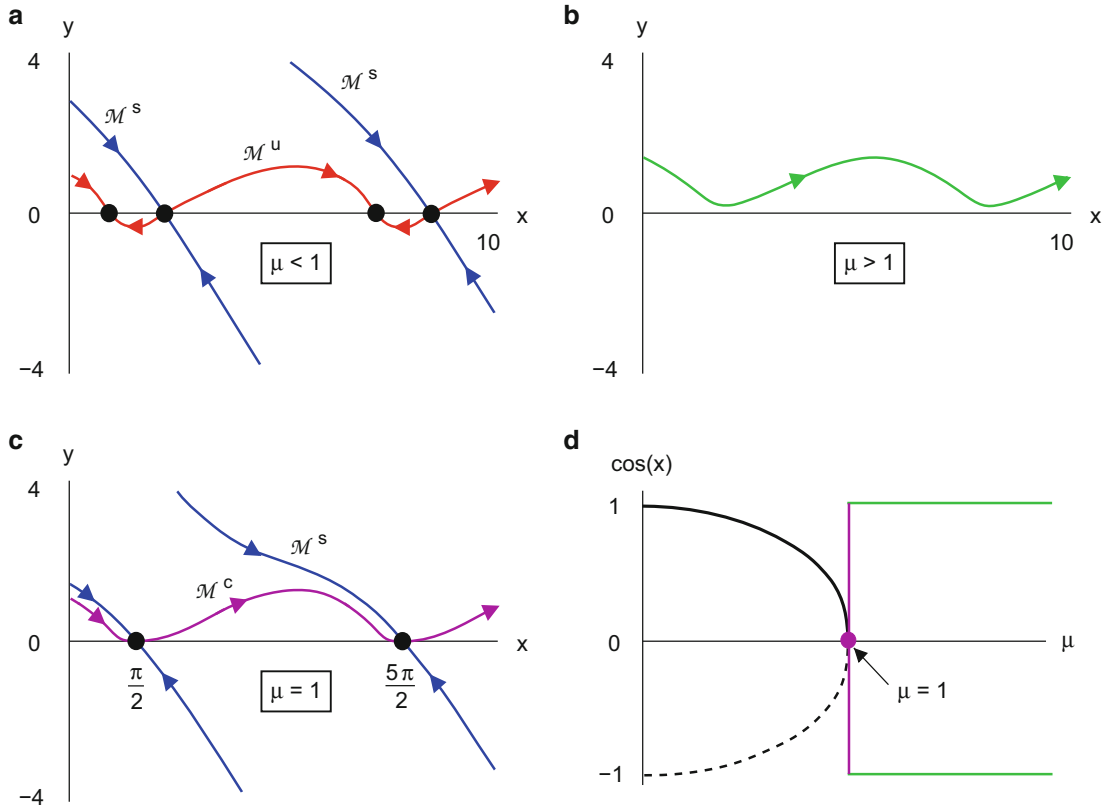

Figure 9.6: *Phase-plane plots and bifurcation diagram for the saddle-node bifurcation of limit cycles in the torqued pendulum equations* (9.8) *as* μ *is varied, assuming* $\beta = 1.3$. *In Panel (c), purple represents a center manifold.*

the pendulum starts from rest at $t = -\infty$, undergoes a complete revolution, and then converges back to the equilibrium as $t \to \infty$. As we saw in Figure 9.4(c), a homoclinic global center manifold is the signature of an SN-limit-cycle bifurcation.

The phase portraits in Figures 9.6(a-b) conform to the SN-limit-cycle-bifurcation pattern: if μ is decreased slightly below 1, the degenerate equilibrium splits into two equilibria, a saddle and a node, and if μ is increased slightly, then the equilibrium disappears and a periodic solution appears.

Note that in the bifurcation diagram Panel (d), $\cos x$ is plotted vs. μ rather than x vs. μ. This choice is convenient, because the periodic solution is periodic only because x is reduced modulo 2π.

Figure 9.6 is drawn for the specific friction coefficient $\beta = 1.3$. For all $\beta > 1.191$, a qualitatively identical bifurcation occurs. However, for $\beta < 1.191$, different behavior—called *underdamped*—occurs. This is explored in Exercise 4.

9.3 Poincaré Maps and Stability Loss of Limit Cycles

In the bifurcations of the preceding two sections, a periodic solution exists for a range of a parameter and suddenly disappears at the edge of this range. In the three remaining types of global bifurcation, a periodic solution loses stability but continues to exist. Recall from Section 7.3 that the stability of a periodic solution $\boldsymbol{\gamma}(t)$ was studied through the Poincaré map, which is constructed so that $\boldsymbol{\gamma}(0)$ is a fixed point of $\boldsymbol{\Pi}$. Specifically, a periodic solution of a d-dimensional ODE is asymptotically stable if the eigenvalues of (the differential of) the Poincaré map at the fixed point $\boldsymbol{\gamma}(0)$ satisfy

$$|\lambda_j(\mathbf{D}\boldsymbol{\Pi})| < 1, \qquad j = 1, \ldots, d-1. \tag{9.9}$$

In the context of bifurcation theory, consider a one-parameter family of ODEs $\mathbf{x}' = \mathbf{F}(\mathbf{x}, \mu)$. Suppose that for $\mu = \mu_*$, this equation has a nonhyperbolic periodic solution $\boldsymbol{\gamma}_*(t)$; i.e., one or more eigenvalues of $\mathbf{D}\boldsymbol{\Pi}_*$ lie on the unit circle. We want to consider a minimally degenerate situation, which may occur in one of three generic ways:[2]

- $\mathbf{D}\boldsymbol{\Pi}_*$ has a simple eigenvalue $\lambda = 1$,
- $\mathbf{D}\boldsymbol{\Pi}_*$ has a complex-conjugate pair of eigenvalues of modulus 1, or
- $\mathbf{D}\boldsymbol{\Pi}_*$ has a simple eigenvalue $\lambda = -1$.

As with bifurcation from equilibria, new solutions of the ODE appear in this context. Each case has its own characteristic phenomena, which we illustrate with examples in Sections 9.4–9.6.

Remarks: (i) The linearized map $\mathbf{D}\boldsymbol{\Pi}_*$ is obtained from solving the IVP for the linear ODE

$$\mathbf{x}'' + A(t)\mathbf{x} = \mathbf{0}, \tag{9.10}$$

where $A(t)$ is the periodic matrix-valued function $A(t) = \mathbf{DF}(\boldsymbol{\gamma}_*(t), \mu_*)$. Such problems are precisely the focus of Floquet theory. *(ii)* Bifurcation of fixed points of maps defined on Euclidean space can be studied divorced from any connection to a Poincaré map. Fixed points of a map lose stability through the same three mechanisms above. Although we probe a few bifurcation phenomena for maps in Exercises 13 and 14 in this chapter and in Section 10.6.1, we refer you to Chapter 10 of [81] or to [17] for more serious study.

[2]The first two mechanisms are closely analogous to steady-state bifurcation and to Hopf bifurcation from equilibria, respectively. The third has no analogue in bifurcation from equilibria.

9.4 Mutual Annihilation of Two Limit Cycles[3]

9.4.1 An Academic Example

Consider a one-parameter family of ODEs on $\mathbb{R}^2 \sim \{\mathbf{0}\}$, written in polar coordinates,

$$\begin{aligned} r' &= (r-1)^2 + \mu, \\ \theta' &= 1. \end{aligned} \tag{9.11}$$

If $\mu = 0$, then $\{r = 1\}$ is a periodic orbit of (9.11), and in Exercise 2 we ask you to show that the lone eigenvalue of the Poincaré map for this orbit equals 1.

What if μ is perturbed slightly? If $-1 < \mu < 0$, then (9.11) has two periodic orbits $\{r = 1 \pm \sqrt{-\mu}\}$, as illustrated in Figure 9.7(a), the inner one stable and the outer one unstable. However, if $\mu > 0$, there are none. As μ passes through zero, the two periodic solutions meet and "annihilate" each other, behavior that is summarized in the bifurcation diagram[4] of Figure 9.7(b).

Although we use the standard term "annihilation," this language might be misleading. Actually, the stable and unstable periodic orbits in Figure 9.7(b) fit together to make a smooth surface in $\mathbb{R}^3$.

Let us examine (9.11) with Poincaré maps. Choosing a Poincaré section $\{\theta = 0, a_1 < r < a_2\}$, we solve the IVP for (9.11) with initial conditions

$$r(0) = b, \qquad \theta(0) = 0$$

to obtain $\Pi(b, \mu) = r(2\pi)$. Periodic solutions of (9.11) near $\{r = 1\}$ are in one-to-one correspondence with fixed points of Π, i.e., with solutions of

$$\Pi(b, \mu) = b. \tag{9.12}$$

When $\mu < 0$, $\Pi(\cdot, \mu)$ has two fixed points, as shown in Panel (c) of Figure 9.7. As $\mu \to 0$, the fixed points approach each other, and when $\mu > 0$, $\Pi(\cdot, \mu)$ has no fixed points, as shown in Panel (d). This bifurcation of periodic solutions is completely analogous to saddle-node bifurcation of equilibria.

9.4.2 The Denatured Morris–Lecar Equation

In Figure 9.5(d) above, which is a bifurcation diagram for the denatured Morris–Lecar equations (9.7), a mutual annihilation bifurcation occurs at $I = I_{\text{mutan}}$. Specifically,

[3] In this bifurcation, $\lambda = 1$ is an eigenvalue of the Poincaré map. Although mutual annihilation is the most common bifurcation that occurs in this situation, it is not the only possibility; cf. Exercise 3.

[4] As $\mu \to -1$, the inner orbit shrinks to the origin. On a superficial level this behavior resembles Hopf bifurcation, but this description is inappropriate, because (9.11) is singular at the origin.

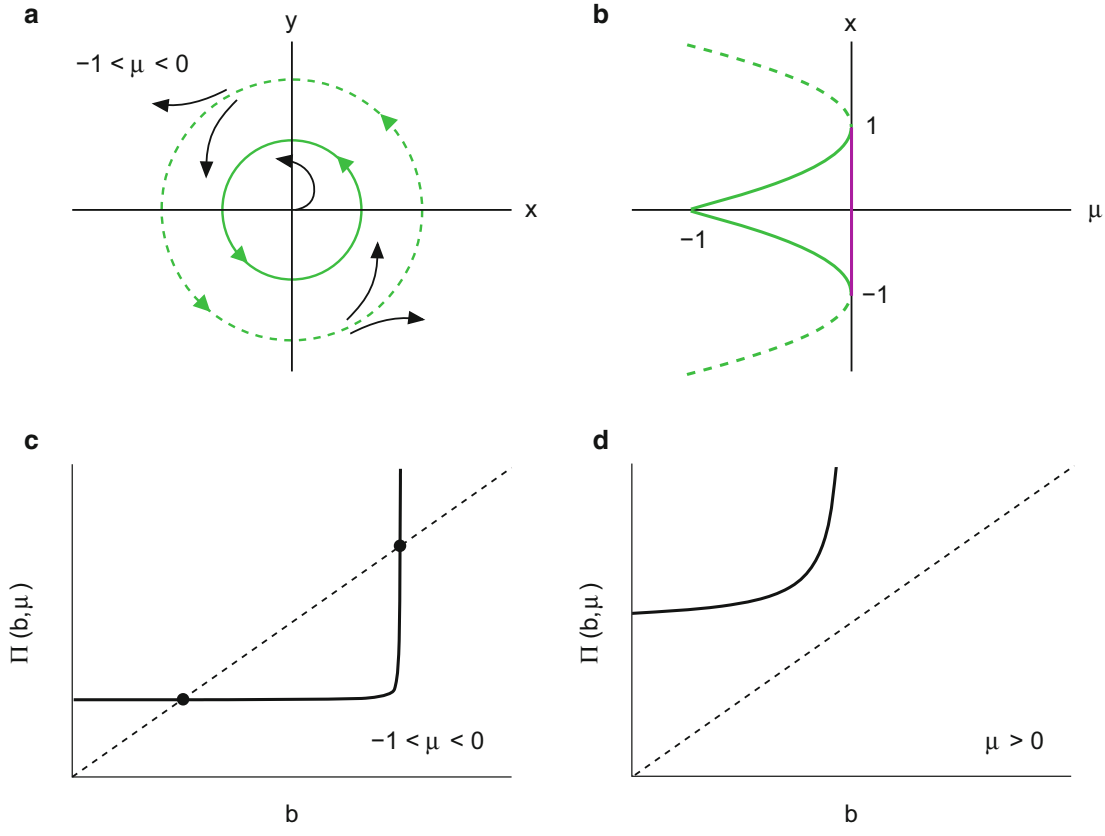

Figure 9.7: *(a) Phase-plane plot for* (9.11) *with* $\mu < 0$*, so the equation has two limit cycles. (For* $\mu > 0$*, not shown, it has no periodic solutions.) (b) Bifurcation diagram for this equation. (c,d) Graphs of the Poincaré maps for this equation.*

for I slightly below the value I_{mutan}, the equations have two periodic solutions, one stable and one unstable; these meet and annihilate one another as I increases through I_{mutan}. The dedicated reader might want to check (numerically) that the eigenvalue of the Poincaré map tends to 1 as $I \to I_{\text{mutan}}$.

9.4.3 Phase-Locking in Coupled Oscillators

Recall from Chapter 7 the coupled-oscillators equations (7.6) on the torus $\mathbb{T}^2$,

$$\begin{aligned} \theta_1' &= \omega_1 + K_1 \sin(\theta_2 - \theta_1), \\ \theta_2' &= \omega_2 + K_2 \sin(\theta_1 - \theta_2). \end{aligned} \tag{9.13}$$

As we saw there, if

$$\left| \frac{\omega_1 - \omega_2}{K_1 + K_2} \right| < 1, \tag{9.14}$$

then (9.13) has two phase-locked periodic solutions, which have the form (7.8). To rephrase this information in the language of bifurcation theory: if starting from parameters that satisfy (9.14), one of the parameters is varied so that the inequality is violated, these two periodic solutions meet and annihilate each other. Incidentally, if you reexamine the stability calculation of Section 7.3.1, you will see that the eigenvalue of the Poincaré map for the stable periodic orbit tends to 1 as the bifurcation point is approached.

In fact, (9.13) exhibits infinitely many mutual annihilation bifurcations. As ω_2/ω_1 varies, the two variables get entrained in various rational frequency ratios, not just $1:1$ as above. The birth and death of every entrainment interval occurs through a mutual annihilation bifurcation. This behavior, which is explored in Section 10.4, is common for ODEs on the torus.

9.5 Hopf-Like Bifurcation to an Invariant Torus

In this bifurcation, sometimes called a Neimark–Sacker bifurcation, a pair of complex-conjugate eigenvalues of the Poincaré map cross the unit circle. In particular, the Poincaré map is at least two-dimensional, which means that the governing ODE must be at least three-dimensional.

9.5.1 An Academic Example

This example embeds the basic illustration of Hopf bifurcation, (8.71), in three dimensions in a construction similar to (7.20). Consider a family of ODEs (with bifurcation parameter μ) written in cylindrical coordinates and defined on $\{(r, \theta, z) : r > 0\}$,

$$\begin{aligned} \theta' &= 1, \\ \begin{bmatrix} r' \\ z' \end{bmatrix} &= \begin{bmatrix} \mu & -\omega \\ \omega & \mu \end{bmatrix} \begin{bmatrix} r-1 \\ z \end{bmatrix} - R^2(r,z) \begin{bmatrix} r-1 \\ z \end{bmatrix}, \end{aligned} \tag{9.15}$$

where ω is a nonzero constant and $R(r,z) = \sqrt{(r-1)^2 + z^2}$. For all μ, it has the periodic solution

$$\theta(t) = t \pmod{2\pi}, \qquad r(t) = 1, \qquad z(t) = 0. \tag{9.16}$$

Relative to the Poincaré section $\{\theta = 0\}$, the Poincaré map for (9.16) involves only the r, z-subsystem of (9.15), and its differential can be found by solving the linearization of this subsystem. Thus, $\mathbf{D}\Pi_\mu = \exp(2\pi C_\mu)$, where C_μ is the matrix in real canonical form

$$C_\mu = \begin{bmatrix} \mu & -\omega \\ \omega & \mu \end{bmatrix}.$$

Now C_μ has eigenvalues $\mu \pm i\omega$, so $\mathbf{D}\Pi_\mu$ has eigenvalues $e^{2\pi\mu}e^{\pm 2\pi i\omega}$. Provided $\omega/2$ is not an integer, these are a pair of complex conjugates, and as μ increases, they cross the unit circle when $\mu = 0$.

How do solutions change when μ moves across zero? If $\mu < 0$, then r and z, whose behavior is independent of θ, decay to $(1,0)$; thus nearby trajectories of (9.15) converge to the periodic trajectory (9.16). When $\mu > 0$, this periodic solution becomes unstable, and trajectories of (9.15) converge to a curve on the torus $\{(r,\theta,z) : R^2(r,z) = \mu\}$, specifically to a solution

$$\theta(t) = t, \quad r(t) = 1 + \sqrt{\mu}\cos(\omega t + \alpha), \quad z(t) = \sqrt{\mu}\sin(\omega t + \alpha), \tag{9.17}$$

where α is an arbitrary phase angle. (See Figure 9.8.) If ω is rational, every bifurcated solution (9.17) is periodic; otherwise, the solution is an (aperiodic) skew line. Which case occurs does not vary with μ.

In general, Theorem 3.5.2 of Guckenheimer and Holmes [33] formulates a rigorous result about bifurcation from a periodic orbit to flow on an invariant torus. As discussed in Section 10.4, the behavior of ODEs on a torus is typically far more complicated than what is presented above. Thus, in bifurcations of this type it may not be possible to determine the nature of the flow on the invariant torus without numerical solutions.

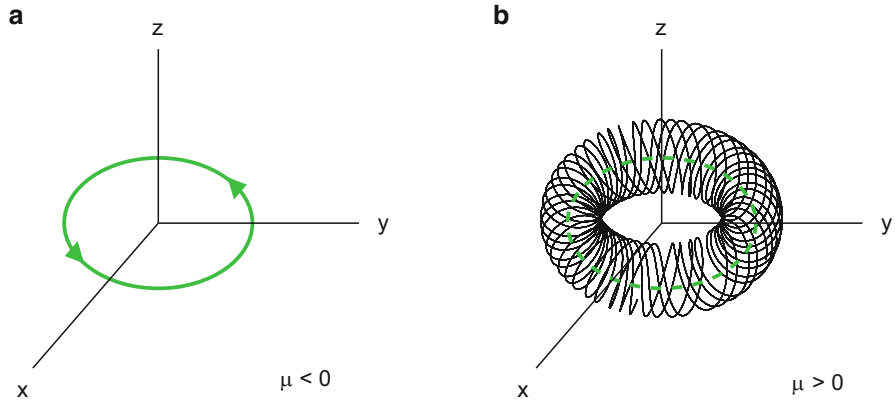

Figure 9.8: *Trajectories of* (9.15) *before and after the bifurcation at* $\mu = 0$. *In Panel (b), it is hard to see, but the (green) periodic orbit inside the torus is unstable and therefore shown as dashed.*

9.5.2 The Periodically Forced van der Pol Equation

Suppose, as in Section 7.9.2, a periodic force $\Gamma \cos \omega t$ is applied in van der Pol's equation. This situation may be described by the autonomous ODE

$$\begin{aligned} x' &= y, \\ y' &= -x - \beta(x^2 - 1)y + \Gamma \cos z, \\ z' &= \omega, \end{aligned} \tag{9.18}$$

where z is reduced modulo 2π. Consider (9.18) as a bifurcation problem with Γ as the bifurcation parameter.

If $\beta = 0$, equation (9.18) admits the explicit solution

$$x(t) = \frac{\Gamma}{1-\omega^2} \cos \omega t, \qquad y(t) = -\omega \frac{\Gamma}{1-\omega^2} \sin \omega t, \qquad z(t) = \omega t \pmod{2\pi}, \tag{9.19}$$

with the same period $2\pi/\omega$ as the forcing. (Cf. Exercise 1.13.) This solution is Lyapunov stable but not asymptotically stable.

In Section 7.9.2 we argued that for every Γ and $\beta > 0$, (9.18) has at least one $2\pi/\omega$-periodic solution. However, if Γ and β are large, there may also be some extremely complicated solutions, possibly chaotic (e.g., see Section 2.1 of [33]). Therefore, we consider only one case, with $\beta = 0.1$, $\omega = 0.3$, and Γ restricted to moderate values. For example, if $\Gamma = 1.3$, the green curve in Figure 9.9(a) shows (the projection into the x, y-plane of) one such periodic solution, which, moreover, is asymptotically stable. Indeed, reflecting the fact that β is small, this solution doesn't differ greatly from the corresponding solution (9.19) of the frictionless problem, which is shown in black.

At first it might seem surprising that a solution of (9.18) could be asymptotically stable, because the van der Pol equation has "negative friction" for $|x| < 1$, which is a destabilizing effect. However, the solution spends more time in the stable region $|x| > 1$ than in the unstable region, so on balance stability wins out.

But consider the effect of reducing the forcing coefficient Γ. Having a smaller force will decrease the amplitude of x in the solution, increasing the fraction of time spent in the unstable region. In other words, you might expect the periodic solution to lose stability as Γ decreases. Panels (b) and (c) confirm this suspicion; they show three-dimensional pictures of the flow before and after the bifurcation, which occurs at $\Gamma \approx 1.285$ when a pair of complex-conjugate eigenvalues of the Poincaré map cross the unit circle. The apparent torus in Panel (c) consists of many distinct curves; the solution of (9.18) is computed over a long time interval, and intervals $\{2n\pi < z < (2n+1)\pi\}$ on the solution curve are translated back to $[0, 2\pi]$.

Remarks: (i) In Exercise 7 we ask you to verify that the eigenvalues of an appropriate Poincaré map cross the unit circle. *(ii)* Fourier analysis is needed for a fuller understanding of the bifurcation, but this lies beyond the scope of the present book. (Chapter III of [8] gives a compact introduction to Fourier methods in the context of dynamical systems. See also Section 8.5 of [18].)

9.5.3 Other Examples of Bifurcation to an Invariant Torus

The Rosenzweig–MacArthur predator–prey model, which exhibits a Hopf bifurcation, was introduced in Exercise 8.22. Let us modify it to allow for a seasonal variation in the linear growth rate for the prey:

$$\begin{aligned} x' &= (1 + \gamma \cos \nu t - x)x - \frac{xy}{1+\sigma x}, \\ y' &= E\frac{xy}{1+\sigma x} - \mu y. \end{aligned} \tag{9.20}$$

Provided μ is fairly large, (9.20) has a stable $2\pi/\nu$-periodic solution. When μ gets sufficiently small, the Hopf bifurcation of the Rosenzweig–MacArthur model makes itself felt as a bifurcation of this periodic solution of (9.20) to an invariant torus. We invite you to explore this behavior in Exercise 8.

The above bifurcations of (9.18) and (9.20) depend on having time-varying forcing. In many fluid-mechanics problems, bifurcation to an invariant torus occurs even with time-independent external driving, as we discuss briefly in Section 9.10.2.

9.6 Period-Doubling

In this bifurcation, an eigenvalue of the Poincaré map passes through -1. In a two-dimensional autonomous ODE, the Poincaré map of a periodic solution cannot have

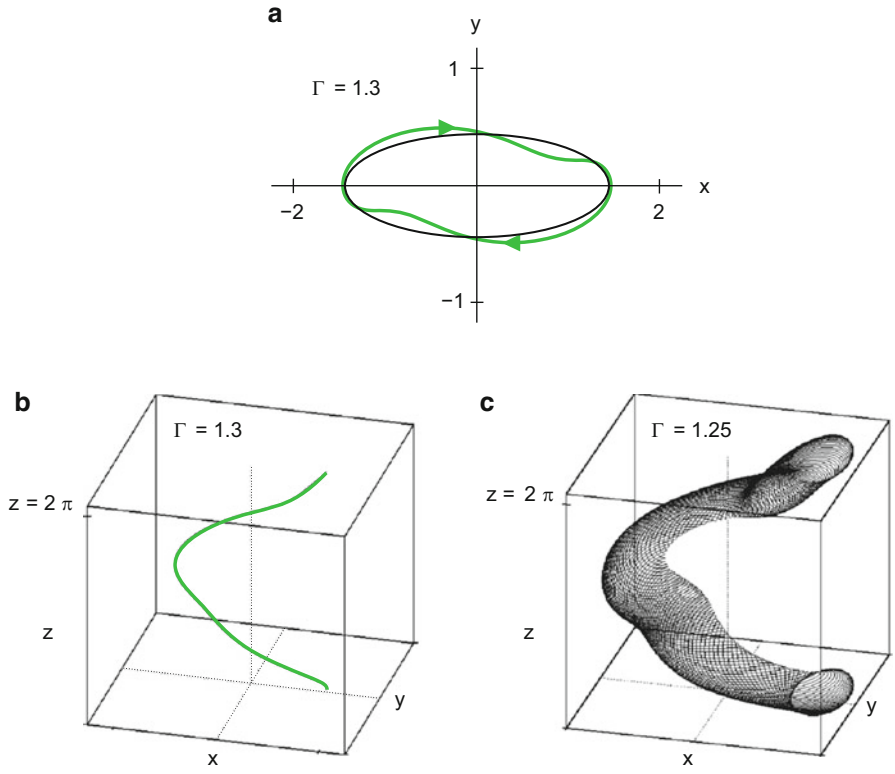

Figure 9.9: *Trajectories of* (9.18) *with* $\beta = 0.1$ *and* $\omega = 0.3$. *Panel (a) shows the two-dimensional projection of the stable periodic solution with* $\Gamma = 1.3$ *(in green), along with the solution* (9.19) *of the frictionless equation (in black). Panels (b) and (c) show perspective drawings of three-dimensional flow before and after the bifurcation at* $\Gamma \approx 1.285$. *In the figures,* z *is plotted reduced modulo* 2π*; consequently, in Panel (c) we may see the outline of the invariant torus.*

a negative eigenvalue: if a trajectory starts outside a periodic orbit, it must stay outside. Thus, an (autonomous) ODE exhibiting this bifurcation must be at least three-dimensional.

In fact, the simplest examples of period-doubling bifurcation occur for *maps*, not necessarily related to any ODE, i.e., in the language of Section 7.3.3, for discrete dynamical systems. Before proceeding to examples with ODEs, we first introduce period-doubling in mappings on the line.

9.6.1 Academic Example 1: Mappings

Let $\Psi : \mathbb{R} \times [0, \infty) \to \mathbb{R}$ be the one-parameter family of mappings given by

$$\Psi(x, \mu) = -\mu \frac{x}{\sqrt{1 + x^2}}. \tag{9.21}$$

Observe that for all μ, the origin $x = 0$ is a fixed point of $\Psi(\cdot, \mu)$. This trivial fact may also be seen graphically from Figures 9.10(a,b). To assess its stability, calculate that

$$D\Psi(0, \mu) = \frac{\partial \Psi}{\partial x}(0, \mu) = -\mu.$$

Thus, according to Theorem 7.3.7, the fixed point is asymptotically stable if $0 \le \mu < 1$ and unstable if $\mu > 1$. The (unique) eigenvalue of $D\Psi(0, \mu)$ passes through -1 as μ crosses 1.

What new phenomena appear as a result of this loss of stability? There are no new fixed points; it is apparent from Figure 9.10(b) that $x = 0$ is the only fixed point of $\Psi(\cdot, \mu)$, no matter how large the value of μ. But let us consider the iterate,

$$\Psi^2(x, \mu) = \Psi(\Psi(x, \mu), \mu) = \mu^2 \frac{x}{\sqrt{1 + (\mu^2 + 1)x^2}},$$

which is graphed in Figures 9.10(c,d). For all μ, the origin is of course a fixed point of $\Psi^2(\cdot, \mu)$. If $\mu < 1$, this is the only fixed point of $\Psi^2(\cdot, \mu)$, but when μ passes through unity, $\Psi^2(\cdot, \mu)$ acquires *two new fixed points*. Analytically, you may calculate that $\pm\sqrt{\mu^2 - 1}$ are fixed points of $\Psi^2(\cdot, \mu)$ when $\mu > 1$.

To rephrase this behavior in suggestive language for ODEs, let us consider the discrete dynamical system

$$x_{n+1} = \Psi(x_n, \mu), \quad n = 0, 1, 2, \dots . \tag{9.22}$$

We may regard the fixed point $x = 0$ as a periodic trajectory of this dynamical system *with period one*. When $\mu > 1$, $\Psi(\pm\sqrt{\mu^2 - 1}) = \mp\sqrt{\mu^2 - 1}$, so we may regard $\{\pm\sqrt{\mu^2 - 1}\}$ as a periodic trajectory of (9.22) *with period two*. The bifurcation mantra is that when a fixed point of a mapping loses stability because an eigenvalue of the differential passes through -1, expect period-doubled solutions to appear.

9.6.2 Cardiac Alternans

Following a purely academic example, let's consider an example of a meaningful, even potentially dangerous, period-doubling bifurcation in a mapping, cardiac alternans. Your heart's ability to pump blood efficiently relies upon an incredibly well coordinated electrical signaling system. Information about electrical activity and heart rhythm can be extracted from routine, noninvasive electrocardiograms (ECGs). No doubt you have seen output from an ECG. A schematic trace of the output from an ECG lead appears in Figure 9.11(a). It is less likely that you have seen a recording of electrical activity from an *individual* cardiac muscle cell. Figure 9.11(b) illustrates the voltage measured across a single cell membrane as a function of time during three consecutive heartbeats. Note that the voltage is elevated during a portion of each beat; such elevations are known as *action potentials*. As indicated in Figure 9.11(b), let APD_n denote the duration of the nth action potential, i.e., the length of time that the voltage exceeds a threshold value. The dashed vertical lines in Figure 9.11 indicate how APDs correlate with certain features of an ECG recording.

Phenomenological models[5] for APDs can be based on a property known as *restitution*: as observed by Nolasco and Dahlen [61], each APD can be expressed (approximately) as a function of the time the heart has to rest following the previous action potential, what is called the *diastolic interval*. In symbols, using the notation indicated in Figure 9.11(b),

$$\mathrm{APD}_{n+1} = F(\mathrm{DI}_n). \tag{9.23}$$

A common approximation for the restitution function F has the form

$$F(\mathrm{DI}) = C_1 - C_2 e^{-\mathrm{DI}/\tau}, \tag{9.24}$$

where C_1, C_2, τ are constants. For example, the data of [34] was fit with

$$C_1 = 392, \quad C_2 = 525, \quad \tau = 40, \tag{9.25}$$

where all variables and parameters are measured in milliseconds.

Assuming such a model, we suppose that stimuli for action potentials arrive periodically, with a uniform spacing B between consecutive stimuli. Thus, $\mathrm{DI}_n = B - \mathrm{APD}_n$. Substituting into (9.23) and defining $\Psi(\mathrm{APD}, B) = F(B - \mathrm{APD})$, we may describe the sequence $\{\mathrm{APD}_n\}$ as the trajectory of a discrete dynamical system

$$\mathrm{APD}_{n+1} = \Psi(\mathrm{APD}_n, B), \quad n = 0, 1, 2, \dots, \tag{9.26}$$

[5]For modeling at a more fundamental level, we refer you to Keener and Sneyd [47], who give a beautiful introduction to electrophysiology from a mathematician's perspective.

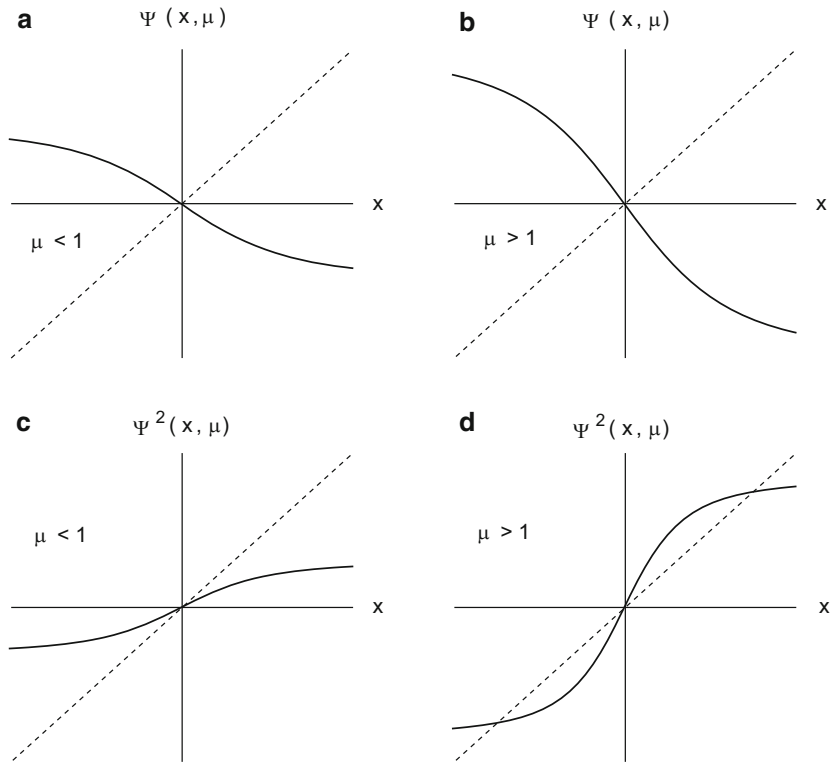

Figure 9.10: *Panels (a,b): graphs of the mapping* (9.21) *for* $\mu < 1$ *and* $\mu > 1$, *with the diagonal* $\{y = x\}$ *indicated by a dotted line. These panels show that the origin is the only fixed point of* $\Psi(\cdot, \mu)$, *no matter how large* μ *may be. Panels (c,d): graphs of the iterated map* $\Psi^2(\cdot, \mu)$. *These panels show that while for* $\mu < 1$, *the iterated map has only one fixed point, for* $\mu > 1$ *it has three. The nonzero fixed points of* $\Psi^2(\cdot, \mu)$ *in Panel (d) may be interpreted as period-two trajectories of the discrete dynamical system* (9.22).

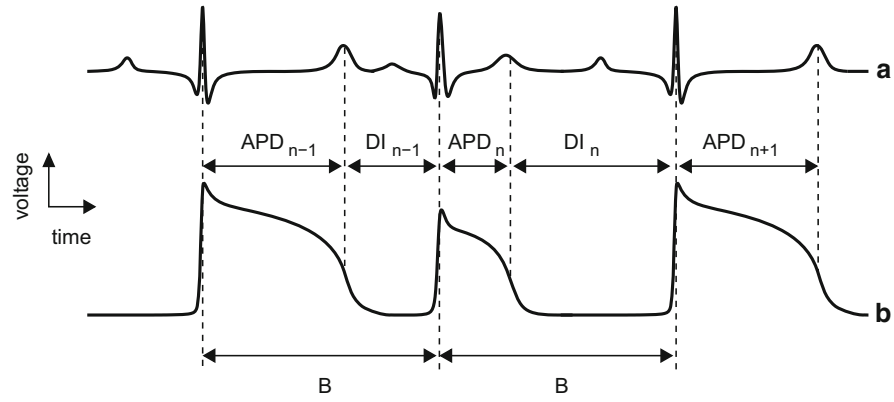

Figure 9.11: *(a) Hypothetical recording of one lead of an electrocardiogram over three heartbeats. (b) Action potentials in an individual cardiac cell.*

where we consider B a bifurcation parameter. We shall not specify a precise domain for the function Ψ, although some physiological constraints must be satisfied.[6]

As suggested by the graphs in Figure 9.12(a), for every B within the physiological range, $\Psi(\cdot, B)$ has a unique fixed point APD_* satisfying the transcendental equation

$$\mathrm{APD}_* = C_1 - C_2 e^{-(B-\mathrm{APD}_*)/\tau}.$$

Is it stable? If B is large (slow stimulation), the slope of $\Psi(\cdot, B)$ at the fixed point is close to zero; in particular, it has absolute value less than 1, so the fixed point is stable. However, as B decreases (more rapid stimulation), the slope of $\Psi(\cdot, B)$ at the fixed point becomes more negative, suggesting that the fixed point could become unstable as B decreases. In fact, for the mapping (9.24) with the parameters (9.25), a period-doubling bifurcation occurs at $B \approx 455$ ms, as shown in the bifurcation diagram Figure 9.12(b). The period-2 response induced by this bifurcation, known as *alternans*, is an abnormal rhythm that is viewed as a precursor to potentially fatal arrhythmias.

Incidentally, mapping models of restitution have been derived via asymptotics from ODE models of the electrical behavior of cardiac cells [56]. In Exercise 10, you are asked to explore such an ODE model numerically, including finding the period-doubling bifurcation to alternans.

[6]For instance, a stimulus will fail to elicit an action potential if DI is too short. Such considerations are important for modeling, but for the purposes of this subsection they are a distraction that we sweep under the rug.

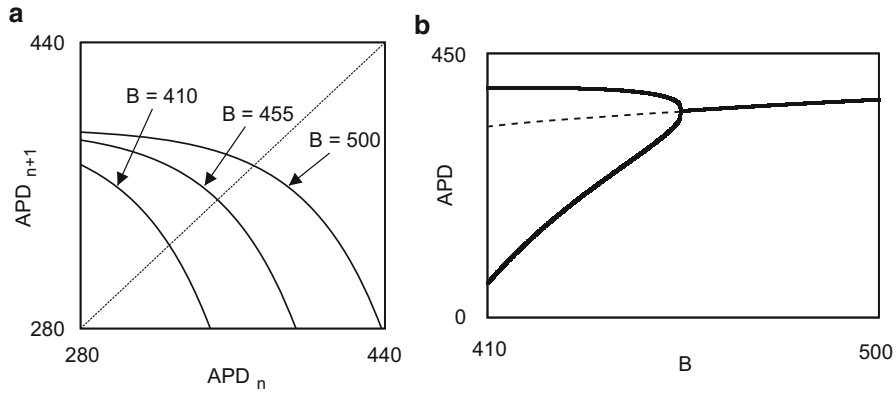

Figure 9.12: *(a) Graphs of* $\Psi(APD, B) = F(B - APD)$ *from* (9.24) *for three choices of* B*. Intersections with the diagonal line are fixed points. (b) Loss of stability of the fixed point at* $B \approx 455$ *ms and bifurcation of a period-doubled response.*

9.6.3 Academic Example 2: ODEs

We introduce period-doubling in ODEs with the system written in cylindrical coordinates on the domain $\{(r, \theta, z) : r > 0\}$,

$$\begin{aligned} \theta' &= 1, \\ \begin{bmatrix} r' \\ z' \end{bmatrix} &= (\Omega - I + \mu M(\theta)) \begin{bmatrix} (r-1) \\ z \end{bmatrix} - R^2(r,z) \begin{bmatrix} (r-1) \\ z \end{bmatrix}, \end{aligned} \tag{9.27}$$

where in the linear term, I is the 2×2 identity matrix,

$$\Omega = \begin{bmatrix} 0 & -1/2 \\ 1/2 & 0 \end{bmatrix}, \quad \text{and} \quad M(\theta) = \begin{bmatrix} \cos^2(\theta/2) & \sin(\theta/2)\cos(\theta/2) \\ \sin(\theta/2)\cos(\theta/2) & \sin^2(\theta/2) \end{bmatrix},$$

and in the nonlinear term, $R(r,z) = \sqrt{(r-1)^2 + z^2}$. For all μ, formula (9.16) defines a 2π-periodic solution of this ODE. The following lemma addresses its stability.

Lemma 9.6.1. *The Poincaré map for* (9.16) *has negative real eigenvalues,* $-e^{2\pi(\mu-1)}$ *and* $-e^{-2\pi}$*.*

Proof. We define the Poincaré map based on the transverse section $\{\theta = 0 \pmod{2\pi}\}$. To calculate the eigenvalues of $\mathbf{D\Pi}_\mu$, we need to solve the linearized r, z-subsystem

$$\begin{bmatrix} w_r' \\ w_z' \end{bmatrix} = (\Omega - I + \mu M(t)) \begin{bmatrix} w_r \\ w_z \end{bmatrix}, \tag{9.28}$$

where we have substituted $\theta(t) = t$ from (9.16). First consider the simpler system

$$\begin{bmatrix} w_r' \\ w_z' \end{bmatrix} = (\Omega - I) \begin{bmatrix} w_r \\ w_z \end{bmatrix}, \tag{9.29}$$

with $M(t)$ omitted. Two linearly independent solutions of this equation are

$$e^{-t} \begin{bmatrix} \cos t/2 \\ \sin t/2 \end{bmatrix}, \qquad e^{-t} \begin{bmatrix} -\sin t/2 \\ \cos t/2 \end{bmatrix}.$$

Preparing to reinsert $M(t)$ into the problem, we observe that

$$M(t) \begin{bmatrix} \cos t/2 \\ \sin t/2 \end{bmatrix} = \begin{bmatrix} \cos t/2 \\ \sin t/2 \end{bmatrix}, \qquad M(t) \begin{bmatrix} -\sin t/2 \\ \cos t/2 \end{bmatrix} = \mathbf{0}.$$

In other words, $M(t)$ has eigenvalues 1 and 0, and the eigenvectors of $M(t)$ rotate exactly in synchrony with the solutions of (9.29). Therefore

$$\mathbf{v}_1(t) = e^{(\mu-1)t} \begin{bmatrix} \cos t/2 \\ \sin t/2 \end{bmatrix}, \qquad \mathbf{v}_2(t) = e^{-t} \begin{bmatrix} -\sin t/2 \\ \cos t/2 \end{bmatrix}$$

solve (9.28). Since $\mathbf{v}_1(0) = \mathbf{e}_r$ and $\mathbf{v}_2(0) = \mathbf{e}_z$, we conclude that

$$\mathbf{D\Pi}_\mu \, \mathbf{e}_r = \mathbf{v}_1(2\pi) = -e^{2\pi(\mu-1)}\mathbf{e}_r, \qquad \mathbf{D\Pi}_\mu \, \mathbf{e}_z = \mathbf{v}_2(2\pi) = -e^{-2\pi}\mathbf{e}_z.$$

Thus, $\mathbf{e}_r, \mathbf{e}_z$ are eigenvectors of $\mathbf{D\Pi}_\mu$ with the claimed eigenvalues. □

Note that an eigenvalue of $\mathbf{D\Pi}_\mu$ passes through -1 when $\mu = 1$. Thus, the periodic solution (9.16) is stable if $\mu < 1$ but becomes unstable as μ increases beyond this limit. Because of the artificial simplicity of (9.27), we can actually exhibit new solutions that appear when $\mu > 1$:

$$\begin{bmatrix} \theta(t) \\ r(t) \\ z(t) \end{bmatrix} = \begin{bmatrix} t + \alpha \\ 1 + \sqrt{\mu - 1}\, \cos(t/2 + \alpha) \\ \sqrt{\mu - 1}\, \sin(t/2 + \alpha) \end{bmatrix}, \tag{9.30}$$

where α is an arbitrary phase angle. These solutions are periodic, but with period 4π, twice the period of (9.16). If $\alpha = 0$, then at time zero, (9.30) starts at a point in the Poincaré section $\{\theta = 0\}$ just "outside" the loop formed by (9.16). As time increases, (9.30) moves along with (9.16) while rotating in the r, z-plane around the point $(1, 0)$. (Cf. Figure 9.13.) At time 2π, the trajectory returns to the Poincaré section, but it has completed only half of a rotation around $(1, 0)$. During the remainder of a period, $2\pi \le t \le 4\pi$, the trajectory completes the rotation around $(1, 0)$ and returns

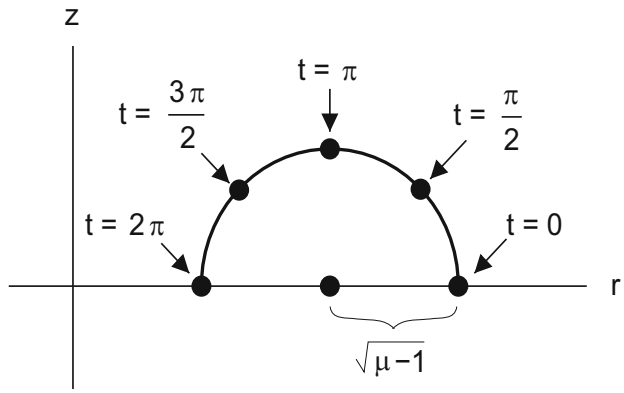

Figure 9.13: *Schematic of the motion in the r, z-plane (θ suppressed) of the period-doubled solution (9.30) with $\alpha = 0$ during the first half of its period, $0 \le t \le 2\pi$. In the second half of its period, it completes the revolution around $(1, 0)$.*

to its starting point as it crosses the Poincaré section.[7]

We invite you to compute with (9.27) to see that when $\mu > 1$, virtually all solutions of this equation are asymptotic, as $t \to \infty$, to one of the solutions (9.30).

Perhaps surprisingly, this example is more robust than might be apparent. For example, suppose we slightly change the rotation speed in Ω or perturb $M(\theta)$ in some way. This will alter the eigenvalues of the Poincaré map, but these (distinct) eigenvalues vary continuously with the perturbation. Therefore, we may deduce that as μ varies, one eigenvalue of the Poincaré map remains real and close to zero, while the other passes through -1 somewhere near $\mu = 1$. In other words, in the perturbed problem the periodic solution (9.16) experiences the same loss of stability. Although we can't calculate eigenvalues or solutions in the perturbed problem, a general theoretical result (cf. Theorem 3.5.1 of Guckenheimer and Holmes [33]) guarantees that a stable period-doubled solution still bifurcates. In Exercise 9 we ask you to perturb (9.27) and verify this claim with appropriate computations.

9.6.4 A Periodically Forced Pendulum

Consider a pendulum subjected to periodic forcing; after nondimensionalization, the motion may be described by the ODE

$$x'' + \beta x' + \sin x = \Gamma \cos \omega t. \tag{9.31}$$

As the forcing amplitude Γ increases, the long-term behavior of solutions of (9.31) changes through multiple bifurcations, including period-doubling bifurcations. In

[7]Do you have enough artistic talent to make a perspective drawing that does justice to the three-dimensionality of this trajectory? If so, please send it to us, and we will put it on the web page, along with a grateful acknowledgment.

fact, this system admits such a bewildering variety of solutions that we consider it only for a limited range of Γ with the specific parameters[8]

$$\beta = 0.5, \qquad \omega = 1, \tag{9.32}$$

which yields the fairly simple behavior discussed below.

Incidentally, let us record a symmetry of (9.31): if $x(t)$ satisfies this equation, so does

$$\tilde{x}(t) = -x(t + \pi/\omega). \tag{9.33}$$

For small Γ, (9.31) has a unique $2\pi/\omega$-periodic solution, say $\gamma(t)$, and it is globally attracting. The phase portrait of such a solution for $\Gamma = 1.8$ is shown in Figure 9.14(a). Given the parameters (9.32) and making the approximation $\sin x \approx x$, we may estimate that

$$\gamma(t) \approx A \sin t \qquad \text{where } A = 2\Gamma = 3.6. \tag{9.34}$$

Although the approximation is qualitatively correct, nonlinear effects are already sufficient to undermine its quantitative accuracy. These solutions are invariant under the symmetry (9.33); thus, the orbit in Figure 9.14(a) is invariant under reflection through the origin in $\mathbb{R}^2$.

The first bifurcation of (9.31), which occurs when $\Gamma = \Gamma_{\text{pitch}} \approx 2$, breaks the symmetry (9.33). The invariant 2π-periodic solution continues to exist but is unstable; two new solutions with the same period 2π appear at the bifurcation.[9] One new solution, for $\Gamma = 2.2$, is shown in Figure 9.14(b). The other solution, not shown, is the image of the first under the symmetry (9.33); its orbit is the reflection of the orbit in the figure through the origin. As Γ passes Γ_{pitch}, an eigenvalue of the Poincaré map *(How is this defined for* (9.31)*?)* crosses $+1$. (Cf. Exercise 15.)

The next bifurcation is the reason we chose this example. As Γ crosses $\Gamma_{\text{per-dbl}} \approx 2.318$, the (asymmetric) trajectory illustrated in Figure 9.14(b) undergoes a period-doubling bifurcation. A typical trajectory after the bifurcation, which has period 4π, is shown in Figure 9.14(c). These phase-plane plots make a convincing argument that an eigenvalue of the Poincaré map has passed through -1, but the dedicated reader may wish to verify this behavior numerically.

As Γ further increases, additional period-doubling bifurcations appear. For example, a period-quadrupled solution is illustrated in Figure 9.14(d). However, we do not pursue this behavior further.

---

[8]Note that this forcing is at the resonant frequency for the linearization of (9.31). This fact is not important for the phenomena we study; it merely simplifies formula (9.34) for the approximate solution.

[9]This bifurcation illustrates that, as mentioned above, mutual annihilation of limit cycles is not the only possibility when an eigenvalue of the Poincaré map crosses $+1$. Cf. Exercise 3.

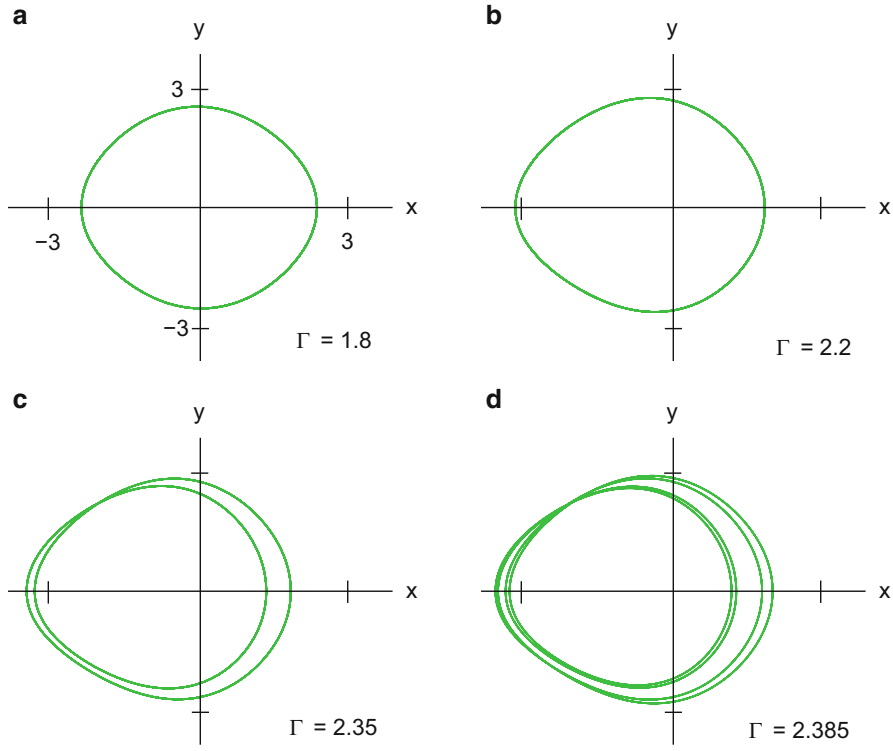

Figure 9.14: *Bifurcations of the forced pendulum* (9.31) *as* Γ *is varied, assuming* $\beta = 1/2$ *and* $\omega = 1$. *As described in the text, the panels indicate that a symmetry-breaking bifurcation occurs between* $\Gamma = 1.8$ *and* $\Gamma = 2.2$, *a period-doubling bifurcation between* $\Gamma = 2.2$ *and* $\Gamma = 2.35$, *and another period-doubling bifurcation between* $\Gamma = 2.35$ *and* $\Gamma = 2.385$.

Mostly we are relying on the computer to demonstrate these bifurcations, since rigorous analysis is unrewardingly technical. In Exercise 15, we relate the bifurcations to parametric resonance, which provides some analytical insight into the phenomena.

9.6.5 Rössler's Equation

Rössler's equation

$$\begin{aligned} x' &= -y - z, \\ y' &= x + ay, \\ z' &= b + z(x - c), \end{aligned} \tag{9.35}$$

provides the most elegant demonstration of period-doubling in an ODE. It is not derived from any application; rather, Rössler created it as a simplification of the Lorenz equation, a model of a model. In Exercise 8.17 you showed that (9.35) has

a saddle-node bifurcation when $c = c_* = 2\sqrt{ab}$ at which two equilibria appear. If $a = b < \sqrt{2}$, all three eigenvalues of $\mathbf{DF}_*$ lie on the imaginary axis, so the principle of exchange of stability provides no information. Lacking theoretical guidance, we blithely plow ahead with computations.

If $a = b = 0.2$, equation (9.35) has stable periodic solutions as soon as $c > 2a$. Initially, as c increases, the amplitude and the basin of attraction of these solutions grow. This trend is illustrated in Figure 9.15(a,b), which shows projections of solution trajectories into the xy-plane for $c = 1, 2$. However, if c is increased to 3 as in Figure 9.15(c), the projection of the limit cycle makes two loops around the origin. (Of course in three dimensions, the trajectory does not cross itself.) The trajectory is still periodic, but it has roughly twice the period as when $c = 2$. For most investigators this is adequate confirmation of a period-doubling bifurcation, but you can remove any vestige of doubt by computing that an eigenvalue of the Poincaré map tends to -1 as $c \to 2.832$. (Cf. Exercise 11.)

Another period-doubling bifurcation occurs as c continues to increase, at $c \approx 3.837$. The period of the solution shown in Figure 9.15(d) has period approximately *four* times that in Figure 9.15(b).

And the action continues: there is, in fact, an *infinite sequence* of period-doubling bifurcation as c increases! The bifurcation diagram in Figure 9.16(a) gives an overview of the bifurcations. When $c < 2.832$, the x component of the periodic solution has exactly one local minimum and one local maximum per period; these are graphed as functions of c in the figure. After the first period-doubling, when $2.832 < c < 3.837$, $x(t)$ has two local minima and two local maxima, all of which are graphed in the figure. The number of local minima and local maxima continues to double at each bifurcation. The bifurcations accumulate at a point with $c \approx 4.3$. Beyond that point, there are infinitely many local extrema of $x(t)$; some of these are indicated by small black dots in the figure, but they form a blur rather than recognizable curves.

Figure 9.16(b) shows a portion of a trajectory computed for $c = 5$, over the range $0 < t < 900$. This bounded aperiodic behavior is typical of *deterministic chaos*.[10] The geometry of the trajectories, including the Poincaré map of (9.35), is explored in Sections 3.4 and 4.1 of Part 2 of [1]. (Their analysis is summarized in Section 12.3 of [81].)

Chaotic behavior reappears briefly in Exercise 13 and in Sections 9.7 and 10.6 of this book, but we do not explore chaos beyond these minor skirmishes. Part III of [81] and [17] give readable introductions to chaos.

[10]We use this term informally without defining it precisely. Some features of chaotic behavior are explored in Section 10.6.

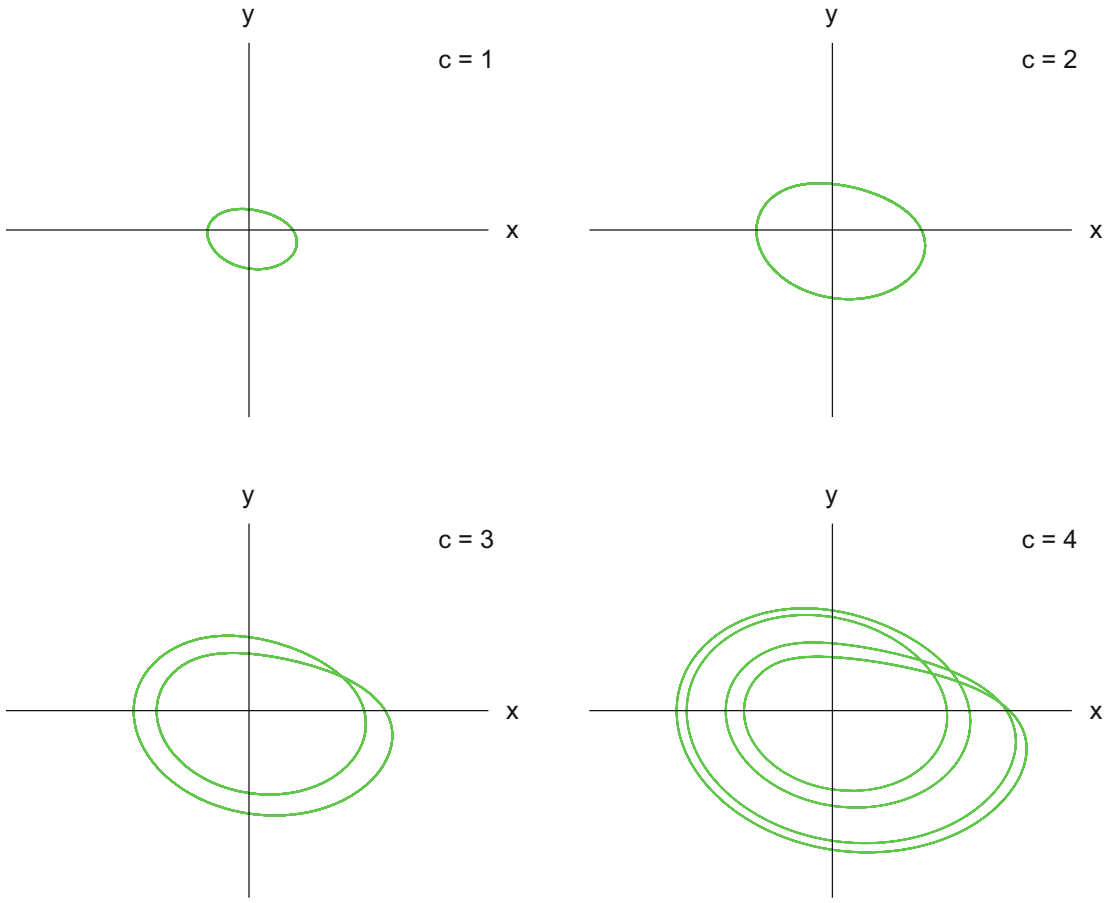

Figure 9.15: *Projections into the x, y-plane of solutions of the Rössler system (9.35), assuming $a = b = 0.2$. Note the period-doubling bifurcations in this system as c increases.*

9.6.6 Other Examples

The period-doubling cascade we just saw is not an uncommon phenomenon. The simplest instance of it occurs in a discrete dynamical system, the quadratic map

$$\Psi(x, \mu) = \mu x(1 - x) \tag{9.36}$$

on the unit interval $[0, 1]$. In Exercise 13 we invite you to use the computer to discover this phenomenon for yourself.

The Lorenz equations exhibit several period-doubling cascades (cf. Chapter 4 of [75]), but in a range of the bifurcation parameter ρ well beyond what we consider in the next section.

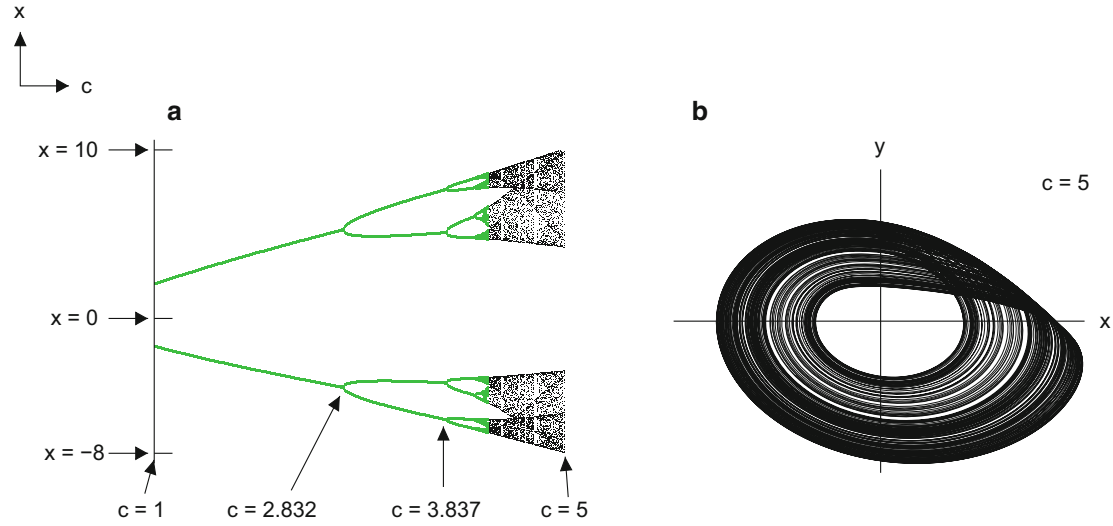

Figure 9.16: *(a) Bifurcation diagram for the period-doubling cascade in the Rössler system (9.35), assuming $a = b = 0.2$. (b) Projection of a typical orbit in the chaotic regime.*

9.7 The Onset of Chaos in the Lorenz Equations

Following well-established tradition in this subject, we study the Lorenz equations

$$\begin{array}{lrcl} (a) & x' & = & \sigma(y - x), \\ (b) & y' & = & \rho x - y - xz, \\ (c) & z' & = & -\beta z + xy, \end{array} \tag{9.37}$$

with ρ as the bifurcation parameter and the other two parameters fixed: $\sigma = 10$ and $\beta = 8/3$. The origin is globally stable if $\rho < 1$; if ρ is increased to, say, 28, virtually all solutions of these equations are chaotic. (A typical solution in the chaotic regime is illustrated in Figure 9.17(a).) This radical change in behavior occurs through several bifurcations of familiar types that are indicated schematically in Figure 9.17(b). The present section guides you through these transitions.

The *first bifurcation* is the supercritical pitchfork bifurcation at $\rho = \rho_{\text{pitch}} = 1$, discussed in Section 8.1. At this bifurcation, two nontrivial equilibria

$$\mathbf{P}_{\pm} = (\pm\sqrt{\beta(\rho - 1)}, \pm\sqrt{\beta(\rho - 1)}, \rho - 1)$$

bifurcate from the origin. After the bifurcation, the origin is a saddle point equilibrium with a one-dimensional unstable manifold $\mathcal{M}_u$. For example, Figure 9.18(a) shows the projection of $\mathcal{M}_u$ into the x, y-plane for ρ a little beyond the bifurcation.

Tracking the asymptotic behavior of $\mathcal{M}_u$ as ρ increases provides the key to understanding the subsequent bifurcations. Because of symmetry, we need track only

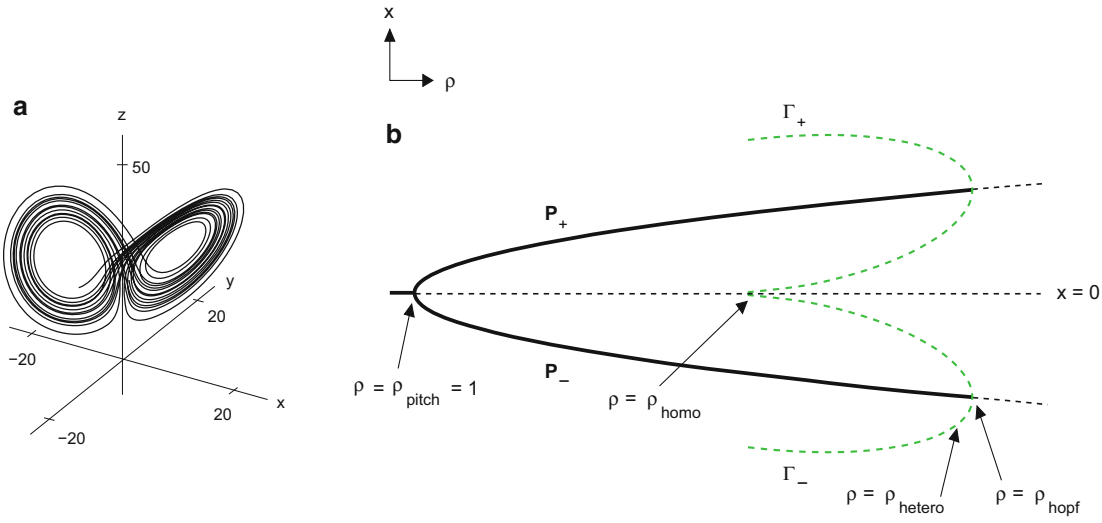

Figure 9.17: *(a) A typical trajectory for the Lorenz system* (9.37) *assuming* $\rho = 28$, $\sigma = 10$, *and* $\beta = 8/3$. *Computations indicate that such orbits are not periodic. (b) A bifurcation diagram for* (9.37). *The bifurcations at* $\rho_{\text{homo}} \approx 13.926$, $\rho_{\text{hetero}} \approx 24.06$, *and* $\rho_{\text{hopf}} \approx 24.74$ *are discussed in the text. Note that the value of* ρ *used in Panel (a) lies beyond all the bifurcations in the diagram.*

Range of ρ	Asymptotic behavior
$\rho_{\text{pitch}} < \rho < \rho_{\text{homo}}$	$\mathcal{M}_u^{(+)}$ converges to $\mathbf{P}_+$
$\rho_{\text{homo}} < \rho < \rho_{\text{hetero}}$	$\mathcal{M}_u^{(+)}$ converges to $\mathbf{P}_-$
$\rho_{\text{hetero}} < \rho$	$\mathcal{M}_u^{(+)}$ never converges

Table 9.1: *Asymptotic behavior of the unstable manifold through the origin for* (9.37) *in various ranges of the bifurcation parameter.*

half of $\mathcal{M}_u$. Let $\mathcal{M}_u^{(+)}$ be the half of $\mathcal{M}_u$ that as $t \to -\infty$, approaches the origin from the first quadrant.[11] As shown in Figures 9.18(a,b) and noted in Table 9.1, before the second bifurcation, $\mathcal{M}_u^{(+)}$ converges to $\mathbf{P}_+$ as $t \to \infty$. (For the record: the convergence of $\mathcal{M}_u^{(+)}$ to $\mathbf{P}_+$ is a spiral only if $\rho > 1.3456$; for smaller ρ, the convergence is monotone.)

As ρ increases, $\mathcal{M}_u^{(+)}$ swings back closer and closer to the origin before spiraling into $\mathbf{P}_+$, until at $\rho = \rho_{\text{homo}} \approx 13.926$ it approaches the origin again asymptotically as $t \to \infty$ (cf. Figure 9.18(c)). This is the *second bifurcation*, a homoclinic bifurcation. After the bifurcation, $\mathcal{M}_u^{(+)}$ approaches the *other* equilibrium, $\mathbf{P}_-$, asymptotically.

[11] To be completely accurate, we should say that the *projection* of $\mathcal{M}_u^{(+)}$ approaches the origin from the first quadrant. Here and below we gloss over this technical point to simplify the syntax.

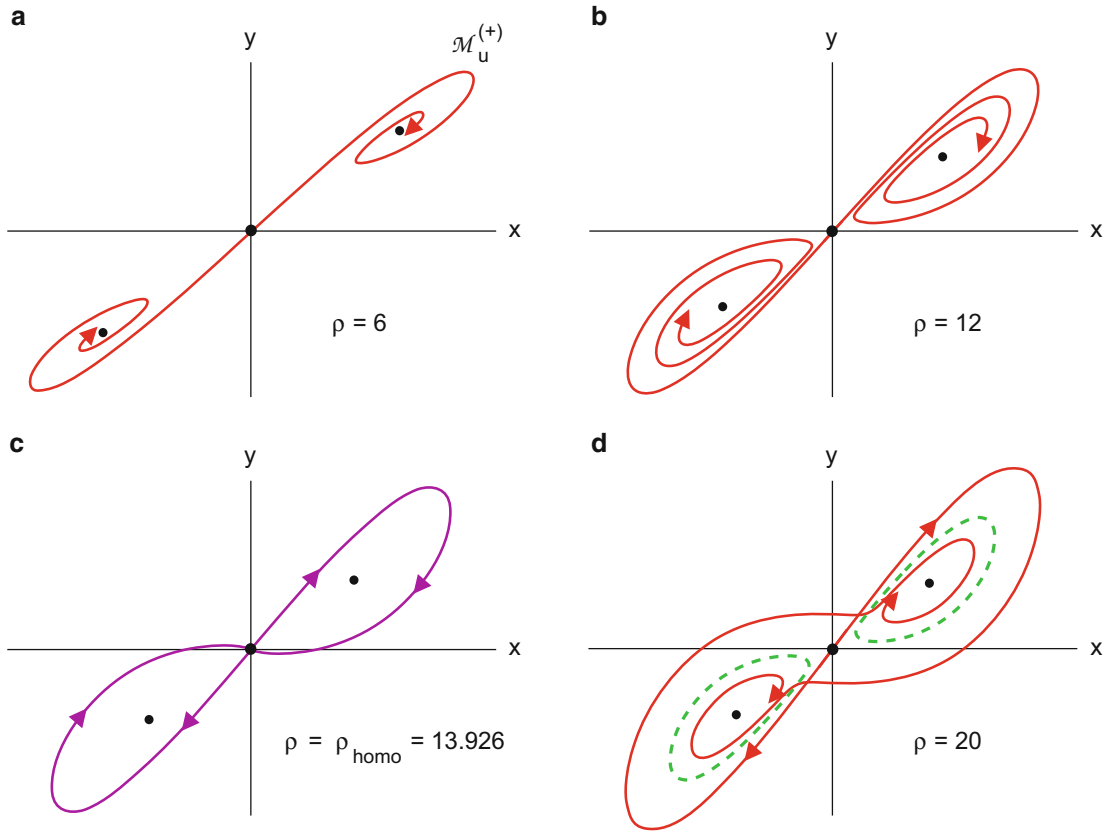

Figure 9.18: *Projections of unstable manifolds for the Lorenz system* (9.37). *Panels (a,b): when* $1 < \rho < \rho_{\text{homo}}$, $\mathcal{M}_u^{(+)}$ *connects to the stable equilibrium* $\mathbf{P}_+$. *(c) For* $\rho = \rho_{\text{homo}} \approx 13.926$, $\mathcal{M}_u^{(+)}$ *forms a homoclinic orbit. (d) For* $\rho = 20$, $\mathcal{M}_u^{(+)}$ *takes a longer excursion through phase space before connecting to* $\mathbf{P}_-$. *Note the unstable limit cycles in this range of* ρ.

The homoclinic bifurcation creates unstable periodic orbits Γ_+ and Γ_- whose projections lie in the first and third quadrants, respectively (cf. Figure 9.18(d)). The figure suggests that

$$\mathcal{M}_u^{(+)} \text{ converges to } \mathbf{P}_- \text{ "inside" } \Gamma_-, \tag{9.38}$$

but since the flow is three-dimensional, this statement needs to be clarified, a nontrivial but essential task. The linearized Poincaré map $\mathbf{D\Pi}$ of Γ_- has two eigenvalues, both real, one stable (magnitude < 1) and one unstable (magnitude > 1). Thus, the stable and unstable manifolds of Γ_-, let us call them $\mathcal{N}_s$ and $\mathcal{N}_u$, are two-dimensional, with one dimension corresponding to motion around Γ_- and the other to the stable and unstable eigenvectors of $\mathbf{D\Pi}$, respectively. Trajectories in one half of $\mathcal{N}_u$, let us call it $\mathcal{N}_u^{(\text{in})}$, spiral into $\mathbf{P}_-$ as $t \to \infty$, and in the opposite limit, they spiral outward to Γ_-. Thus, we may view $\mathcal{N}_u^{(\text{in})}$ as a sort of "membrane" stretched over Γ_- that contains the point $\mathbf{P}_-$. Since $\mathcal{N}_u^{(\text{in})}$ is an invariant manifold for (9.37), it forms an impenetrable barrier for the flow. We may now clarify (9.38): for ρ slightly beyond ρ_{homo}, the invariant manifold $\mathcal{M}_u^{(+)}$ from the origin bumps up against $\mathcal{N}_u^{(\text{in})}$ and then spirals into $\mathbf{P}_-$. This is shown in perspective in Figure 9.19(a).

As ρ is increased beyond ρ_{homo}, $\mathcal{M}_u^{(+)}$ bumps up against $\mathcal{N}_u^{(\text{in})}$ closer and closer to Γ_- before spiraling into $\mathbf{P}_-$. This trend culminates at $\rho = \rho_{\text{hetero}} \approx 24.06$, when $\mathcal{M}_u^{(+)}$ approaches the periodic orbit asymptotically.[12] This is the *third bifurcation*, the bifurcation that initiates chaos. If $\rho > \rho_{\text{hetero}}$, then $\mathcal{M}_u^{(+)}$ bumps up against $\mathcal{N}_u$ *outside* of Γ_- and, as indicated in Figure 9.19(d), moves farther away from $\mathbf{P}_-$ as it revolves around that point; after a few turns, it sails off toward $\mathbf{P}_+$. It then takes a few turns around $\mathbf{P}_+$ as it recedes from that point and flies back toward $\mathbf{P}_-$, where it approximately repeats its earlier behavior. This alternation between revolution around and recession from the two centers $\mathbf{P}_\pm$ continues indefinitely, in an aperiodic manner, as shown in Figure 9.17(a).

After ρ passes ρ_{hetero}, $\mathcal{M}_u^{(+)}$ and all nearby trajectories are chaotic. However, for $\rho_{\text{hetero}} < \rho < \rho_{\text{hopf}} \approx 24.74$, equation (9.37) is bistable: although many solutions are chaotic, solutions with initial conditions near $\mathbf{P}_\pm$ will still spiral inward to the (stable) equilibrium. Bistability ends when $\rho = \rho_{\text{hopf}}$: the equilibria lose stability through the *fourth bifurcation*, a subcritical Hopf bifurcation (cf. Exercise 21 in Chapter 8). This initiates a range of ρ in which virtually all trajectories are chaotic.

Even after ρ_{hopf}, equation (9.37) retains its ability to surprise. For still larger values of ρ the equation undergoes many more bifurcations, including multiple period-doubling cascades. Finally, it settles down when $\rho > 313$; in this range virtually all solutions have the same simple asymptotic behavior, a stable periodic orbit (see Chapter 7 of [75]).

[12] In other words, Γ_- is the ω-limit of points on $\mathcal{M}_u^{(+)}$. You might find it helpful to look at Exercise 5, which gives an analytically more tractable example of such behavior.

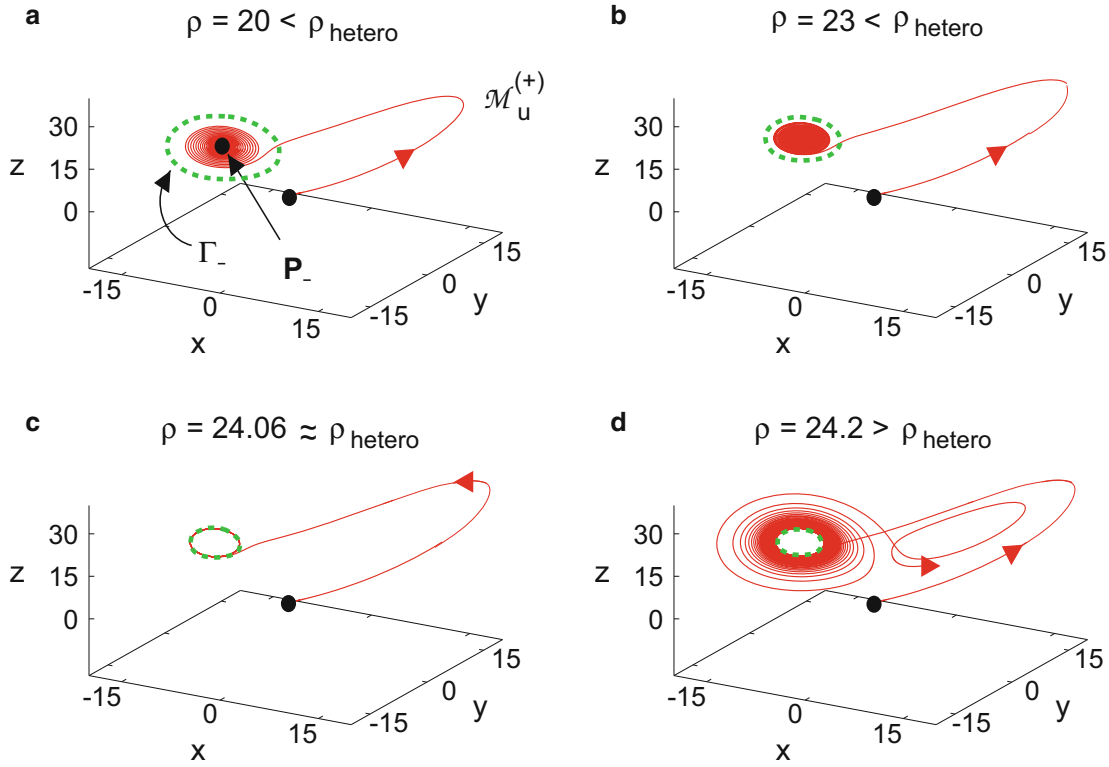

Figure 9.19: *Asymptotic behavior of $\mathcal{M}_u^{(+)}$ for various values of $\rho > \rho_{\text{homo}}$. In all panels, starting from the origin (the black dot), $\mathcal{M}_u^{(+)}$ makes a partial loop around $\mathbf{P}_+$ (equilibrium not shown) before entering the third quadrant. Panels (a,b): $\mathcal{M}_u^{(+)}$ bumps up against the unstable manifold $\mathcal{N}_u$ "inside" Γ_- and converges to $\mathbf{P}_-$ as $t \to \infty$. (c) $\mathcal{M}_u^{(+)}$ is asymptotic to the periodic orbit Γ_-; this is the onset of chaos. (d) $\mathcal{M}_u^{(+)}$ bumps up against $\mathcal{N}_u$ "outside" Γ_- and recedes from $\mathbf{P}_-$; after a few loops it gets shot back toward Γ_+, where it behaves similarly (cf. Figure 9.17(a)).*

9.8 Bursting in the Denatured Morris–Lecar Equations

Many neurons and other excitable cells exhibit a behavior known as *bursting oscillations*. This behavior is characterized by an alternation between intervals of rapid oscillations (in the transmembrane potential) with quiescent intervals, as illustrated in Figure 9.20(a).

Ermentrout and Terman [23] construct physiologically realistic ODE models for bursting based on the real Morris–Lecar equations.[13] Here, choosing simplicity over

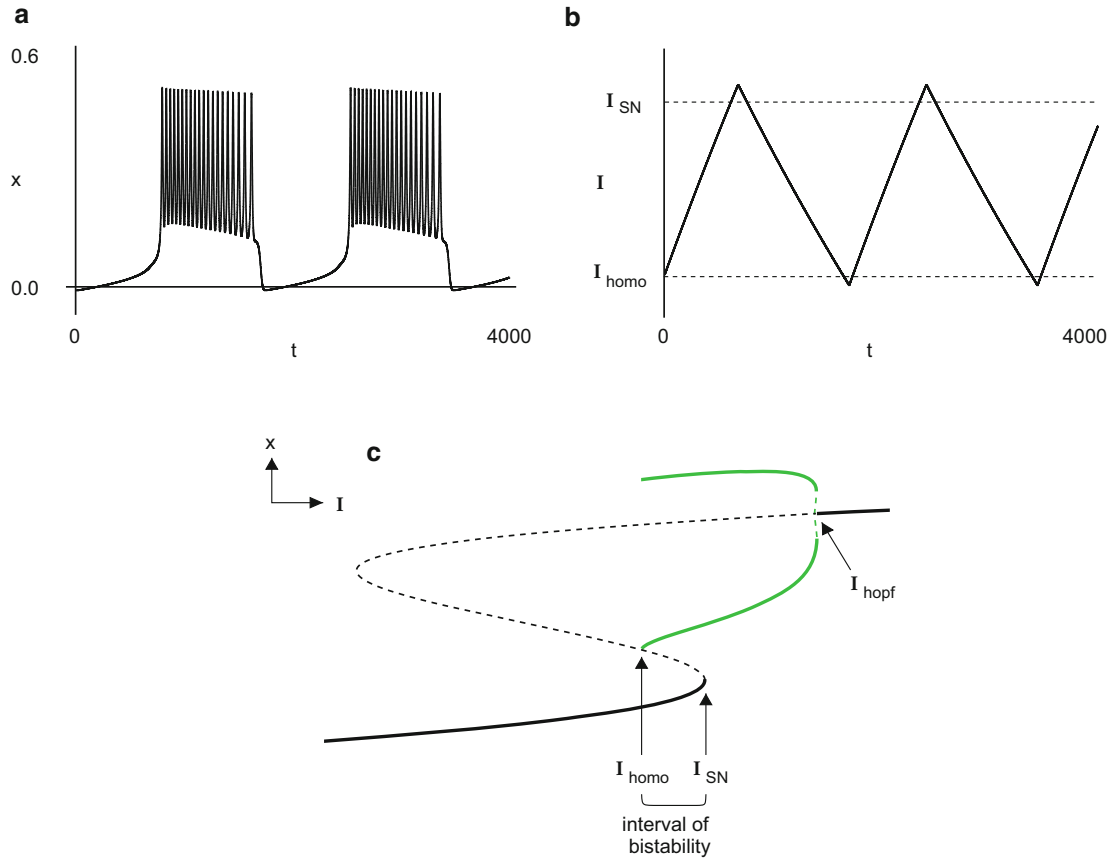

Figure 9.20: *(a) Bursting in the denatured Morris–Lecar equations* (9.39), (9.41) *with parameters* (9.40) *and* $\varepsilon = 0.0002$. *(b) A graph of* $I(t)$ *in the simulation of Panel (a). (c) Bifurcation diagram for* (9.39) *with parameters* (9.40). *On the horizontal axis,* I, *which is treated as a parameter, not a dynamic variable, ranges from* 0 *to* 0.022. *A homoclinic bifurcation occurs at* $I_{\text{homo}} \approx 0.01237$ *and a saddle-node bifurcation at* $I_{\text{SN}} \approx 0.01484$. *(The subcritical Hopf bifurcation, not related to bursting, is located at* $I_{\text{hopf}} \approx 0.01916$.*)*

[13] If you are interested in neuroscience, you should definitely consult their book and/or [96]. For example, there are different types of bursting, and we model only one of them.

realism, we construct such behavior based on the denatured Morris–Lecar equations,

$$\begin{aligned} x' &= x^2(1-x) - y + I, \\ y' &= Ae^{\alpha x} - \gamma y, \end{aligned} \tag{9.39}$$

but with slightly different parameters from those considered in Chapter 8, namely with

$$A = 0.0041, \quad \alpha = 5.27, \quad \gamma = 0.31. \tag{9.40}$$

A bifurcation diagram for (9.39) with these parameter values is shown in Figure 9.20(c). Although this figure appears rather similar to Figure 8.19 in Chapter 8, the two cases differ in how the periodic solutions terminate as I is decreased: in Figure 8.19, through an SN-limit-cycle bifurcation, but in Figure 9.20(c), through a homoclinic bifurcation.

The system (9.39) is bistable if $I_{\text{homo}} < I < I_{\text{SN}}$, where $I_{\text{homo}} \approx 0.01237$ and $I_{\text{SN}} \approx 0.01484$ are identified in Figure 9.20(c). Specifically, I_{homo} refers to the homoclinic bifurcation at which periodic solutions terminate, and I_{SN} refers to the saddle-node bifurcation at which the lower branch of (stable) equilibria terminates. In this range of I, both a periodic solution and an equilibrium are stable. This bistability can give rise to bursting if the current I is allowed to vary *slowly* with time, depending on the voltage x.

To elaborate, we refer to Figure 9.20(b), which shows the current $I(t)$ in the simulation represented in Figure 9.20(a). Because the current evolves slowly, the simulation may be analyzed approximately as a solution of (9.39) under quasistatic variation of I. At the starting time, $I(0) = I_{\text{homo}}$ and $x(0)$ equals the equilibrium on the lower branch for this value of I. As t increases, $I(t)$ slowly increases, and x will approximately track the bottom equilibrium in Figure 9.20(c) with $I = I(t)$, as long as $I(t) < I_{\text{SN}}$. This corresponds to the quiescent phase of bursting. However, when $I(t)$ crosses I_{SN}, the equilibrium disappears, so the system must jump to new behavior; specifically, it jumps to the stable periodic solution shown in the bifurcation diagram Figure 9.20(c). This corresponds to the rapidly oscillatory phase of bursting. Although I then starts to decrease, these oscillations will continue as long as $I(t) > I_{\text{homo}}$. When $I(t)$ finally decreases below I_{homo}, the system will jump back to the equilibrium, and the process starts over. In other words, bursting results from the system (9.39) alternating between resting and spiking as quasistatic changes in I move it around a hysteresis loop in Figure 9.20(c).

To obtain this behavior in an autonomous system of ODE, we add I as a third variable to (9.39) subject to the equation

$$dI/dt = \varepsilon(I_{\text{asym}}(x) - I), \tag{9.41}$$

where $\varepsilon = 0.0002$ and $I_{\text{asym}}(x)$ is a smoothed-out version of the step function

$$H(x) = \begin{cases} 1/30 & \text{if } x < v_{\text{switch}}, \\ 0 & \text{if } x > v_{\text{switch}} \end{cases} \qquad (v_{\text{switch}} = 0.05).$$

The evolution of I is slow, because ε is small. Parameters in the step function were chosen to achieve the following: if $x < v_{\text{switch}}$ (as it is on the lower branch of equilibria), then I tends to the asymptotic value $1/30 > I_{\text{SN}}$; and if $x > v_{\text{switch}}$ (as it is on the periodic solutions), then I tends to the asymptotic value $0 < I_{\text{homo}}$. In the simulation we used

$$I_{\text{asym}}(x) = \frac{1}{60}\left[1 + \tanh\left(\frac{.05 - x}{.001}\right)\right].$$

9.9 Exercises

After the core exercises there are sections on computations to support the text and on bifurcation in certain one-dimensional maps.

9.9.1 Core Exercises

The core exercises address the following issues:

Limit sets in global bifurcation	1
Calculation of eigenvalues of **DΠ**	2
Exploring various bifurcation phenomena	3–6

1. For each of the flows shown in Figures 9.2(c), 9.3(c), and 9.4(c), decide which of the following statements is true:

   (i) The purple orbit[14] is the α-limit of points inside it.
   (ii) The purple orbit is the ω-limit of points inside it.
   (iii) The purple orbit is neither an α- nor ω-limit of points inside it.
   (iv) The purple orbit is both an α- and an ω-limit of points inside it.

2. For the periodic solution of (9.11) with $\mu = 0$, apply Theorem 4.6.1 to show that $\lambda = 1$ is the eigenvalue of its Poincaré map.

[14]Strictly speaking, we should say the purple (homoclinic) orbit *and* the equilibrium, but please cut us some slack.

3. Using (9.11) as a guide, construct an academic example of an ODE with a pitchfork-like bifurcation of periodic solutions from a periodic solution. Do likewise for transcritical bifurcation.

   Hint: Compare the r-equation in (9.11) with the normal form for saddle-node bifurcation in Table 8.1.

   Remark: The first bifurcation in Figure 9.14 is a nonacademic example of a pitchfork-like bifurcation of periodic solutions from a periodic solution.

4. *Introduction:* As we saw in Section 7.4, the torqued pendulum (9.8) has a stable periodic solution if $\mu > 1$. In Section 9.2.3 we showed that for a representative *large* value of the friction coefficient β (which is called overdamped), the periodic orbit terminates through an SN-limit-cycle bifurcation at $\mu = 1$. However, if friction is small (underdamped), the periodic orbit continues to exist for a range of μ below 1 and terminates through a homoclinic bifurcation at $\mu = \mu_{\text{homo}} < 1$. In the range $\mu_{\text{homo}} < \mu < 1$, the motion is bistable: both the periodic orbit and one of the equilibria are stable. In the present exercise you explore this behavior.

   Figure 9.21 shows equilibria, stable and unstable manifolds, and periodic solutions of (9.8) *without our usual color conventions* for three values of μ, assuming $\beta = 0.2$.

   (a) Identify which equilibria are sinks, saddles, or sources.

   (b) Identify which trajectories are stable manifolds, unstable manifolds, homoclinic orbits, or periodic solutions.

   (c) Place arrowheads to indicate the direction of flow.

   (d) Draw a bifurcation diagram that summarizes this behavior.

   Challenge: Figure out what's special about the phase-plane plot of the torqued pendulum at the border between underdamped and overdamped behavior, at $\beta \approx 1.191$.

5. *Introduction:* This problem gives a simple example of a heteroclinic orbit that connects a saddle point to an unstable periodic orbit. This behavior in the example can be derived without numerics, in contrast to the bifurcation at $\rho = \rho_{\text{hetero}}$ in the Lorenz equation (9.37).

   (a) Verify that for every $\mu > 0$, the system given in cylindrical coordinates by

   $$\begin{aligned} r' &= (r^2 + 1 - 3z)(r^2 - 1 - 3z)r, \\ \theta' &= 1 - z(1 + r\sin\theta), \\ z' &= \mu z(z-1), \end{aligned} \tag{9.42}$$

   has a saddle-point equilibrium at $(r, \theta, z) = (\sqrt{2}, 0, 1)$ with a one-dimensional unstable manifold $\mathcal{M}_u$ and an unstable periodic orbit $\{r = 1,\ z = 0\}$.

   (b) Show that there is a value of $\mu > 0$ such that $\mathcal{M}_u$ is asymptotic to the periodic orbit as $t \to \infty$.

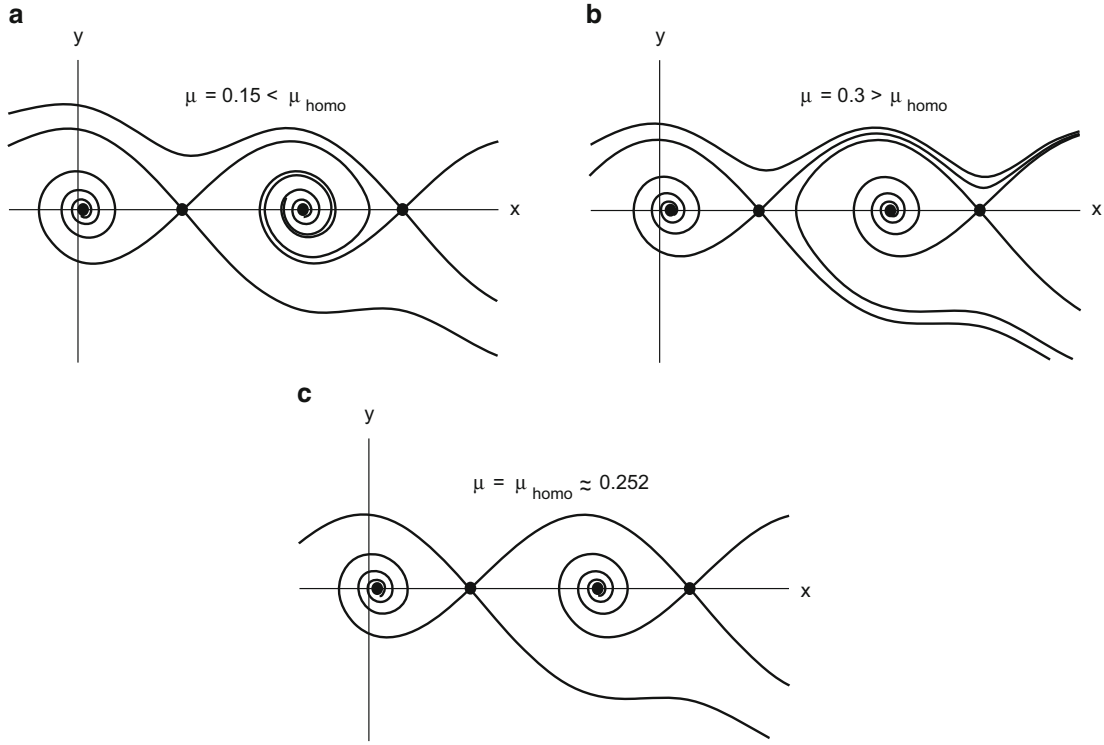

Figure 9.21: *Stable/unstable manifolds and periodic orbits for the underdamped torqued pendulum equations with* $\beta = 0.2$. *In Exercise 4 we ask you to supply the appropriate colors and draw the bifurcation diagram.*

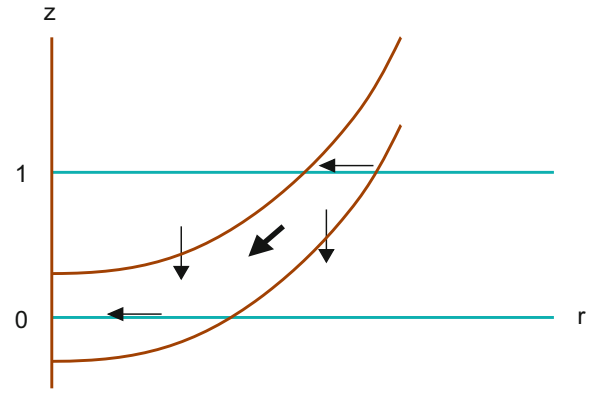

Figure 9.22: *Part of the flow-quadrant diagram for the r, z-subsystem of* (9.42).

Hint: First consider the r, z-subsystem. Calculate its nullclines, which do not depend on μ, and verify the flow-quadrant diagram of Figure 9.22. In particular, show that the subsystem has saddle points at $(\sqrt{2}, 1)$ and $(1, 0)$, and show that in the quadrilateral bounded by the r-nullclines $z = (r^2 \pm 1)/3$ and the z-nullclines $z = 0$ and $z = 1$, flow is directed into the third quadrant.

Now ask how the unstable manifold of the r, z-subsystem through $(\sqrt{2}, 1)$, which we also designate by $\mathcal{M}_u$, varies with μ. Argue that if μ is small, $\mathcal{M}_u$ converges to the asymptotically stable equilibrium at the origin. Then argue that if μ is large, $\mathcal{M}_u$ is swept out to infinity. Conclude by continuity that for an intermediate value, say $\mu = \mu_*$, $\mathcal{M}_u$ converges to the saddle point at $(1, 0)$. Finally, put θ back into the problem and argue that when $\mu = \mu_*$, the unstable manifold $\mathcal{M}_u$ for the full problem is asymptotic to the periodic orbit.

6. *Introduction:* Rewrite (7.106), the equation for a vertically vibrated pendulum, as a first-order autonomous system

$$\begin{aligned} x' &= y, \\ y' &= -\beta y + (1 - \alpha\omega^2 \cos z) \sin x, \\ z' &= \omega, \end{aligned} \tag{9.43}$$

where we reduce z modulo 2π. For all ω, this equation has the periodic solution

$$x(t) \equiv y(t) \equiv 0, \quad z(t) = t. \tag{9.44}$$

According to Section 7.10, provided α is not too large, this solution is stable if $\omega > \sqrt{2}/\alpha$.

(a) Just on the basis of readily observed properties, try to figure out what kind of bifurcation accompanies this loss of stability:

(i) Homoclinic or heteroclinic bifurcation

(ii) SN-limit-cycle bifurcation

(iii) Supercritical pitchfork-like bifurcation ($\lambda(\mathbf{D\Pi}_*) = +1$)

(iv) Subcritical pitchfork-like bifurcation ($\lambda(\mathbf{D\Pi}_*) = +1$)

(v) Transcritical-like bifurcation ($\lambda(\mathbf{D\Pi}_*) = +1$)

(vi) Mutual annihilation bifurcation ($\lambda(\mathbf{D\Pi}_*) = +1$)

(vii) Bifurcation to an invariant torus ($\lambda(\mathbf{D\Pi}_*) = e^{\pm i\omega}$)

(viii) Period-doubling bifurcation ($\lambda(\mathbf{D\Pi}_*) = -1$)

Hint: Two readily observed properties are that the periodic solution continues to exist after it loses stability and (9.43) has the reflectional symmetry $(x, y, z) \mapsto (-x, -y, z)$.

(b) What new solutions, if any, appear at the bifurcation?

9.9.2 Computations to Support Claims in the Text

7. *Introduction:* Using Exercise 7.10 as a guide, define a Poincaré map $\mathbf{\Pi} : \mathbb{R}^2 \to \mathbb{R}^2$ whose fixed points locate $2\pi/\omega$-periodic solutions of (9.18).

   (a) Compute the fixed point of $\mathbf{\Pi}$ for $\Gamma = 1.3$.

   (b) Decreasing Γ slowly, track the fixed point as a function of Γ and verify that the eigenvalues of $\mathbf{D\Pi}$ cross the unit circle when $\Gamma \approx 1.285$.

   Hint: Although your software package may compute eigenvalues of the Poincaré map automatically, it is useful to realize how simple it is to compute these from scratch. Suppose $\boldsymbol{\gamma}(t)$ is a T-periodic solution of a d-dimensional autonomous ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$. As in Section 4.8, consider the IVP

$$\begin{array}{rcll} \mathbf{x}' & = & \mathbf{F}(\mathbf{x}), & \mathbf{x}(0) = \boldsymbol{\gamma}(0), \\ X' & = & \mathbf{DF}(\mathbf{x})X, & X(0) = I, \end{array} \tag{9.45}$$

   with unknowns the d-dimensional solution vector $\mathbf{x}(t)$ and a $d \times d$ matrix $X(t)$. Then $X(T) = \mathbf{D}\boldsymbol{\varphi}(T, \boldsymbol{\gamma}(0))$, and Proposition 7.3.5 can be invoked to extract the eigenvalues of $\mathbf{D\Pi}(\boldsymbol{\gamma}(0))$.

8. Verify numerically that (9.20) has periodic solutions that as μ varies, bifurcate to an invariant torus.

   Discussion: If you want a suggestion, here are some possible parameter values:

$$\gamma = 1/4, \quad \nu = 1, \quad \sigma = 3, \quad E = 1.$$

   If $\gamma = 0$, then (9.20) undergoes a Hopf bifurcation at $\mu = 1/6$. Thus for small γ, you would expect the bifurcation to an invariant torus to occur for μ not too far from $1/6$.

9. Make one or more perturbations of (9.27) and check numerically that the perturbed equations still exhibit period-doubling.

10. *Introduction:* Here is a simplified model of the cardiac action potential [56]:

$$\begin{aligned} \frac{dv}{dt} &= \frac{h}{\tau_{\text{in}}} v^2(1-v) - \frac{v}{\tau_{\text{out}}} + J_{\text{stim}}(t), \\ \frac{dh}{dt} &= \begin{cases} (1-h)/\tau_{\text{open}}, & \text{if } v < v_{\text{crit}}, \\ -h/\tau_{\text{close}}, & \text{if } v \geq v_{\text{crit}}, \end{cases} \end{aligned} \tag{9.46}$$

where v represents voltage across a cell membrane and h is a "gate" variable; i.e., by "opening" and "closing," depending on whether the voltage exceeds a threshold v_{crit}, it regulates the inward current in the voltage equation. The variables v and h are scaled to vary between 0 and 1. In solving (9.46), use the following choices for parameters:

$$\begin{aligned} \tau_{\text{in}} &= 0.1 \text{ ms}, \quad \tau_{\text{out}} = 2.4 \text{ ms}, \quad v_{\text{crit}} = 0.13, \\ \tau_{\text{open}} &= 30 \text{ ms}, \quad \tau_{\text{close}} = 150 \text{ ms}. \end{aligned}$$

You may regard $J_{\text{stim}}(t)$ as a periodic *impulsive*[15] stimulus; i.e., every B milliseconds, the stimulus instantaneously raises v by some fixed amount, say 0.25. (Thus, these jumps make the voltage discontinuous; the ODE holds on the intervals between stimuli.)

(a) Using numerical simulations with $B = 400$ ms, $v(0) = 0.5$, and $h(0) = 0.9$, solve the equations for a few periods of the stimulus until v and h are periodic with period B; plot this steady response v vs. t.

(b) Repeat the computation under quasi-static reduction of B until you observe period-doubling; i.e., the period of the steady response v vs. t becomes $2B$.

Discussion: For $B = 400$ ms, you will find that the steady response v vs. t has one shark-fin-shaped elevation of v per period of the stimulus. We may define the APD as the time that v exceeds v_{crit}. Before period-doubling, all APD's are identical; after, they have a long–short alternating pattern. The associated cardiac rhythm, called T-wave alternans, is considered abnormal and potentially dangerous.

11. If $a = b = 0.2$, the Rössler equation (9.35) has a unique periodic solution for $0.4 < c < 2.832$. Compute the eigenvalues of the Poincaré map of this periodic solution for a range of c below 2.832 and show that one of them tends to -1 as $c \to 2.832$.

12. Show numerically that (9.39) with parameters (9.40) undergoes a homoclinic bifurcation at $I \approx 0.01237$.

[15] If it bothers you that the periodic forcing term $J_{\text{stim}}(t)$ is impulsive or that the right-hand side of the dh/dt equation has jumps, feel free to employ smooth alternatives. If you do so, be sure that over each period, $J_{\text{stim}}(t)$ is nonzero only briefly, perhaps on the order of one to three milliseconds, but integrates to something on the order of say 0.5 (in order to provide a sufficient kick to v).

9.9.3 Bifurcation in a Quadratic Map

13. Use the computer to investigate for yourself the remarkable sequence of period-doubling bifurcations in the logistic map

$$\Psi(x) = \mu x(1-x), \tag{9.47}$$

which for $\mu \leq 4$ maps the unit interval $[0,1]$ into itself.

Discussion: You can read about this behavior in any number of references (e.g., Chapter 10 of [81]), but it's more fun to discover it yourself. You can get started analytically. If $\mu > 1$, then $x = 1 - 1/\mu$ is a nonzero fixed point of Ψ. This fixed point is stable if $1 < \mu < 3$, but it loses stability at $\mu = 3$ through a supercritical period-doubling bifurcation. You can find analytically the 2-cycle that bifurcates at $\mu = 3$ and show that it is stable for $3 < \mu < 1 + \sqrt{6}$. At this point, you probably want to turn the job over to the computer. Here's a specific question to address: if the nth period-doubling bifurcation occurs at μ_n, estimate the limit μ_∞ of the sequence μ_n.

After doing this exercise, you might find it stimulating to read about the universal behavior captured by Feigenbaum's renormalization theory, which is covered in Section 10.7 of [81].

14. *Introduction:* For $\mu_\infty < \mu \leq 4$, where μ_∞ is defined in (the discussion following) the previous exercise, the orbits of the logistic map (9.47) are mostly aperiodic, but with some exceptions. Most notably, there is a range of μ in which Ψ has stable period-3 orbits. Although this behavior was discovered using the computer, you can show that such orbits exist using just a *calculator*! With $\mu = 3.835$, let

$$\alpha_1 = 0.152074, \qquad \alpha_2 = 0.494514, \qquad \alpha_3 = 0.958635.$$

You may calculate that $\Psi(\alpha_k) \approx \alpha_{k+1}$, where for $k = 3$ we define $\alpha_4 = \alpha_1$.

Evaluate $\Psi'(\alpha_k)$ and use this information to construct short closed intervals $\mathcal{I}_k$, $k = 1, 2, 3$, around each point α_k such that $\Psi(\mathcal{I}_k) \subset \text{Int } \mathcal{I}_{k+1}$.

Discussion: It follows that the sequence $\Psi^{3n}(\alpha_1)$ contains a subsequence that converges to a point in $\mathcal{I}_1$ that has period 3 under iteration by Ψ. You could calculate that

$$\Psi'(\alpha_1)\Psi'(\alpha_2)\Psi'(\alpha_3) \approx -0.394972,$$

and with a little more work you could show that for some $\varepsilon > 0$,

$$|\Psi'(x_1)\Psi'(x_2)\Psi'(x_3)| < 1 - \varepsilon \tag{9.48}$$

for all x_k such that $x_k \in \mathcal{I}_k$. From this it would follow that the period-3 trajectory you found above is unique and stable.

However, far more rewarding than verifying (9.48) is to read about Sarkovskii's theorem, which identifies truly remarkable behavior that is implied by the existence of such a period-3 orbit; see, for example, Section 1.10 of [17].

9.9.4 PHD Exercises

15. *Introduction:* In this exercise we seek intuition regarding the bifurcations in (9.31). Eigenvalues of the Poincaré map for a periodic solution $\gamma(t)$ of (9.31) may be calculated by solving the linearized equation

$$w'' + \beta w' + [\cos\gamma(t)]\, w = 0. \tag{9.49}$$

Consider only small Γ in (9.31) with parameter values (9.32). Using (9.34) and making the approximation $\cos\gamma \approx 1 - \gamma^2/2$, (9.49) reduces to

$$w'' + w'/2 + [(1 - A^2/4) + (A^2/4)\cos 2t]\, w = 0, \tag{9.50}$$

where we have used the identity $\sin^2 t = (1-\cos 2t)/2$. As A increases, there are two possible mechanisms for instability in this equation: (i) the coefficient $1 - A^2/4$ could become too negative for the periodic forcing to counteract, as studied in (7.99), or (ii) periodic forcing, even though detuned from the natural frequency in (9.50), could cause exponential growth, as described for (7.83) in Exercise 7.16. It is not obvious which effect comes into play first, but we may deduce from our computations in Section 9.6.4 that mechanism (i) causes the first instability.[16] Let's try to locate where this bifurcation occurs.

Use Proposition 7.10.3 to estimate that (9.50) acquires solutions with exponential growth when $A \approx 2.16$.

Hint: Change variables in (9.50) to $\tau = 2t$ to obtain

$$\frac{d^2w}{d\tau^2} + \frac{1}{4}\frac{dw}{d\tau} + \left[\frac{1}{4}\left(1 - \frac{A^2}{4}\right) + \frac{A^2}{16}\cos\tau\right] w = 0,$$

whose trigonometric term has the same frequency as (7.100). Ignoring friction for the moment, apply Proposition 7.10.3 with $\varepsilon = A^2/32$ to derive the condition

$$\frac{1}{4}\left(1 - \frac{A^2}{4}\right) = -2\left(\frac{A^2}{32}\right)^2$$

for the onset of solutions with exponential growth, and solve this equation for A. A posteriori, check that the value of β in (7.100) implicit in making this estimate, i.e., defined by $\beta(A^2/32) = 1/4$, is in an acceptable $\mathcal{O}(1)$ range.

Remark: Comparing this result with Figure 9.14 and recalling how crude the above approximations were, we can only say, "not bad."

16. *Introduction:* Although the periodic solution of (9.31) that is invariant under (9.33) loses stability as Γ increases, it still continues to exist; of course, it can't be computed simply by

[16] How can we deduce this? In the first place, note that $A \approx 2.4$ in Figure 9.14(a), and at the bifurcation, it will be slightly larger, i.e., large enough to make $1 - A^2/4$ negative. Moreover, reflecting on the discussion following Exercise 7.16, we can see that for mechanism (i), the destabilizing eigenvalue of the Poincaré map would be $+1$, while for mechanism (ii), it would be -1; since the period is *not* doubled at the first bifurcation, we conclude that mechanism (i) is the driver. Period-doubling bifurcations of course do occur later, but they are less amenable to analysis; we do not attempt to locate them, even approximately.

letting $t \to \infty$ in a simulation. However, feedback can be used to compute such an unstable solution. Specifically, consider the equation[17]

$$x'' + \beta x' + \sin x = \Gamma \cos \omega t - C[x(t) + x(t - \pi/\omega)], \tag{9.51}$$

where $C > 0$. The new term in (9.51) may be interpreted as a fictitious force that pushes $x(t)$ toward $-x(t - \pi/\omega)$. For functions that are invariant under (9.33), this term vanishes identically; thus, a solution of (9.51) that is invariant under (9.33) also satisfies (9.31).

Choosing arbitrary initial data for $0 \le t \le \pi/\omega$, solve (9.51), say with $\Gamma = 2.2$ and the parameters (9.32). Experiment until you find a value of the coefficient C such that your solution is invariant under (9.33).

Remark: Incidentally, after the period-doubling bifurcation, the unstable solutions of (9.31) with the undoubled period $2\pi/\omega$ can be found by solving an analogue of (9.51) with a fictitious force

$$-C[x(t) - x(t - 2\pi/\omega)].$$

9.10 Pearls of Wisdom

9.10.1 Remarks on Heteroclinic Orbits

The term "heteroclinic" is used to describe an orbit of an ODE that connects two equilibria. Heteroclinic orbits are important when both equilibria are saddle points: thinking for the moment in two dimensions, such an orbit is simultaneously the unstable manifold of one saddle point and the stable manifold of the other. For example, reflecting this situation, the heteroclinic orbit in Figure 9.2(c) is shown in purple, a combination of red for unstable and blue for stable. As is typical, the connection occurs only for one specific value of the bifurcation parameter, and this value separates parameter regimes with different phase portraits.

By contrast, an orbit that connects a saddle point to a sink has little significance. For example, two orbits that connect equilibria in Figure 9.4(a) are shown in red as they are the unstable manifold of the saddle point $(r, \theta) = (1, +\arccos(1 + \mu))$, but they have no special status as regards their other endpoint: they are only two of infinitely many orbits approach the sink $(1, -\arccos(1 + \mu))$ asymptotically. Moreover, this connection exists for a range of the bifurcation parameter, with no change in phase portraits involved.

In higher dimensions, an orbit that connects two saddle points, say $\mathbf{P}_1$ and $\mathbf{P}_2$, has significance for bifurcation if the dimension of the unstable manifold of $\mathbf{P}_1$ plus the

[17]Note that this is an ODE *with delay*, a class of equations that is discussed in Section 10.5. Such equations exhibit some new phenomena; in particular, initial data must be specified along an entire time interval of length π/ω. However, given such initial data, equations with delay can still be solved numerically using software such as XPPAUT.

dimension of the stable manifold of $\mathbf{P}_2$ equals exactly the dimension of the ambient space. (Incidentally, the analogue of this condition for a homoclinic orbit is always satisfied: the dimensions of the unstable and stable manifolds of a hyperbolic saddle point add up to the dimension of the ambient space.)

As Exercise 5 demonstrates, heteroclinic orbits can connect other invariant sets besides equilibria, such as a periodic orbit. In the Lorenz equations (9.37), a more interesting example, if $\rho = \rho_{\text{hetero}}$, then the one-dimensional unstable manifold $\mathcal{M}_u^{(+)}$ is asymptotic to the periodic orbit Γ_- as $t \to \infty$. (Cf. Figure 9.19(c).) And as we saw above, a bifurcation occurs at this value of ρ.

9.10.2 Bifurcation in Fluid-Mechanics Problems

As we mentioned in Section 9.5.3, bifurcation to an invariant torus occurs in many fluid-mechanics problems,[18] even without periodically varying external driving. A representative example is provided by *Taylor–Couette flow*. This term refers to fluid flow in the annular region between long concentric rotating cylinders, where a rotation speed, say of the inner cylinder, is taken as the bifurcation parameter. For small rotation rates, the flow is purely circumferential and is essentially independent of z (the coordinate along the axis). Multiple bifurcations occur as the speed is increased:

- first, to steady flow with circulation in cells, called *Taylor vortices*;
- next, to a periodic flow, called *wavy vortices*;
- then, to an aperiodic flow called *modulated wavy vortices*.

The third bifurcation is a bifurcation to an invariant torus. Look online to learn more about these phenomena.

In fact, more instabilities continue to appear as the speed of rotation is increased further, contributing to the transition to turbulent flow. Before the 1970s, it was believed that fully turbulent flow ensued only after an infinite number of bifurcations to increasingly complex flow had occurred. However, this view was supplanted by the revolutionary proposal of Ruelle and Takens [69] that turbulent flow arrived abruptly, after just a few such bifurcations, when the differential equations admitted chaotic solutions. This story makes fascinating reading; see, for example, [30].

9.10.3 Routes to Chaos

Although we don't deal with chaos in a serious way, it has nevertheless cropped up in our work. Let us note three different routes through which chaos appeared:

[18] Although the description of fluid motion requires PDEs—the Navier–Stokes equation—in fact the analysis of bifurcation in many PDEs is closely analogous to the analysis of bifurcation in ODEs.

- A period-doubling cascade, in Rössler's equation, Section 9.6.
- An abrupt global bifurcation, in the Lorenz equation, Section 9.7.
- Repeated Hopf-like bifurcations, in fluid mechanics, Section 9.10.2.

One further route that can be understood with the concepts from this book deserves to be mentioned: a *homoclinic tangle*. The homoclinic bifurcations of Section 9.1 provide a useful point of reference. In these bifurcations, the stable and unstable manifolds of a saddle point intersect in a complete orbit that approaches the saddle point asymptotically in both limits $t \to \pm\infty$. In a homoclinic tangle, the stable and unstable manifolds of a *periodic solution* intersect transversely in an orbit that is asymptotic to the periodic solution in both limits $t \to \pm\infty$. Equivalently, in terms of the Poincaré map $\mathbf{\Pi}$ of the periodic orbit, the stable and unstable manifolds of the associated fixed point of $\mathbf{\Pi}$ intersect transversally.[19] Such an intersection, which is robust under perturbation, entails geometry so complicated as to defy the imagination; for example, consult Sections 5.1, and 5.2 in Part 3 of [1] or Section 15.2 of [39], especially Figure 15.2.

In many references, the existence of a homoclinic tangle in the periodically forced Duffing equation,

$$x'' + \beta x' - x + x^3 = \Gamma \cos \omega t,$$

is used to find chaotic solutions; the derivation in Section 15.3.1 of [39] is among the more readable.

[19]Note the contrast: the stable and unstable manifolds of an equilibrium of an ODE cannot intersect transversely.

Chapter 10

Epilogue

In this epilogue, we offer brief overviews of several ODE topics not covered in the main body of the text.

10.1 Boundary Value Problems

10.1.1 An Overview Through Examples

Early on, we saw that ODEs have many solutions. Mostly in this book we have used initial conditions to select one specific solution from the many. However, other means of specifying a unique solution also arise in practice and need to be studied. In these problems, typically the independent variable is a spatial coordinate—*not time*—say ranging over an interval $0 \le x \le \ell$. A specific solution is determined by requiring it to satisfy *boundary conditions* involving information from *both* ends of the interval.

Let us illustrate possible boundary value problems (BVPs) in a simple case, say for the linear second-order scalar ODE

$$\frac{d^2y}{dx^2} - y = c, \qquad 0 \le x \le \ell, \tag{10.1}$$

where c is a constant. For any choices of the coefficients a_k, b_k, the relations

$$a_0y'(0) + a_1y(0) = a_2, \qquad b_0y'(\ell) + b_1y(\ell) = b_2, \tag{10.2}$$

are appropriate candidates for boundary conditions, provided we avoid trivialities by requiring that a_0 and a_1 not both be zero, and likewise for b_0, b_1. Thus, in contrast to the IVP, there are many different viable BVPs based on a single ODE.

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8_10

Here is a more profound difference between IVPs and BVPs: even for a perfectly benign ODEs like (10.1), some boundary conditions can yield a problem with no solution, and *it may not be feasible to determine whether such behavior occurs in a specific problem.* For an example in which this determination is possible, consider (10.1) subject to

$$y(0) = a, \qquad y'(\ell) - 2y(\ell) = b. \tag{10.3}$$

The general solution of (10.1) is

$$y(x) = -c + C_1 e^x + C_2 e^{-x},$$

and the boundary conditions require that

$$-c + C_1 + C_2 = a, \qquad 2c + (1-2)e^{\ell} C_1 + (-1-2)e^{-\ell} C_2 = b. \tag{10.4}$$

For every $\ell \neq (\ln 3)/2$, the coefficient matrix in this 2×2 linear system is invertible, so (10.4) has a unique solution. However, for the exceptional value of ℓ, this matrix is singular. In this case, if

$$2c - b = \sqrt{3}(c + a), \tag{10.5}$$

then the BVP (10.1), (10.2) has solutions, but they are not unique, and if (10.5) is not satisfied, the BVP does not have any solutions.

BVPs for *nonlinear* ODEs also occur. For example, the following BVP arises in a calculus-of-variations[1] problem:

$$yy'' - (y')^2 = 1, \qquad 0 < x < \ell, \tag{10.6}$$

subject to

$$y(0) = a, \qquad y(\ell) = b, \tag{10.7}$$

where a and b are positive constants. This problem relates to the shape assumed by a soap film stretched over circular rings of radius a and b, as sketched in Figure 10.1. The unknown $y(x)$ specifies the radius of a rotationally symmetric surface as a function of position along the axis. One seeks a function $y(x) > 0$, subject to the boundary conditions (10.7), to minimize the integral

$$\int_0^{\ell} y(x)\sqrt{1 + (y')^2}\, dx, \tag{10.8}$$

which equals ($1/2\pi$ times) the area of such a surface of revolution. The BVP poses a necessary condition for $y(x)$ to minimize the integral.

Reflecting its nonlinearity, this BVP exhibits new behavior. For simplicity, let's suppose that in the boundary conditions, $b = a$. We claim that (10.6), (10.7) has *two* solutions iff $0 < \ell < \ell_0 \approx 1.3255a$; these two solutions merge for $\ell = \ell_0$, and for

[1] See Chapter 3 of [98] for a brief introduction to this subject or [93] for more thorough treatment.

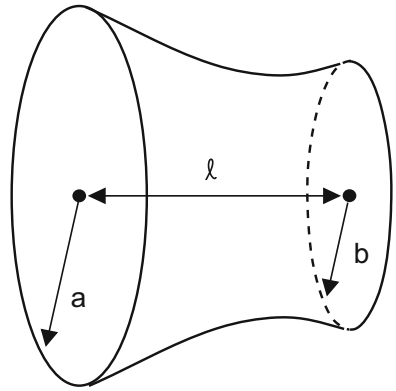

Figure 10.1: *Physical problem leading to BVP* (10.6)–(10.7)*: Find a rotationally symmetric minimal surface stretched over two rings, of radii* a *and* b*, separated by a distance* ℓ*.*

$\ell > \ell_0$, the BVP has *no* solutions.[2] (In Exercise 2 we guide you through a proof of this claim.) Two solutions of (10.6), (10.7), assuming $\ell < \ell_0$, are sketched in Figure 10.2. Physically, the solution in Panel (a) describes a stable shape that a soap film may assume in this geometry. The other solution, which is unstable, cannot be realized in experiments. If starting from the stable solution in Panel (a), ℓ is increased beyond the critical value ℓ_0, the soap film ruptures and contracts to two disconnected disks bounded by the circles.

There are many variants of boundary conditions beyond the simple cases in (10.2). For example,

- BVPs may be formulated for higher-order ODEs, of course with more boundary conditions imposed. For example, small lateral deflections of a uniform beam[3] (cf. Figure 10.3) may be described, after some scaling, by the ODE

$$\frac{d^4y}{dx^4} = -\rho, \qquad 0 \le x \le \ell, \tag{10.9}$$

  where ρ is proportional to its density (mass per unit *length*). One of many possible sets of boundary conditions for (10.9) is

$$y(0) = y''(0) = 0, \qquad y(\ell) = y''(\ell) = 0.$$

[2]Using the terminology of Chapter 8, we may say that the BVP undergoes a saddle-node bifurcation as ℓ passes through ℓ_0.

[3]Bending involves highly nonlinear effects, but under certain simplifying assumptions, including that deflections are small, the equations reduce to this linear problem; cf. Section 9.4 of [7].

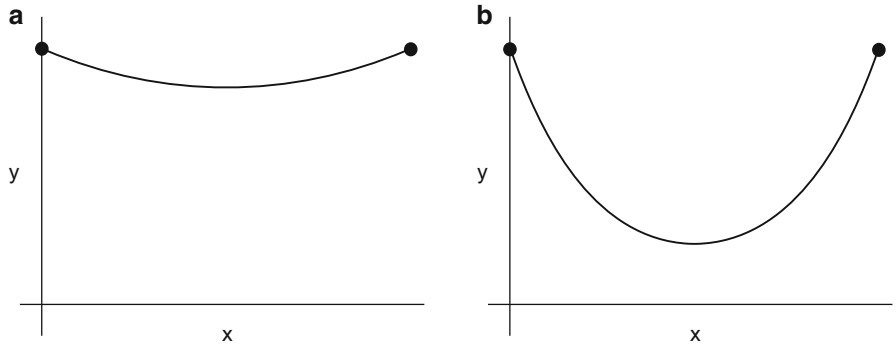

Figure 10.2: *Two catenaries representing different solutions of BVP* (10.6)–(10.7)*, where we have assumed* $a = b = \ell$*. The solution in Panel (a) can be observed in experiments with soap films. By contrast, the solution in Panel (b) has only mathematical significance. It represents a critical point of the area integral* (10.8) *that is not a local minimum.*

These relations, called *simply supported boundary conditions*, are illustrated in the figure. The ends of the beam are held at a fixed height, and the beam is free to rotate about the support points.

- A single boundary condition may involve the solution at both ends of the interval. The most common case of this is *periodic boundary conditions*; for a second-order ODE, these are

$$y(\ell) = y(0), \qquad y'(\ell) = y'(0).$$

(See Exercise 4.)

- An ODE may be posed on an infinite interval with constraints on the behavior of the solution at infinity playing the role of a boundary condition.[4] For example, the unique solution of (10.1) on $0 \le x < \infty$ subject to

$$y(0) = a, \quad y(x) \text{ bounded as } x \to \infty,$$

is $y(x) = (a + c)e^{-x} - c$.

- Sometimes, a specific solution of an ODE may be selected based in part on an integral of the solution over an entire interval $0 \le x \le \ell$, not just values

[4]Looking for points on the stable manifold of a saddle point of a nonlinear system of ODEs may be viewed as imposing a boundary condition at $t = \infty$, albeit not one that picks out a unique solution.

Figure 10.3: *Physical problem leading to ODE* (10.9)*: a beam bending under its own weight.*

at the endpoints. For example, consider a uniform flexible chain[5] of length L, as in Figure 10.4, suspended at its ends, hanging under its own weight. If the tension in the chain is eliminated, the shape $y(x)$ of the chain satisfies the ODE (cf. Section 10.3 of [28])

$$\frac{d}{dx}\left\{\frac{y''}{\sqrt{1+(y')^2}}\right\} = 0. \tag{10.10}$$

Two obvious boundary conditions come from the supports at the end of the chain, say

$$y(0) = a, \qquad y(\ell) = b. \tag{10.11}$$

Since (10.10) is of third order, to complete the formulation of the problem we need a third auxiliary condition, which is provided by the length constraint

$$\int_0^\ell \sqrt{1+(y'(x))^2}\,dx = L. \tag{10.12}$$

In Exercise 5 we lead you through the solution of (10.10) subject to these conditions in case $b = a$.

10.1.2 Eigenvalue Problems

As we saw above, both existence and uniqueness fail for the BVP (10.1), (10.3) in the one case that $\ell = (\ln 3)/2$. One can dispel some of the mystery of such seemingly random misbehavior by studying the eigenvalue problem. In the eigenvalue problem derived from (10.1), (10.3), one asks for what values of the parameter λ does there exist a nonzero solution of the BVP on the interval $0 \le x \le \ell$,

$$y'' - y = \lambda y; \qquad y(0) = 0, \quad y'(\ell) - 2y(\ell) = 0. \tag{10.13}$$

[5]This is the traditional application; to be more current, you might prefer to think of a long-distance power cable.

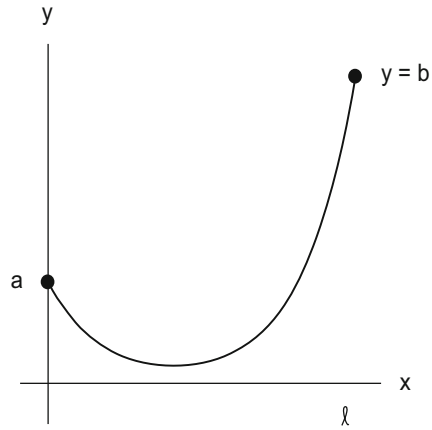

Figure 10.4: *Physical problem leading to BVP* (10.10), (10.11), (10.12)*: a hanging chain.*

This equation is analogous to the equation $A\mathbf{v} = \lambda\mathbf{v}$ for eigenvalues and eigenvectors of a $d \times d$ matrix A. The matrix defines a linear transformation on $\mathbb{R}^d$. Pursuing the analogy, we use the ODE in (10.13) to define a linear operator

$$L[y] = y'' - y$$

on an infinite-dimensional space of functions that incorporates the *homogeneous* boundary conditions[6]

$$\mathsf{X} = \{y \in \mathcal{C}^2([0,\ell]) : y(0) = 0,\ y'(\ell) - 2y(\ell) = 0\}.$$

Then, we may rewrite (10.13) compactly as $L[y] = \lambda y$. Making the obvious definitions, we call a *nonzero* solution of (10.13) an eigenfunction of L, and λ an eigenvalue.

Eigenvalue problems may be formulated for many linear differential operators. For example, restricting our attention to second-order operators (with variable coefficients)

$$L[y] = c_0(x)y'' + c_1(x)y' + c_2(x)y,$$

we append (homogeneous) boundary conditions to obtain

$$L[y] = \lambda y; \qquad a_0y'(0) + a_1y(0) = 0, \quad b_0y'(\ell) + b_1y(\ell) = 0, \tag{10.14}$$

where a_0 and a_1 are not both zero and likewise for b_0 and b_1. Such eigenvalue problems have been thoroughly studied, because they arise from solving PDEs with the technique known as *separation of variables*. (This method is explained, for example, in Section 4.1 of [80] and Chapter 5 of [10].) Typically, (10.14) will have infinitely

[6]For L to be linear, the boundary conditions must be homogeneous. The theory can be extended to inhomogeneous boundary conditions, but that is a digression we do not pursue.

many eigenvalues. For general coefficients $c_k(x)$, the task of calculating these eigenvalues ranges from difficult to impossible. Nevertheless, because of the importance of this problem in applications, over the decades and centuries eigenvalues have actually been calculated for surprisingly many specific examples. Moreover, for a few operators in which the coefficients c_k are constant, this calculation is fairly transparent.

As an example, let us calculate the eigenvalues of the simplest problem of the form (10.14),

$$y'' = \lambda y; \qquad y(0) = 0, \quad y(\ell) = 0. \tag{10.15}$$

If $\mu^2 = \lambda$, where λ may be complex but (for the moment) is nonzero, the general solution of the ODE (10.15) is

$$y(x) = C_1 e^{\mu x} + C_2 e^{-\mu x}.$$

Applying the boundary condition at $x = 0$, we conclude that $C_2 = -C_1$. Applying the boundary condition at $x = \ell$, we conclude that either $C_1 = C_2 = 0$ or $e^{\mu\ell} = e^{-\mu\ell}$. $C_1 = 0$ yields only the zero solution, so for λ to be an eigenvalue, we must have the latter alternative; this means that $\mu\ell = n\pi i$ for some integer n, and squaring μ gives $\lambda = -(n\pi/\ell)^2$. We return to consider $\lambda = 0$. In this case, the general solution of the ODE is $y(x) = C_1 + C_2 x$, and the boundary conditions imply that $C_1 = C_2 = 0$, so zero is not an eigenvalue. To conclude, we have calculated that (10.15) has an infinite sequence $\{\lambda_n\}$ of (real) eigenvalues that tend to negative infinity,

$$\lambda_n = -(n\pi/\ell)^2; \quad n = 1, 2, 3, \dots \ . \tag{10.16}$$

The eigenfunction corresponding to λ_n is $\sin(n\pi x/\ell)$.

We return to the difficulties with (10.1), (10.3). In Exercise 6, we ask you to show that $\lambda = 0$ is an eigenvalue of (10.13) if and only if $\ell = (\ln 3)/2$. In other words, solving (10.1), (10.3) is problematic precisely when zero is an eigenvalue of L. (Would it be pedantic to point out the analogy with solving a finite-dimensional linear system $A\mathbf{x} = \mathbf{b}$?)

10.2 Stochastic Population Models

ODEs describe deterministic systems: the complete evolution of a model follows once its initial state is specified. Stochastic differential equations (SDEs) expand such models to include the effects of randomness. A tiny corner of SDEs—birth–death processes—makes contact with this book. Specifically, birth–death processes provide a more faithful representation for the evolution of some systems that we have described by "bathtub models," such as population growth in ecology or gene

networks. For serious study of stochastic processes, we refer you to [26, 48] or [49]. Here, we merely indicate briefly how the most basic of ODEs,

$$x' = \alpha x, \tag{10.17}$$

can be embedded in a probabilistic model.

In the simplest stochastic analogue of (10.17), one studies a time-dependent random variable, say $X(t)$. This variable, which assumes integer values, 0,1,2,..., may jump up or down by one through "births" and "deaths" that occur independently at random times, according to the axioms below. Depending on the population at time t, the probabilities[7] of a jump in a short interval $(t, t+h)$ are given by

$$\begin{array}{lllll} \text{(a)}\ \mathbb{P}\{X(t)=n & \& & X(t+h)=n+1\} & = & [n\beta h + o(h)] \quad \mathbb{P}\{X(t)=n\}, \\ \text{(b)}\ \mathbb{P}\{X(t)=n & \& & X(t+h)=n-1\} & = & [n\delta h + o(h)] \quad \mathbb{P}\{X(t)=n\}, \end{array} \tag{10.18}$$

where β and δ are positive constants, the "birth" and "death" rates, respectively. Here is the thinking behind these formulas. The $o(h)$-correction terms[8] in (10.18) are related to multiple events, e.g., two or more individuals giving birth, a new-born individual also giving birth, both a birth and a death occurring during the interval. If h is small, the probability of any such multiple event is $o(h)$ small. Now modulo multiple events, the population will increase by precisely 1 during the interval $(t, t+h)$ iff one of n individuals gives birth during the interval, and for each individual,

$$\mathbb{P}\{\text{Individual } j \text{ gives birth}\} = \beta h + o(h). \tag{10.19}$$

Equation (10.18a) synthesizes these ideas. Similarly with (10.18b) for deaths. Incidentally, the probability of no change in the population is given by

$$\mathbb{P}\{X(t)=n \quad \& \quad X(t+h)=n\} \quad = \quad [1 - n(\beta+\delta)h + o(h)] \quad \mathbb{P}\{X(t)=n\}.$$

The connection of (10.18) to the ODE (10.17) is made through the *expected value* of the probabilistic model,

$$\mathbb{E}\{X(t)\} = \sum_{n=0}^{\infty} n\mathbb{P}\{X(t)=n\}. \tag{10.20}$$

Specifically, we have the following statement:

Proposition 10.2.1. *The expected value $x(t) = \mathbb{E}\{X(t)\}$ is differentiable and satisfies the ODE (10.17), where $\alpha = \beta - \delta$.*

---

[7] If you are familiar with probability theory, you may recognize that (10.18) could be formulated more simply in terms of conditional probabilities.

[8] The precise values of the $o(h)$-correction terms here and below are different. Part of the power of the order notation is to allow such sleight of hand in rigorous analysis.

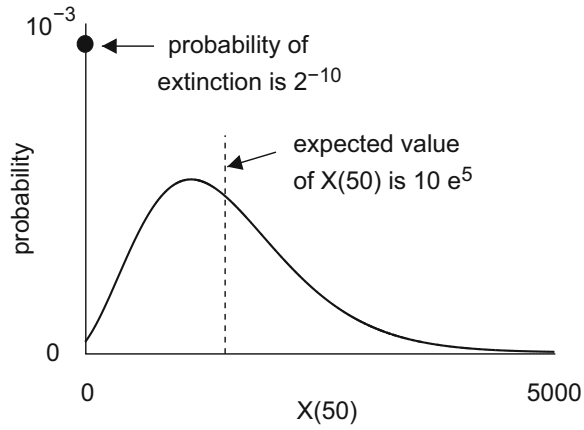

Figure 10.5: *Probability distribution at time $t = 50$ of the birth–death process (10.18), where $\beta = 0.2$, $\delta = 0.1$, and $X(0) = 10$. Apart from extinction, the discrete random variable $X(t)$ is characterized approximately by a continuous probability distribution function $P(x)$; thus, if $x > 0$, then $P(x) \approx \mathbb{P}\{x \le X(50) < x+1\}$. Note that the probability of extinction $\mathbb{P}\{X(50) = 0\}$ is much greater than $\lim_{x\to 0} P(x)$.*

Idea of proof. Without worrying about convergence issues, we prepare to form a difference quotient:

$$x(t+h) = \sum_{n=0}^{\infty} n\mathbb{P}\{X(t+h) = n\}. \tag{10.21}$$

Now by (10.18) and independence,

$$\begin{aligned} \mathbb{P}\{X(t+h) = n\} \;=\; & [1-(\beta+\delta)nh] \quad \mathbb{P}\{X(t) = n\} \\ & +\beta(n-1)h \quad \mathbb{P}\{X(t) = n-1\} \\ & +\delta(n+1)h \quad \mathbb{P}\{X(t) = n+1\} + o(h). \end{aligned} \tag{10.22}$$

On substituting (10.22) into (10.21) and shifting the summation index—do this carefully; it's what gives the RHS of (10.23)—we find that

$$x(t+h) - x(t) = (\beta-\delta)\sum_{n=0}^{\infty} nh\,\mathbb{P}\{X(t) = n\} + o(h). \tag{10.23}$$

The result follows on dividing by h and taking the limit $h \to 0$. □

The behavior in the stochastic model is far more complicated than in the ODE (10.17). In the first place, even if $\beta > \delta$, the stochastic model may suffer extinction; i.e., there may be a time T such that $X(T) = 0$. Indeed, if $X(0) = n$ (and if $\beta > \delta$), then the probability of eventual extinction equals $(\delta/\beta)^n$. Of course, this probability tends to zero as $n \to \infty$.

Under any circumstances, the stochastic model specifies the population only through a peaked probability distribution, such as sketched in Figure 10.5. The expected value $\mathbb{E}\{X(t)\} = X(0)e^{\alpha t}$ provides an estimate for the position of the peak. Thus, as t increases, the peak moves exponentially fast towards higher populations. The standard deviation of the distribution, the square root of its variance, provides a measure of the width of the distribution, or the size of deviations from $\mathbb{E}\{X(t)\}$ to be anticipated. Because the standard deviation is large, the expected value $\mathbb{E}\{X(t)\}$ lies to the right of the peak of the distribution.

Even though this stochastic model is a lot more complicated than the ODE, it still seems painfully simplistic as a description of real populations.

10.3 Numerical Methods: Two Sobering Examples

10.3.1 Stiff ODEs

Consider Euler's method applied to the test problem

$$x' = -Mx, \qquad x(0) = 1, \tag{10.24}$$

where $M > 0$. The approximation y_n to $x(nh) = e^{-Mnh}$ is $y_n = (1 - Mh)^n$. If M is large, this approximation is completely off track until h becomes very small; indeed, unless $h < 2/M$, the approximate solution is a *growing* exponential that oscillates on the grid scale. As this example illustrates, it is difficult for a naive numerical method to track a solution of an ODE that decays rapidly.

As regards (10.24), you may be thinking, "So what? Just take appropriately small steps to follow the exact solution to zero and then quit." Well, the real difficulty comes from, for example, fast–slow systems like the one in Exercise 4.23. In such systems, the overall evolution proceeds at a relatively slow rate, but some components of the solution evolve very quickly. (In numerical contexts the term *stiff* is used to describe a system in which different components evolve at widely divergent rates.) If you use a naive numerical method on a stiff system, you may be forced to choose a ridiculously small step size determined by the fastest time scale in the problem. If so, an enormous number of steps are required to compute the solution over quite modest times, which wastes computing resources and introduces unnecessary round-off errors.

Faced with a stiff system, you may be able to sidestep the problem by making an asymptotic approximation that eliminates the fast time scale. But in any case, there are numerical methods specifically designed to cope with stiffness. The simplest of these is the *backward* or *implicit Euler method*:

$$y_{n+1} = y_n + hF(y_{n+1}). \tag{10.25}$$

Note that y_{n+1} is not given explicitly by (10.25) but must be determined by solving an (in general nonlinear) implicit equation for it; thus each step using (10.25) involves more work than using the ordinary Euler method. On the other hand, (10.25) avoids the instability of Euler's method observed above. For example, applying (10.25) to the test problem (10.24) (in which solving the *linear* implicit equation at each step is no obstacle), we obtain

$$y_n = \frac{1}{(1+Mh)^n};$$

no matter how large $M > 0$ may be, the approximate solution decays docilely to zero, regardless of the size of h.

While the implicit Euler method has excellent stability properties, its accuracy is limited; as with the ordinary Euler method, the error in the numerical solution is approximately proportional to h; this behavior is called first-order accuracy. More accurate numerical schemes are given in Quarteroni et al. [65] and [45]. We do not attempt to summarize any of these methods, because we believe that you should first learn some of the supporting theory and see demonstrations of how they perform when applied to specific problems.

10.3.2 Unreasonable Behavior of Reasonable Methods

Here's an idea for a cheap way to increase accuracy compared to Euler's method: given a scalar ODE $x' = F(x)$, consider the formula

$$x((n+1)h) \;=\; x((n-1)h) + \int_{(n-1)h}^{(n+1)h} F(x(s))\,ds$$

and estimate the integral by a one-term Riemann sum using the midpoint of the interval $((n-1)h, (n+1)h)$,

$$\int_{(n-1)h}^{(n+1)h} F(x(s))\,ds = 2hF(x(nh)) + \mathcal{O}(h^3). \tag{10.26}$$

Neglecting the $\mathcal{O}(h^3)$-error, we obtain the recurrence relation called the "midpoint method":

$$y_{n+1} = y_{n-1} + 2hF(y_n). \tag{10.27}$$

This is an example of a *multistep method*; to compute y_{n+1}, you need to know y_n *and* one or more earlier approximates.[9]

[9] If you are trying to solve an IVP with (10.27), y_0 can be taken from the initial conditions, but y_1 must be determined by some other method. In principle, you could calculate y_1 with a higher-order one-step method, but there is no need: you do not sacrifice the higher-order accuracy of (10.27) if *for one step only* you compute y_1 with the less-accurate Euler approximation, $y_1 = y_0 + hF(y_0)$.

By way of comparison, Euler's method can be derived from a one-term Riemann sum using the left-hand endpoint of an interval,

$$\int_{nh}^{(n+1)h} F(x(s))\,ds = hF(x(nh)) + \mathcal{O}(h^2).$$

Since the error term in (10.26) is of higher order in h, we might expect the midpoint method to be more accurate than Euler's method. Let's see how the midpoint method performs on the test problem (10.24), say with $M = 1$, because stiffness is not the issue here. For this test problem, (10.27) reduces to

$$y_{n+1} = y_{n-1} - 2hy_n. \tag{10.28}$$

Looking for solutions of (10.28) of the form $y_n = r^n$, we derive the characteristic equation $r^2 + 2hr - 1 = 0$. Thus, the general solution of (10.28) is

$$y_n = C_+ r_+^n + C_- r_-^n, \tag{10.29}$$

where

$$r_\pm \;=\; -h \pm \sqrt{h^2+1} \approx \begin{cases} 1-h & (\text{for } r_+), \\ -1-h & (\text{for } r_-). \end{cases}$$

To approximate x at time t, we must take $n = [t/h]$ steps, where square brackets indicate the "greatest integer" function. With this many steps, as $h \to 0$, the first term in (10.29) is approximately

$$C_+(h)(1-h)^{[t/h]} \approx C_+(h)e^{-t},$$

where we have applied the formula from calculus $\lim_N (1 + x/N)^N = e^x$. Thus, provided the coefficient C_+ determined by the initial conditions y_0, y_1 satisfies $C_+(h) \approx 1$, this term by itself gives a tolerable approximation to the solution e^{-t} of the test problem. By contrast, the second term in (10.29) is approximately

$$C_-(h)(-1-h)^{[t/h]} \approx \pm C_-(h)e^t,$$

and this term is a disaster—it grows exponentially, and it oscillates on the grid scale. You might hope that the coefficient C_- determined by the initial conditions tends to zero as $h \to 0$, and in a strict mathematical sense, if $y_1 = 1 - h + \mathcal{O}(h^2)$, this is true; hence, if (10.28) were implemented with *infinite-precision arithmetic*, the approximate solution computed from (10.28) would converge to the solution of the test problem. In practice, however, this fact is irrelevant. Even assuming that $y_1 = r_+ y_0$, so that $C_- = 0$, if (10.27) is implemented on the computer, round-off errors constantly feed the problematic term in (10.29).

The moral: First off, don't compute with this method, except possibly to check whether it's as bad as we say. More broadly, there are risks in creating numerical

methods on your own. While we encourage experimentation, this area has many pitfalls for the unwary. It is prudent to inform yourself about the collective wisdom regarding numerical solutions of ODEs that has been developed over many years by a large, talented research community.[10] This expertise is available at your fingertips, among other places, in the references.

10.4 ODEs on a Torus: Entrainment

As we saw in Exercise 7.7, linear flow on a torus, the ODE

$$\begin{aligned} \theta_1' &= \omega_1, \\ \theta_2' &= \omega_2, \end{aligned} \tag{10.30}$$

where θ_1, θ_2 are reduced modulo 2π, presents a straightforward dichotomy: If ω_1/ω_2 is rational, every orbit is periodic, while if ω_1/ω_2 is irrational, every orbit is aperiodic and dense in the torus. For nonlinear ODEs on a torus, the situation is far more complex. Example 4 in Chapter 7,

$$\begin{aligned} \theta_1' &= \omega_1 + K_1 \sin(\theta_2 - \theta_1), \\ \theta_2' &= \omega_2 - K_2 \sin(\theta_2 - \theta_1), \end{aligned} \tag{10.31}$$

provided a hint of this. We saw that for a range of parameters, specifically if

$$\left| \frac{\omega_1 - \omega_2}{K_1 + K_2} \right| < 1, \tag{10.32}$$

then (10.31) has the two periodic solutions (7.8) in which the phases of the two variables are locked together or entrained. In fact, for many parameter ranges, the variables in (10.31) become entrained at different frequency ratios, and this is the generic behavior for ODEs on a torus.[11]

Our goal is to demonstrate the phenomenon rather than to be general. To simplify our task, we restrict attention to an example, a special case of (10.31),

$$\begin{aligned} \theta_1' &= 1, \\ \theta_2' &= \omega - \sin(\theta_2 - \theta_1), \end{aligned} \tag{10.33}$$

on which we impose the initial conditions

$$\theta_1(0) = 0, \qquad \theta_2(0) = b. \tag{10.34}$$

[10] Can you possibly ignore a heavy-duty admonition like this one? We hope not!

[11] Huygens observed entrainment in 1666: over time, the pendulums of two clocks placed on the same board became synchronized, 180° out of phase. You can track down many other instances of it on the Internet.

Let us use flow notation to write $(\theta_1(t), \theta_2(t)) = (t, \varphi(t, b; \omega))$ for the solution to this IVP, where we focus on the nontrivial second component; in this notation we do *not* reduce $\theta_k(t)$ modulo 2π. It follows from (10.33) that

$$b + (\omega - 1)t \le \varphi(t, b; \omega) \le b + (\omega + 1)t. \tag{10.35}$$

Also, computing the derivative using (4.79), we conclude that

$$\frac{\partial \varphi}{\partial \omega}(t, b; \omega) > 0 \qquad \text{for } t > 0. \tag{10.36}$$

The key tool in analyzing the behavior of (10.33) is the rotation number

$$\rho(\omega) = \lim_{T \to \infty} \frac{1}{T} \varphi(T, b; \omega). \tag{10.37}$$

Two basic properties of the rotation number[12] are these:

- The limit in (10.37) exists and is independent of the initial condition b.
- The function $\rho(\omega)$ is continuous and (not strictly) monotone increasing.

This quantity measures the "average" rate at which (the second component of) solutions of (10.33) move around the circle.

Usually, rotation numbers are defined for a homeomorphism of the circle $\Phi : S^1 \to S^1$. In our problem, the underlying homeomorphism is $b \mapsto \Phi(b, \omega)$, where

$$\Phi(b, \omega) = \varphi(2\pi, b; \omega) \qquad (\text{reduced mod } 2\pi);$$

in the first argument of φ on the RHS of this equation, $t = 2\pi$ represents the time needed for θ_1 to advance one period. The following remarkable properties are what make the rotation number so useful:

- If $\rho(\omega)$ is rational, say $\rho(\omega) = m/n$, then the nth iterate of Φ has a fixed point; i.e., there exists $b \in S^1$ such that $\Phi^n(b, \omega) = b \pmod{2\pi}$. If m and n have no common factors, then n is the minimal period of the discrete trajectory $\{\Phi^k(b, \omega) : \ k = 0, 1, \ldots\}$.
- If $\rho(\omega)$ is irrational, then $\{\Phi^k(b, \omega) : \ k = 0, 1, \ldots\}$ is aperiodic and dense in the circle.

If $\rho(\omega) = m/n$, then the iterates $\Phi^k(b, \omega)$, $k = 0, 1, \ldots, n$, of the fixed point complete exactly m circuits around S^1 before returning to b. Some effort is needed to interpret

[12] These properties and the properties quoted below are derived, for example, in Section 1.14 of [17].

this statement for a general homeomorphism on S^1, but in our case, because $\Phi(\cdot,\omega)$ comes from solving an ODE on $\mathbb{R}^2$, the statement has a simple formulation: the fixed point b satisfies

$$\varphi(2n\pi, b; \omega) = b + 2m\pi. \tag{10.38}$$

Conversely, if for a given ω there is at least one b that satisfies (10.38), then $\rho(\omega) = m/n$. (Incidentally, if $\rho(\omega)$ is irrational, then there is a coordinate change on S^1 that reduces $\Phi(\cdot,\omega)$ to rotation[13] by the angle $\rho(\omega)$.)

Let us show how solutions of (10.33) become entrained near values of the frequency parameter such that $\rho(\omega)$ is rational. Suppose that for $\omega = \omega_*$,

$$\rho(\omega_*) = m/n, \tag{10.39}$$

where m and n have no common factors. Figure 10.6(a) shows a hypothetical graph[14] of

$$\varphi(2n\pi, b; \omega_*) - b - 2m\pi \tag{10.40}$$

as a function of b. Because of (10.39), there will be at least one point where (10.38) is satisfied; in the figure, we show two such points, say b_1 and b_2, and the graph crosses the b-axis at them. (Generically, there is an even number of such points.) The figure indicates that if $b_1 < b < b_2$, then $\varphi(2n\pi, b; \omega_*)$ doesn't quite advance a full m revolutions around the circle; similarly, if $0 \le b < b_1$ or $b_2 < b \le 2\pi$, it advances a little more than m revolutions.

To understand entrainment, consider solving (10.38) for ω as a function of b, which the following lemma shows is possible.

Lemma 10.4.1. *For every m, n, and b, there is a unique value of ω that satisfies (10.38), and it varies smoothly with b.*

Proof. It follows from manipulating (10.35) that for all b,

$$\varphi(2n\pi, b; m/n - 1) \le b + 2m\pi \le \varphi(2n\pi, b; m/n + 1),$$

so by continuity, there must be at least one value of $\omega \in [m/n - 1, m/n + 1]$ that solves (10.38). We invoke (10.36) and the implicit function theorem to conclude that this solution is unique and depends smoothly on b. □

Figure 10.6(b) shows a possible graph of the solution ω of (10.38) as a function of b, which is qualitatively consistent with Figure 10.6(a): If $b = b_1$ or b_2, then

[13] A technical issue: the existence of this conjugacy requires that Φ be at least $\mathcal{C}^2$. Of course, this and more are satisfied here.

[14] Real graphs of this type are usually so flat that no structure can be seen. Besides being purely hypothetical, the vertical scale in the figure is enormously expanded.

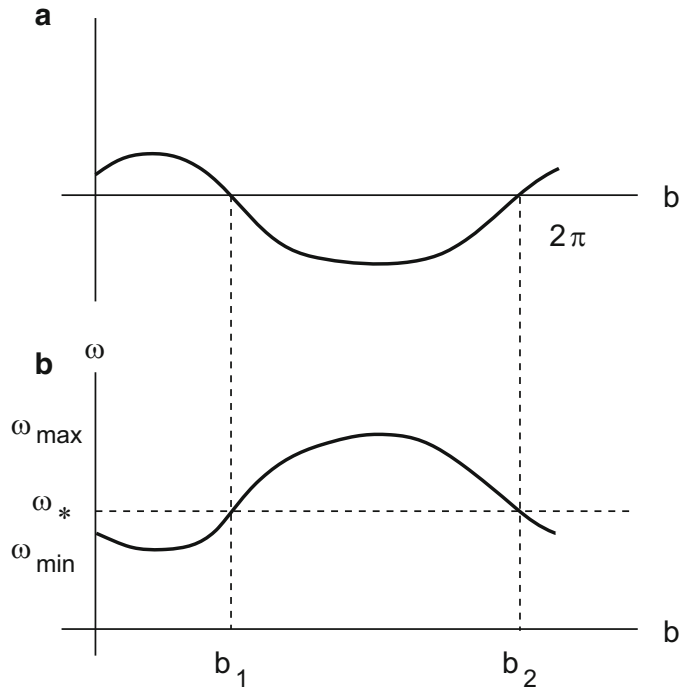

Figure 10.6: *(a) A hypothetical graph of* (10.40)*, the net advance of* $\varphi(2n\pi, b; \omega)$ *as a function of* b*. (b) The solution* ω *of* (10.38) *as a function of* b*, consistent with Panel (a). (Cf. Lemma 10.4.1.)*

$\omega = \omega_*$ satisfies (10.38). If $b_1 < b < b_2$, then a slightly larger value of ω is needed for $\varphi(2n\pi, b; \omega)$ to advance a full m revolutions; if $0 \le b < b_1$ or $b_2 < b \le 2\pi$, a slightly smaller value.

In the figure, $\omega_{\max}$ and $\omega_{\min}$ refer to the maximum and minimum values of ω obtained by solving (10.38). Thus, if $\omega_{\min} \le \omega \le \omega_{\max}$, there is at least one b such that (10.38) is satisfied. Consequently, $\rho(\omega) \equiv m/n$ for all ω in the interval $[\omega_{\min}, \omega_{\max}]$. That is, for ω in this interval, solutions of (10.33) are entrained at a constant frequency ratio.

The graph of the rotation number as a function of ω is simply weird. For every rational number r, this graph has a plateau on which $\rho(\omega)$ has the constant value r, but for every irrational r, the graph passes through r at a single point. Try to plot a function with these properties, or even to convince yourself that such a function is possible. You will see why some poetically inclined researchers call such a graph "the devil's staircase."

10.5 Delay Differential Equations

Let us consider a *delay differential equation* (DDE) that is of first order, scalar, and autonomous,

$$\frac{dx}{dt}(t) = f[x(t-\tau)], \tag{10.41}$$

where $f : \mathbb{R} \to \mathbb{R}$ and $\tau > 0$ is the delay. The key feature of (10.41) is that dx/dt depends on a *past* state of the system. This equation does not include any of the more complicated forms such equations may assume (cf. Exercise 9.16 or more generally [24]), but even this simple example confronts us with two surprises.

The first surprise relates to initial conditions: $x(t)$ may be specified arbitrarily on the interval $0 \le t \le \tau$. Given a function $\phi(t)$ on the interval $[0, \tau]$, let us construct (in forward time) a solution of (10.41) such that

$$x(t) = \phi(t), \qquad 0 \le t \le \tau. \tag{10.42}$$

By the fundamental theorem of calculus, if $t > \tau$, then

$$x(t) = x(\tau) + \int_\tau^t x'(s)ds.$$

But if $\tau \le t \le 2\tau$, the RHS of this equation may be evaluated solely in terms of the initial function $\phi(t)$: shifting the range of integration based on the delay, we obtain

$$x(t) = \phi(\tau) + \int_0^{t-\tau} f[\phi(s)]ds,$$

which defines the solution (uniquely) for $\tau \le t \le 2\tau$. Proceeding iteratively, we may obtain the solution on every interval $[n\tau, (n+1)\tau]$, $n = 1, 2, 3, \ldots$.

Let us rephrase this information: The set of solutions of (10.41) is parametrized by an arbitrary function on $[0, \tau]$. In other words, *the set of solutions of* (10.41) *is infinite-dimensional!* Quite a contrast with ODEs.

This complication makes some aspects of the theory of DDEs a little technical. For example, consider a simple linear DDE of the form (10.41),

$$x'(t) = -ax(t - \tau). \tag{10.43}$$

As with constant-coefficient ODEs, solutions of (10.43) can be sought as linear combinations of exponentials. Substituting into the equation, we calculate that $e^{\lambda t}$ solves (10.43) iff λ satisfies the transcendental characteristic equation

$$\lambda = -ae^{-\lambda\tau}. \tag{10.44}$$

Although this equation cannot be solved explicitly, *it has infinitely many roots* in the complex plane. (See Exercise 8, or fire up your computer.) Every convergent linear combination of the exponential solutions is also a solution.

The second surprise comes from studying solutions of (10.43), assuming $a > 0$, as the delay τ increases. If $\tau = 0$, then all solutions of the ODE (i.e., multiples of e^{-at}) decay monotonically to zero as $t \to \infty$. If $\tau > 0$ is small, specifically, if $\tau < \pi/2a$, then all roots of (10.44) lie in the left half-plane, and solutions of (10.43) still decay. However, if τ exceeds this threshold, then two of the roots move into the right half-plane (*a Hopf bifurcation!*), and more follow for larger τ. After this point, generic solutions of (10.43), which are oscillatory, suffer exponential growth. *Check this numerically.*[15] Since the behavior is generic, you would have to be very unlucky in choosing the initial function $\phi(t)$ not to see it. Analytically, here are a couple of easily verified facts that support the claim of instability:[16]

- If $\tau = \pi/2a$, then $\lambda = \pm i\pi/2\tau$ are roots of (10.44), and $\sin(at)$, $\cos(at)$ are nondecaying oscillatory solutions of (10.43).
- With implicit differentiation of (10.44) you can show that these roots move into the right half-plane if τ increases beyond $\pi/2a$.

Thus, when τ is slightly greater than $\pi/2a$, equation (10.43) admits exponentially growing oscillatory solutions.

The instability can be related to an everyday experience: adjusting the water temperature when you take a shower. When the water is too hot or cold, you rotate

[15] XPPAUT is one example of free software that can solve delay differential equations numerically.

[16] See Section 2.1 of [24] for a more complete analysis.

the dial to make the water cooler or warmer, as desired. However, it takes time for the water to flow from the dial to the shower nozzle, and if the delay is too large, it is easy to overcompensate, and your efforts may result only in a frustrating alternation between scalding and freezing water.

Let us attempt to model this experience with a DDE for the temperature,

$$\frac{dT(t)}{dt} = -a[T(t-\tau) - T_{\text{comf}}], \tag{10.45}$$

where T_{comf} is the comfortable temperature that you wish to achieve. (The equation reduces to (10.43) with the substitution $x = T - T_{\text{comf}}$.) In this model, it is assumed that displaying incredible self-control, you react to a misadjustment in temperature by turning the dial continuously with a speed proportional to the size of the temperature mismatch. The proportionality constant a characterizes your impatience to arrive at a comfortable temperature. If the delay τ vanishes or is small, the temperature simply approaches the desired temperature asymptotically, but if τ is too large (relative to a^{-1}), the temperature oscillates out of control.

DDEs have been used to model various phenomena, including periodic cycles of population, traffic flow patterns, economic trends, the human respiratory-control system, and many others (cf. [24]). Regarding population models, let us mention the Hutchinson–Wright equation[17]

$$\frac{dy}{dt} = ay(t)\left[1 - \frac{y(t-\tau)}{K}\right]. \tag{10.46}$$

If $\tau = 0$, this equation reduces to the logistic equation (1.2). Both the logistic equation and (10.46) with $\tau > 0$ admit $x(t) \equiv K$ as an equilibrium solution. For the logistic equation, this equilibrium is stable. However, by comparing the linearization of (10.46) around the equilibrium with (10.43), you will find that if $\tau > \pi/2a$, then the equilibrium is unstable for (10.46). In contrast to a linear equation, when $\tau > \pi/2a$, generic solutions of (10.46) do not grow indefinitely; rather, they settle down into finite-amplitude periodic oscillations, as you may readily compute. To conclude, the delay-induced instability in (10.46) helped ecologists understand oscillations in some populations that occur even in the absence of predators.

[17]Note that dy/dt depends on *both* a past state and the present state of the system. Consequently, it requires more effort to prove existence for (10.46) than for (10.41).

10.6 A Peek at Chaos

10.6.1 A One-Dimensional Mapping Model

In Exercise 9.13 we considered iteration of the quadratic map $x \mapsto \mu x(1-x)$, which for $0 \le \mu \le 4$ maps the unit interval $[0,1]$ into itself. We saw that this mapping undergoes a period-doubling cascade as $\mu \to \mu_\infty \approx 3.56995$ and that for a range of μ above μ_∞, orbits appear to be chaotic. Exploring these phenomena in detail is beyond the scope of this book. In this subsection our goal is more limited: considering the quadratic map only for the exact value $\mu = 4$,

$$\Psi(x) = 4x(1-x), \qquad 0 \le x \le 1, \tag{10.47}$$

we describe (without proofs) specific chaotic behavior that can be clearly identified in this case. (For proofs, see [17], especially Exercises 6–10 of Section 1.8.)

Although on the surface, the mapping (10.47) has little to do with ODEs, its behavior illustrates chaos in its purest form: the shift operator on symbol space. Analogous behavior has been derived for some ODEs, but it seems to be difficult to derive such results for most ODEs that arise in real applications.

We may define *symbols* for (10.47) as follows: Divide $[0,1]$ into two subintervals $\mathcal{I}_1 = [0, 1/2]$ and $\mathcal{I}_2 = [1/2, 1]$, which overlap at the point $x = 1/2$. We shall call an infinite sequence $s_0, s_1, s_2, \ldots$, where $s_n = 1$ or 2, an *itinerary* of a point $x \in [0,1]$ if

$$x \in \mathcal{I}_{s_0},\ \Psi(x) \in \mathcal{I}_{s_1},\ \Psi^2(x) \in \mathcal{I}_{s_2},\ \ldots\ . \tag{10.48}$$

If $s_0, s_1, s_2, \ldots$ is an itinerary for x, then $s_1, s_2, s_3, \ldots$ is an itinerary for $\Psi(x)$. More formally, Ψ acting on $[0,1]$ is topologically equivalent (as defined in Section 1.7 of [17]) to the shift operator

$$S(s_0, s_1, s_2, \ldots) = s_1, s_2, s_3, \ldots$$

acting on symbol space. Note that we cannot say *the* itinerary of x because of the following ambiguity: if $\Psi^N(x) = 1/2 \in \mathcal{I}_1 \cap \mathcal{I}_2$ for some N, then x has two itineraries that differ in the Nth entry,

$$\begin{aligned} &s_0, s_1, \ldots, s_{N-1},\ 1,\ 2, 1, 1, 1, \ldots\ , \\ &s_0, s_1, \ldots, s_{N-1},\ 2,\ 2, 1, 1, 1, \ldots\ . \end{aligned} \tag{10.49}$$

Although there are only countably many such points, this minor issue slightly complicates the exposition. For example, in the following theorem we want to write $T_N(x)$ for the truncated itinerary $s_0, s_1, s_2, \ldots, s_N$ of x, but to make this notation well defined, we need to adopt a convention such as choosing the first entry in (10.49) in ambiguous cases.

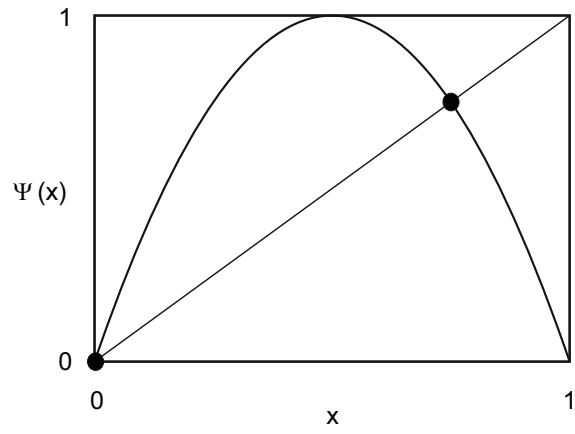

Figure 10.7: *The quadratic map* (10.47). *The fixed points are* $x = 0$ *and* $x = \frac{3}{4}$.

Theorem 10.6.1. *(i) For every infinite sequence* $s_0, s_1, s_2, \ldots$ *, where* $s_n = 1$ *or* 2*, there is a unique* $x \in [0,1]$ *that has this itinerary. (ii) For every* $\varepsilon > 0$*, there is a positive integer* N *such that for all* $x, y \in [0,1]$*,*

$$T_N(x) = T_N(y) \implies |x - y| < \varepsilon. \tag{10.50}$$

For example, the point that has the itinerary $2, 2, 2, \ldots$ is easily identified: it is just the fixed point $x = 3/4$ of Ψ in Figure 10.7. By contrast, it is not readily apparent[18] where the point that has the itinerary $1, 2, 1, 2, \ldots$, is located, let alone the point with itinerary

$$1, 2,\ 1, 2, 2,\ 1, 2, 2, 2,\ 1, 2, 2, 2, 2,\ \ldots\ .$$

Whatever messy sequence may strike your fancy, there is an x that has exactly that itinerary.

Part (ii) of the theorem makes one think of continuity, but (10.50) relates to the continuity of an *inverse* map that associates to an itinerary the point x that has that itinerary. This formula shows the absolute futility of trying to make long-range predictions about itineraries under iteration by Ψ: no matter how small $\varepsilon > 0$ may be, there is an $N = N(\varepsilon)$ such that even knowing that x belongs to an arbitrarily small interval $(b - \varepsilon, b + \varepsilon)$ gives no information whatsoever about the itinerary of x beyond the Nth iterate; all continuations $s_{N+1}, s_{N+2}, s_{N+3}, \ldots$ are possible itineraries. Moreover, $N(\varepsilon) \sim \ln(\varepsilon^{-1})$, so even a huge refinement of accuracy gives only a modest gain in predictability. This mapping displays an extreme form of

[18] But see Exercise 7.

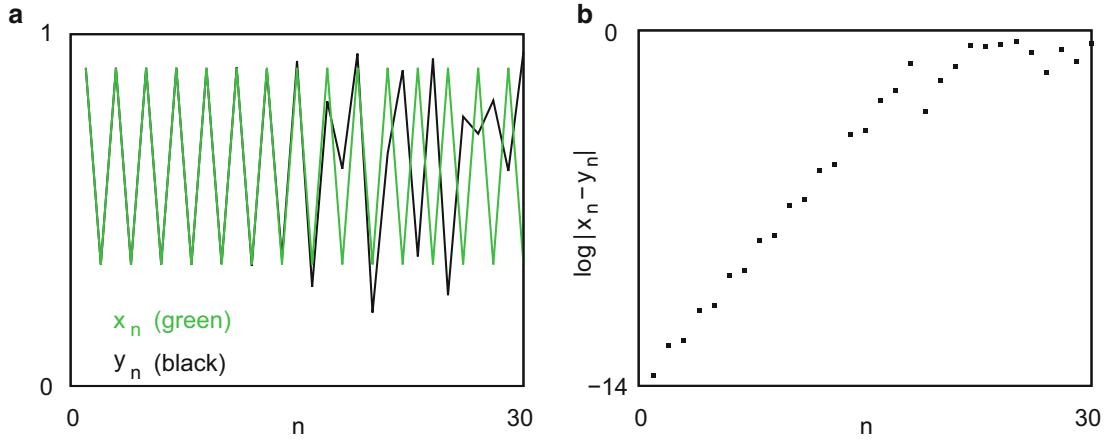

Figure 10.8: *Representations of the divergence of two nearby trajectories under iteration by the quadratic map* (10.47). *(A periodic trajectory is shown in green.)*

sensitive dependence on initial conditions, i.e., evolution in which small perturbations of initial conditions continue to grow until they completely change the long-term trajectory of a system.

Although the discrete dynamical system $x \mapsto \Psi(x)$ is deterministic, the long-term behavior of itineraries appears completely random. The description *chaos* does not seem inappropriate.

It is instructive to view how trajectories that start nearby diverge as the iteration proceeds. For example, Figure 10.8(a) shows two trajectories, with starting points $x = (5 - \sqrt{5})/8 \approx 0.345492$ and $y = x - 10^{-6}$. The first trajectory is periodic with period 2. Until $n = 15$, the two trajectories cannot even be distinguished visually. Nevertheless, as the log plot in Figure 10.8(b) shows, the separation between the two sequences is growing during this time. It continues to grow until around $n = 18$, by which time the separation is comparable to the width of the unit interval. It cannot continue to grow after this point, since both trajectories are contained in the unit interval.

As we have just noted, the separation between two trajectories cannot grow larger than $\mathcal{O}(1)$. One might try to observe growth of the separation over a longer period by choosing starting points that are closer and closer to each other. However, in numerical computations it is impossible to compare two trajectories whose initial separation is less than a lower limit set by round-off errors. We may escape these limitations by introducing the derivative of Ψ, which specifies how much the separation between *infinitesimally close* neighbors is amplified by Ψ. Given a trajectory $\{x_0, x_1, x_2, \ldots\}$, where $x_n = \Psi^n(x)$, for every positive integer N let

$$\mu_N(x) = \frac{1}{N} \sum_{n=0}^{N-1} \ln |\Psi'(x_n)|. \tag{10.51}$$

Then the separation between the Nth iterates of infinitesimally close neighbors of x is the initial separation amplified by a factor of $e^{N\mu_N(x)}$.

What happens to $\mu_N(x)$ as $N \to \infty$? Of course if $\Psi^N(x) = 1/2$ for some N, then $\Psi'(x_N) = 0$, so $\mu_{N+1}(x) = -\infty$, and similarly for subsequent terms. Nevertheless, apart from a few such exceptions, the sequence converges to a finite limit,

$$\mu = \lim_{N\to\infty} \mu_N(x) = \ln 2, \tag{10.52}$$

the same limit for almost every[19] *starting point* x! This convergence is trivial if $x = 3/4$, the fixed point of Ψ, because for this x, every term x_n in the trajectory equals $3/4$, and $|\Psi'(3/4)| = 2$. If $x = (5-\sqrt{5})/8$, then $\ln 2$ is a believable estimate for the average slope in the linear part of the graph in Figure 10.8(b), and in Exercise 7 you derive (10.52) exactly. By contrast, convergence is far from trivial if x generates a trajectory $\{x_0, x_1, x_2, \ldots\}$ that is dense in $[0, 1]$. In this case, $|\Psi'(x_n)|$ wanders up and down—it's nearly 0 when x_n is close to $1/2$ and nearly 4 when x_n is close to 0 or 1—but these variations average out, and $\mu_N(x)$ tends to $\ln 2$.

The limit μ in (10.52), called a *Lyapunov exponent*, gives a quantitative measure of sensitive dependence on initial conditions, which is a hallmark of chaos. That is, in an averaged sense, under each application of Ψ in an iteration, the separation between close neighbors is doubled.

In closing, let us note an important difference between (10.47) and the quadratic map $x \mapsto \mu x(1-x)$ with $\mu < 4$: Although for (10.47) there are points x for which the trajectory $x, \Psi(x), \Psi^2(x), \ldots$ is dense in $[0, 1]$, no such points exist when $\mu < 4$. For the mappings with $\mu < 4$, there are challenging issues about characterizing an attractor to which trajectories converge; similar issues, only partially resolved, also arise for ODEs with chaotic solutions, such as the Lorenz equations.

10.6.2 The Lorenz Equations

In the previous subsection we saw chaotic behavior in a noninvertible map in one dimension. Chaos can occur for an *invertible* map if the dimension is at least two; see, e.g., Section 12.2 of [81] or Section 5.5 of [33]. For ODEs, it follows from the Poincaré–Bendixson theorem, Theorem 7.2.5, that chaos can occur only if the dimension is at least three. Many ODEs in three or more dimensions have chaotic solutions, but in this section we focus exclusively on the Lorenz equations (9.37) as the archetype.

The solutions of the Lorenz equations in Section 9.7 (with $\rho = 28$, $\sigma = 10$, $\beta = 8/3$) certainly *look* chaotic, but what does this really mean? Various authors give

[19] In technical language, the set of exceptional points has Lebesgue measure zero. A brief introduction to measures from a dynamical-systems viewpoint is given in Section 11.4.1 of [67]. However, for this survey, the intuitive associations of this phrase suffice.

different definitions of chaos (and some, like us, don't give any), but sensitive dependence on initial conditions is at the core of every definition. In this section we first explore sensitive dependence on initial conditions in the Lorenz system, and then we briefly address more general issues pertaining to chaos.

(a) Sensitive dependence on initial conditions

The phrase *sensitive dependence on initial conditions* (SDIC), which is reserved for ODEs whose trajectories are bounded, conveys two properties:

P1. Generically,[20] trajectories that start from nearby initial conditions diverge exponentially fast from one another until further growth is limited by the fact that trajectories are bounded.

P2. Even after growth is cut off, generically the evolution of different trajectories is completely uncorrelated.

To clarify this concept, let us consider several examples that have some, but not all, of the requisite behavior.

Example 1: Solutions of the scalar equation $x' = x$ diverge exponentially, but they are unbounded. Likewise for the second-order equation $x'' = x - \beta x'$, which describes the motion of a particle in a linear repulsive force (with friction).

Example 2: If we modify the linear repulsive force by adding a cubic attractive force, we obtain Duffing's equation, which we rewrite as a first-order system

$$\begin{aligned} x' &= y, \\ y' &= x - x^3 - \beta y. \end{aligned} \tag{10.53}$$

Solutions of (10.53) are bounded, and generically solutions with initial conditions *that lie close to the saddle point at the origin* diverge exponentially, at least for a limited time. However, property P1 fails: trajectories not starting close to the origin (i.e., most trajectories) do not exhibit exponential divergence. P2 also fails. Specifically, Figure 10.9 shows the stable manifold $\mathcal{M}_s$ through the saddle point; the complement of $\mathcal{M}_s$ has two components, which are basins of attraction for the two stable equilibria of (10.53) at $(\pm 1, 0)$; all trajectories with initial conditions in the same basin of attraction share the same asymptotic fate, i.e., convergence to one of the equilibria.

Example 3. Consider the following ODE on the cylinder $(\mathbb{R}/2\pi\mathbb{Z} \times \mathbb{R})$,

$$\begin{aligned} x' &= y, \\ y' &= 0. \end{aligned}$$

[20] The caveat "generic" is necessary, because, for example, if two initial conditions happen to lie on the same orbit, the two trajectories will be translates of each other and therefore not diverge.

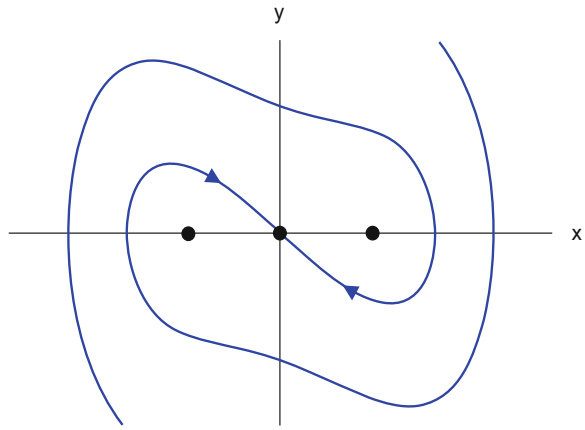

Figure 10.9: *Stable manifold for the saddle point of Duffing's equation* (10.53) *with* $\beta = 1/2$. *This curve separates the basins of attraction of the two stable equilibria,* $(-1, 0)$ *and* $(1, 0)$.

Nearby trajectories diverge as x winds around the circle (at a y-dependent rate), but the separation grows only *linearly* in time, not exponentially as required in P1, so the SDIC label does not apply.

As we said earlier, SDIC is a key property of chaotic ODEs. Consistent with that statement, no one would call any of the above ODEs chaotic.

The exponential divergence in P1 may be quantified with Lyapunov exponents. As in the previous subsection, we introduce differentials so that we can study exponential divergence of trajectories over arbitrarily large time intervals. To elaborate, let $\boldsymbol{\varphi}(t, \mathbf{b})$ be the solution operator for the IVP

$$\begin{array}{rcll} x' &=& \sigma(y - x), & x(0) = b_1, \\ y' &=& \rho x - y - xz, & y(0) = b_2, \\ z' &=& -\beta z + xy, & z(0) = b_3. \end{array} \tag{10.54}$$

The differential $\mathbf{D}\boldsymbol{\varphi}(t, \mathbf{b})$ is a 3×3 matrix with entries $\partial\varphi_j/\partial b_k$. The largest Lyapunov exponent,[21] the most important one, may be characterized as

$$\mu = \lim_{T\to\infty} \frac{\ln \|\mathbf{D}\boldsymbol{\varphi}(T, \mathbf{b})\|}{T}, \tag{10.55}$$

[21]For the three-dimensional Lorenz system, there are actually *three* Lyapunov exponents; see Section 7.2 of [54] or Chapter 29 of [95]. Incidentally, the sum of these three exponents equals the trace of the matrix in (10.57), or -13.6666. Since this number is large and negative, we see from Proposition 7.9.1 that volumes contract rapidly as time evolves in the Lorenz equations.

provided the limit exists and has the same value for almost all[22] initial conditions $\mathbf{b}$. Computations indicate that this proviso is satisfied (look online); based on the graphs of $\|\mathbf{D}\boldsymbol{\varphi}(T, \mathbf{b})\|/T$ in Figure 10.10 (and on other runs out to $T = 100000$), we make the estimate

$$\mu \approx 0.906. \tag{10.56}$$

The crucial information is that $\mu > 0$: generically, at least for a limited time, nearby trajectories will separate on average like the growing exponential $e^{\mu t}$.

The computations reported in Figure 10.10 are fairly straightforward.[23] Recalling the definition of $\boldsymbol{\varphi}(t, \mathbf{b})$ as the solution of (10.54), let $A(t) = \mathbf{DF}(\boldsymbol{\varphi}(t, \mathbf{b}))$, i.e.,

$$A(t) = \begin{bmatrix} -\sigma & \sigma & 0 \\ \rho - z(t) & -1 & -x(t) \\ y(t) & x(t) & -\beta \end{bmatrix}. \tag{10.57}$$

As discussed in Section 4.6.2,

$$\mathbf{D}\boldsymbol{\varphi}(t, \mathbf{b}) = \mathrm{Col}(\mathbf{w}_1(t), \mathbf{w}_2(t), \mathbf{w}_3(t)), \tag{10.58}$$

where $\mathbf{w}_i(t)$ satisfies

$$\mathbf{w}_j' = A(t)\mathbf{w}_j, \qquad \mathbf{w}_j(0) = \mathbf{e}_j. \tag{10.59}$$

To obtain the graph in Figure 10.10, we solved a twelve-dimensional system consisting of (10.54) together with the three associated linearized equations (10.59) and computed the norm of the matrix (10.58). For the record, we used initial conditions equal to $(1, 2, 3)$, but other initial conditions give the same estimate (10.56).

Figure 10.11 illustrates evolution consistent with this Lyapunov exponent. (Cf. Plate 2 of [81].) It summarizes 2500 simulations with initial data distributed over a square patch, of dimension 0.001×0.001, centered on

$$(0.756901, 1.35474, 13.2317)$$

and orthogonal to the trajectory through that point. In the various panels, each red dot shows the position of one of the trajectories at a later time. These points are displayed against a black background, which is the single trajectory starting from the center of the patch computed for $0 \le t \le 100$. The filament-like sets result

[22] Some initial conditions, such as on the stable manifold of the origin, may produce different behavior. In computations, however, this is a nonissue: even if you were so profoundly unlucky as to choose "bad" initial conditions, because of round-off errors as the computation proceeds, exponential growth will leak into your results and then dominate the computation.

[23] Similar calculations were proposed in Exercise 9.7.

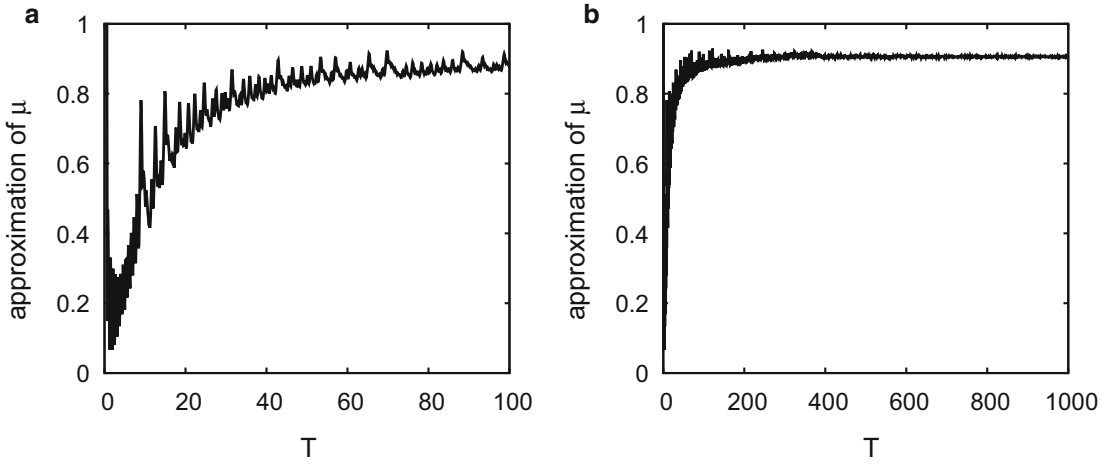

Figure 10.10: *(a) The initial transient of* $\ln \|\mathbf{D}\boldsymbol{\varphi}(T, \mathbf{b})\|/T$. *(b) Convergence to approximately* 0.906 *in the long term.*

from the exponential amplification (in one direction) of initial separations. By time $t = 16$, the different points have distributed themselves over a large fraction of the background trajectory.

Figure 10.11 hints at Property P2, but this behavior is better illustrated in Figure 10.12. The figure shows the x-component of the solution of (10.54) for $0 \le t \le 16$ along the four trajectories starting at the four corners of the patch of initial conditions used in Figure 10.11. For $0 \le t \le 8$, the trajectories have not yet separated very far, and the solutions roughly track one another. However, for $t \ge 12$, they have no connection to one another.

Lorenz proposed the equations (10.54) as an absolutely minimal model derived from weather prediction. Presumably, the PDEs that model global weather have at least the same level of SDIC, if not much greater. Such SDIC has been poetically encapsulated in what's known as the "butterfly effect": hypothetically, in one of its many variants, the disturbance caused by a butterfly flapping its wings in Brazil could grow to the point that after a not-clearly-identified-but-presumably-rather-short period, the path of a tornado in Texas could be altered.

(b) Strange attractors and rigorous results

Computations suggest a compelling description of the asymptotic behavior of solutions of the Lorenz system (10.54). For the standard parameter values, it appears that virtually all trajectories converge to a bounded attractor Λ, which, as shown in Figure 9.17(a), resembles a butterfly. Near a typical point of Λ, say in the middle of one of the "wings," the set may be represented as a cross product of a (two-dimensional) surface with a kind of Cantor set. Thus, both wings consist of infinitely many sheets. This structure is *fractal* (cf. Chapter 11 of [81]), which on an intuitive

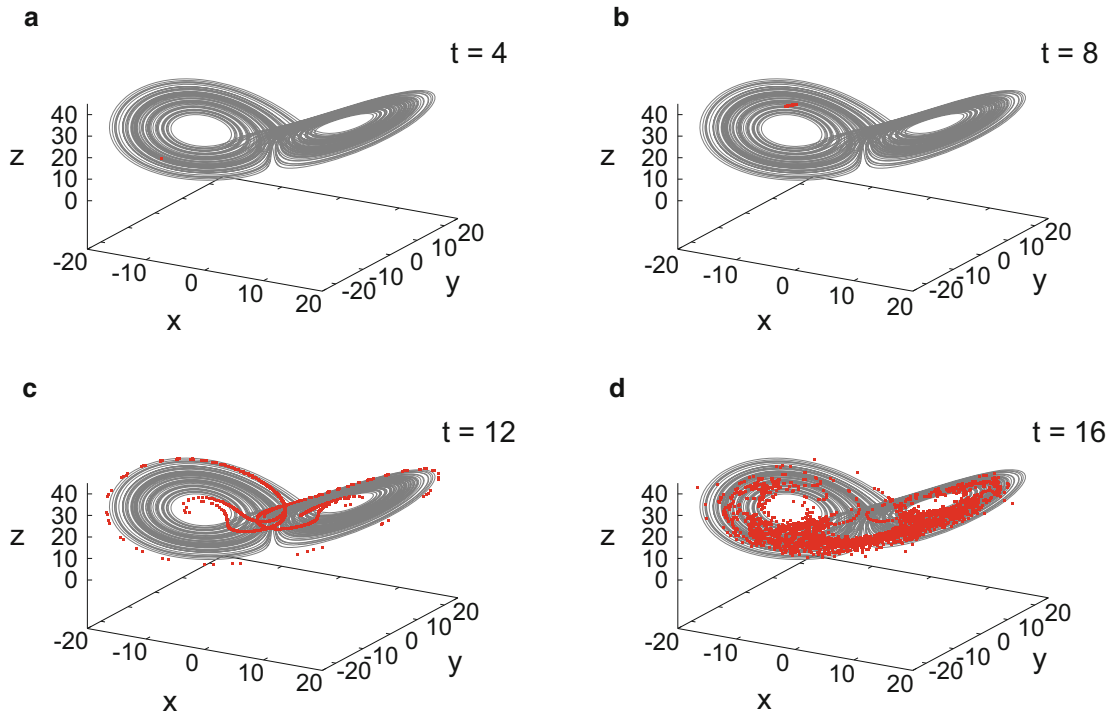

Figure 10.11: *An illustration of exponential divergence of solutions of* (10.54)*: the figure shows the positions at various later times of 2500 trajectories that all start close to one another.*

level means that if a small portion of Λ is viewed on various expanded scales, the results look remarkably like the original set. Such attracting ω-limit sets have come to be called *strange* attractors.

Although the numerical evidence is convincing, proofs would be nice. However, deriving rigorous results has challenged the best and the brightest; more than 50 years on, the story is still not complete. The excellent summary of progress in [91], which contains references to the original papers, is the place to learn more about this work. One line of research derives complete results for geometrically defined *approximations* of the Lorenz system; another line identifies specific chaotic behavior *embedded* in the solution set of the actual equations. This account is fascinating, but since it reports on the best efforts of some very smart people on a hard problem, it is not light reading.

In Chapter 7 we saw zero-dimensional limit sets (e.g., an equilibrium), one-dimensional limit sets (e.g., a periodic orbit), and two-dimensional limit sets (e.g., a torus, the closure of a skew-line orbit). A dimension for the Lorenz attractor Λ may be defined in various ways, but the result is not an integer. All definitions seem to give a dimension slightly greater than 2; the "correlation dimension" turns out to be 2.05 ± 0.01. (See Example 11.5.1 of [81].)

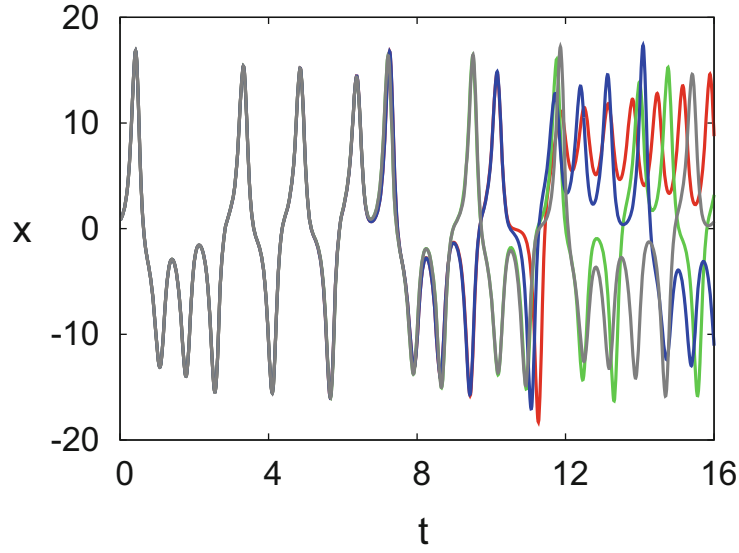

Figure 10.12: *An illustration that even though trajectories start close to one another, they become completely uncorrelated as time evolves.*

10.7 Exercises

10.7.1 Core Exercises

Most core exercises address boundary value problems.

1. Consider a linear second-order BVP

$$c_0(x)y'' + c_1(x)y' + c_2(x)y = f(x), \qquad 0 \le x \le \ell,$$

$$a_0y'(0) + a_1y(0) = a_2, \qquad b_0y'(\ell) + b_1y(\ell) = b_2.$$

   Show that this problem has at most one solution if and only if the homogeneous problem—i.e., with $f(x)$, a_2 and b_2 all set to zero—admits only the zero solution.

2. In this problem you verify the behavior claimed for the BVP for (10.6). We begin with an inspired observation: if we substitute $y(x) = \cosh x$ into (10.6), this equation reduces to the identity $\cosh^2 x - \sinh^2 x = 1$.

   (a) Verify that for every choice of constants C, x_0, with $C \neq 0$, the function

$$y(x) = \frac{1}{C}\cosh[C(x - x_0)] \tag{10.60}$$

   satisfies (10.6).

Remarks: These solutions are constructed from $\cosh x$ using symmetries of (10.6). Can you identify the underlying symmetries? Incidentally, the curve (10.60) is called a *catenary*.

(b) Suppose that $b = a$ in (10.7). Show that (10.60) with $x_0 = \ell/2$ satisfies the boundary condition if and only if C is a solution of

$$\frac{1}{C}\cosh(C\ell/2) = a. \tag{10.61}$$

(c) *Introduction:* Multiplying (10.61) by $2/\ell$, you may rewrite the equation as

$$\frac{2}{C\ell}\cosh(C\ell/2) = \frac{2a}{\ell}. \tag{10.62}$$

Note that the LHS of the rewritten equation has the form $(1/\xi)\cosh\xi$, where $\xi = C\ell/2$.

Graph the function $(1/\xi)\cosh\xi$ to see that it has a positive minimum over $0 < \xi < \infty$, say m. Use calculus to show that

$$m = \min_{0<\xi<\infty} \frac{1}{\xi}\cosh\xi \approx 1.50888.$$

Deduce from (10.62) that if $b = a$, the BVP (10.6), (10.7) has (at least) two solutions if $\ell < 2a/m$ and none (of the above form) if $\ell > 2a/m$.

Challenge: Can you rule out the possibility that the BVP might have other solutions not of the form (10.60)?

3. Determine the midpoint deflection of a beam described by (10.9) with simply supported boundary conditions.

Hint: Equation (10.9) is linear and inhomogeneous. Look for a solution as the sum of a particular solution plus the general solution of the homogeneous equation. The calculations will be simpler if you use a translated coordinate, $\overline{x} = x - \ell/2$, because the beam is symmetric about its midpoint.

4. Consider a solution of an autonomous scalar second-order ODE,

$$y'' = F(y, y'), \qquad 0 < x < \ell,$$

where F is continuous on $\mathbb{R} \times \mathbb{R}$, subject to the boundary conditions

$$y(\ell) = y(0), \quad y'(\ell) = y'(0).$$

(I.e., assume that y'' exists and satisfies the ODE for $0 < x < \ell$ and that both y and y' are continuous on the closed interval $[0, \ell]$.) Show that the periodic extension of such a function to $-\infty < x < \infty$ is $\mathcal{C}^2$ and satisfies the ODE for all x.

5. *Introduction:* In this exercise you construct the solution to the hanging-chain problem, provided that, as required by (10.12), $L > \ell$.

   (a) One family of solutions of (10.10) is $y(x) = A + Bx$, where A, B are constants. Show that every other solution can be written in the form

   $$y(x) = A + \frac{1}{C}\cosh[C(x - x_0)], \tag{10.63}$$

   where A, C, x_0 are constants with $C \neq 0$.

   Hint: Observe from (10.10) that

   $$\frac{y''}{\sqrt{1 + (y')^2}} = C. \tag{10.64}$$

   Argue that you may assume $C \neq 0$. Substitute $z = y'$ into (10.64), solve the resulting first-order separable equation, and derive (10.63) from your solution.

   (b) Assuming $b = a$, deduce from (10.63) and (10.11) that $x_0 = \ell/2$.

   (c) Argue from (10.12) that C must satisfy

   $$\frac{2}{C}\sinh(C\ell/2) = L.$$

   Show that this equation has a unique positive root C, *provided* $L > \ell$.

   (d) Determine the additive constant A from (10.11).

6. Show that $\lambda = 0$ is an eigenvalue of (10.13) if and only if $\ell = (\ln 3)/2$.

   Hint: It suffices to seek eigenvalues such that $\lambda > -1$, so that the general solution of the ODE in (10.13) can be written

   $$y(x) = C_1 \cosh \mu x + C_2 \sinh \mu x,$$

   where $\mu = \sqrt{1 + \lambda}$. Argue that for the solution to satisfy the boundary conditions and not to vanish identically, μ must satisfy

   $$\tanh \mu\ell = \mu/2. \tag{10.65}$$

   Manipulate (10.65) to show that $\mu = 1$ is a root of the equation iff $\ell = (\ln 3)/2$.

   Remark: All the eigenvalues of (10.13) can be characterized as the roots of a transcendental equation similar to (10.65); see Sections 4.3 and 5.3 of [80].

7. *Introduction:* In the terminology of Section 10.6.1, the point $x \in [0, 1]$ with itinerary $1, 2, 1, 2, \ldots$ under the map (10.47) is a period-2 point for Ψ; i.e., it satisfies

   $$\Psi^2(x) = x. \tag{10.66}$$

(a) Write out an equation for (10.66). Two solutions of the resulting quartic equation are $x = 0$ and $x = 3/4$, which are fixed points of Ψ. Factor these out and show that the remaining quadratic equation has roots $x = (5 \pm \sqrt{5})/8$.

(b) Show that

$$\left|\Psi'((5-\sqrt{5})/8)\ \Psi'((5+\sqrt{5})/8)\right| = 4,$$

consistent with (10.52) for the period-2 orbit.

10.7.2 PHD Exercises

8. Show that (10.44) has infinitely many roots in the left half-plane.

Hint: One way to solve this problem is to invoke the principle of the argument from complex analysis. Although this calculation is long and only loosely related to ODE, you might find it interesting to see a nice application of complex analysis in an applied problem. Define the entire function

$$f(\lambda) = ae^{-\tau\lambda} + \lambda. \tag{10.67}$$

According to the principle of the argument (for example, see Section 86 of [11]), if Γ is a closed contour in $\mathbb{C}$ on which $f(\lambda)$ is nonvanishing, then the number of zeros, say N, of $f(\lambda)$ inside Γ equals $(2\pi)^{-1}$ times the variation in $\arg f(\lambda)$ around Γ; in symbols

$$N = \frac{1}{2\pi}\Delta_\Gamma \arg f(\lambda).$$

Apply this principle with Γ equal to the boundary of a large rectangle

$$R = \{\lambda \in \mathbb{C} : -M \le \Re\lambda \le 0,\ |\Im\lambda| \le 2\pi K/\tau\}, \tag{10.68}$$

where K is an integer and M is a large constant. Label the four sides of ∂R as in Figure 10.13. The strategy is to show that

$$\Delta_{\Gamma_1} \arg f(\lambda) \approx 4\pi K \quad \text{and} \quad \Delta_{\Gamma_k} \arg f(\lambda) = \mathcal{O}(1), \quad k = 2, 3, 4, \tag{10.69}$$

where $\mathcal{O}(1)$ denotes a quantity that is bounded independent of K. It follows from (10.69) that the number of zeros of $f(\lambda)$ inside R is at least $2K$ minus a constant, and the desired result may be obtained by letting $K \to \infty$.

How to justify (10.69)? Along Γ_1, the exponential has modulus $|e^{-\tau\lambda}| = e^{\tau M}$. Provided M is chosen large enough, the exponential term swamps the linear term, so $f(\lambda)$ is nonvanishing and

$$\Delta_{\Gamma_1} \arg f(\lambda) \approx \Delta_{\Gamma_1} \arg e^{-\tau\lambda} = 4\pi K.$$

Regarding Γ_2 and Γ_4, since K is an integer, along these sides the exponential in (10.67) is real and positive, so $f(\lambda)$ is nonvanishing and the variation of $\arg f(\lambda)$ is less than π. Divide Γ_3 into three ranges,

$$-2\pi K/\tau \le \Im\lambda \le -2a, \qquad -2a \le \Im\lambda \le 2a, \quad 2a \le \Im\lambda \le 2\pi K/\tau.$$

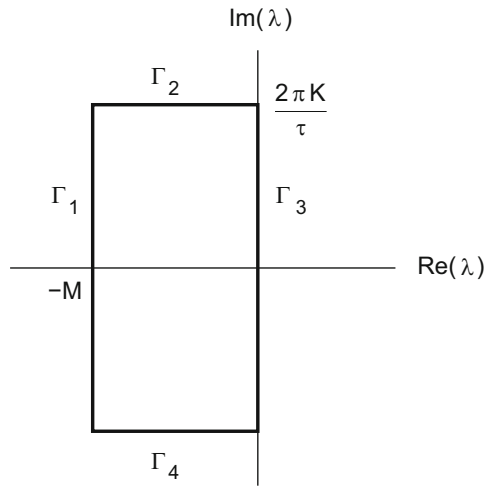

Figure 10.13: *Indexing of the sides of the rectangle* (10.68).

Since λ is pure imaginary along Γ_3, the exponential has modulus 1; thus, in the first and third ranges the linear term λ dominates $ae^{-\tau\lambda}$, so the variation in $\arg f(\lambda)$ along each of these is less than π. For the middle range, depending on a, it might be difficult to derive an explicit bound, but in any case the variation in $\arg f(\lambda)$ along this portion[24] is bounded independent of K.

Remark: With the principle of the argument you can also count how the number of zeros of $f(\lambda)$ in the right half-plane varies with τ. In particular, if $\tau > \pi/2a$, there are at least two such zeros, so generically solutions of (10.43) exhibit exponential growth. Thus, we have a more complete argument for the behavior claimed in Section 10.5.

9. Consider the following modification of the repressilator equations (8.74) in which the decay terms now include a time delay $\tau > 0$:

$$\begin{aligned} x'(t) &= \frac{\mu}{1+y^4} - x(t-\tau), \\ y'(t) &= \frac{\mu}{1+z^4} - y(t-\tau), \\ z'(t) &= \frac{\mu}{1+x^4} - z(t-\tau). \end{aligned} \tag{10.70}$$

Use numerical simulations to show that if $\mu = 1.8$ and $\tau = 0.5$, the system (10.70) has a stable limit cycle. Then, increasing the delay τ quasistatically, compute that a Hopf-like bifurcation to an invariant torus occurs.

Remark: Take a moment to contrast this behavior with that of (8.74), the repressilator without delay. Referring to Section 8.7.2, if $\tau = 0$ and $\mu = 1.8 < 2$,

[24] A minor technicality: depending on a, it might happen that $f(\lambda)$ has zeros along this portion of Γ_3. If so, we may deform the middle portion of Γ_3 to avoid these zeros, and the variation is still $\mathcal{O}(1)$.

solutions tend to the stable equilibrium $x = y = z = x_{\text{eq}}(\mu)$. Hence, this exercise supplies evidence of a delay-induced instability: an equilibrium that is stable for small nonnegative τ loses stability at some $\tau = \tau_{\text{bif}} > 0$. A stable limit cycle exists for τ slightly larger than τ_{bif}, but evidently that limit cycle suffers a loss of stability once τ crosses some other threshold $\tau = \tau_{\text{torus}} > \tau_{\text{bif}}$ in a bifurcation to an invariant torus.

10. *Introduction:* In the elastica (10.71), which is considered in the Pearls, s is arc length along the beam and $\theta(s)$ is the angle between the tangent to the beam and the x-axis. Thus, by calculus, the height $y(s)$ of the point a distance s along the beam is given by

$$y(s) = y(0) + \int_0^s \sin(\theta(s'))\, ds'.$$

Deduce from (10.71) that provided $\mu > 0$, the two ends of the beam are at the same height; i.e., $y(0) = y(\pi)$.

10.8 Pearls of Wisdom

10.8.1 The Elastica

Another nonlinear BVP, the *elastica*, is the granddaddy of all bifurcation problems—Euler [25] introduced it in 1744. This problem, which can be formulated variationally, models the buckling of an elastic beam under a compressive load, as sketched in Figure 10.14. After scaling, it may be written (cf. Section VII.2(a) of [31] or Section 12.1 of [92])

$$\frac{d^2\theta}{ds^2} + \mu \sin\theta = 0, \qquad \theta'(0) = \theta'(\pi) = 0, \tag{10.71}$$

where $\mu > 0$ is the scaled load. The variables in (10.71) are somewhat subtle: s is arc length along the beam, and $\theta(s)$ is the angle between the tangent to the beam and the x-axis (i.e., $\arctan(dy/dx)$) at the indicated position. For all μ, $\theta(s) \equiv 0$ is a solution of (10.71). If $0 < \mu < 1$, this solution is unique. However, as μ crosses 1, new solutions of (10.71) appear through a pitchfork bifurcation. Note that the linearized operator $L[\theta] = \theta'' + \mu\theta$ derived from the equation has a zero eigenvalue (with null eigenfunction $\cos s$) precisely when $\mu = 1$; this is the infinite-dimensional analogue of our observation in Chapter 8 that an equilibrium of an ODE can bifurcate only when the linearization of the ODE is singular. As indicated in Figure 10.14(c), the bifurcating solutions of (10.71) correspond to buckled states of the beam.[25]

[25]It is not obvious, but the boundary conditions in (10.71) imply that the ends of the beam are at the same height, as indicated in the figure; see Exercise 10.

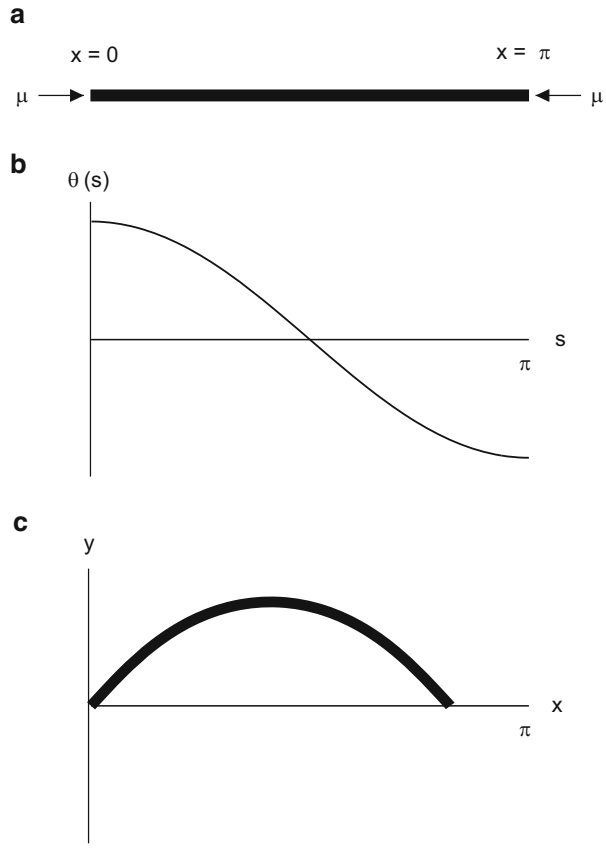

Figure 10.14: *(a) A schematic diagram of the elastica modeled by* (10.71)*: buckling of a beam under a compressive load. (b) The null eigenfunction* $\theta(s) = \cos s$ *of the linearization of* (10.71) *when* $\mu = 1$*. (c) A buckled configuration in the physical* x, y*-plane with* $\mu = 1.1$*. The reflected function* $-y(x)$ *is also a solution.*

10.8.2 A Bit More on Chaos

Here is some minimal guidance for supplementing the inadequate treatment of chaos in this book. Part III of Strogatz [81] and Devaney [17] give readable introductions to chaos that complement each other. Guckenheimer & Holmes [33] is one of many good sources for more advanced theory; in particular, Chapter 2 introduces, with a minimum of fuss, four elementary examples that display chaos. Fourier analysis provides a different, invaluable, perspective on chaos; Chapter III of [8] and Section 8.5 of [18] give relatively compact discussions of the key ideas as they apply to dynamics.[26]

Rather than attempt to survey the vast chaos literature, let us mention just one specific result, due to Wiggins [94], that may help entice you to pursue the subject. Really, this gem, which identifies symbol-like behavior in an ODE, takes our breath away. It concerns a vibrated pendulum,

$$x'' + \varepsilon\delta x' + (1 + \gamma \sin \varepsilon t) \sin x = 0, \tag{10.72}$$

where $\varepsilon > 0$ is small, meaning that the vibration is slow and damping is correspondingly weak. In dimensional terms, the hypothesis that $|\gamma| < 1$ means that the acceleration of the pivot is always less than that of gravity; the sign of γ doesn't matter, since it can be flipped by a translation of time by π/ε. A proof of this result, which is surprisingly readable in case $\delta = 0$, is given in [39].

Theorem 10.8.1. *If $0 < |\gamma| < 1$ and $0 \le \delta < |\gamma|/2$, then there is a positive constant ε_1 with the following property: for every ε with $0 < \varepsilon < \varepsilon_1$ and for every infinite sequence $\{m_1, m_2, m_3, \ldots\}$ of positive integers, there is a solution of (10.72) in which the pendulum makes exactly m_1 complete clockwise rotations, followed by exactly m_2 complete counterclockwise rotations, followed by exactly m_3 complete clockwise rotations, etc.*

Note that oscillations not involving a complete rotation may be part of the above solution.

Call us naive, but ... We hope that as you finish this book, even after 500 pages, you are at least a little sorry that it's over.

[26] The use of Fourier series in PDEs is covered, for example, in [80].

Appendix A

Guide to Commonly Used Notation

A.1 Letter Choices

In a book of this size, it is impossible to avoid using the same letter to represent different types of mathematical objects in different contexts. Nevertheless, a lot of information is conveyed by the choice of notation in a formula. Although not every use of a letter is covered below, we indicate our more common usages.

Lower case Latin letters. Letters before u in the alphabet usually indicate a parameter; letters after this may indicate a variable. Vector quantities are set in **boldface** (even $\mathbf{0}$ for the zero vector). Some specific recurring notation:

- $\mathbf{b}$: an initial condition (normal type b is also used for a coefficient of friction).
- d: the dimension of a vector.
- $g(x,t)$: the reduced function resulting from the Lyapunov–Schmidt reduction of a bifurcation problem (in Chapter 8).
- j,k,n denote integers to enumerate something. The letter j often enumerates the components of a vector or entries of a matrix, so j runs between 1 and d. The letter k is used to enumerate other types of lists, so its range depends on context. (But k also denotes a spring constant.) The letter n is usually used to enumerate the entries of a sequence or series, in which case n runs from 1 to ∞.
- t: the independent variable in an ODE, usually thought of as time.

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8

- $\mathbf{v}, \mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z}$ may denote either a vector in $\mathbb{R}^d$ or a vector-valued function of t. Scalar versions of these mathematical objects are written with the same letters in regular type. Typically $\mathbf{w}$ is used as the unknown in the linearization of a nonlinear equation at an equilibrium.

Uppercase Latin letters. Some specific recurring notation:

- B: used for an *open* ball; thus $B(\mathbf{b}, \delta) = \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x} - \mathbf{b}| < \delta\}$.
- C, K, M: constants. The letter C is used for the constants of integration in solving an ODE; K and M are typically large constants; M may be related to the maximum of a continuous function over a compact set.
- $\mathbf{D}$: indicates derivatives in the Jacobian of a vector-valued function, $\mathbf{DF}(\mathbf{x})$.
- E, H, V: energy in one form or another. The letter H is used for the energy in Hamiltonian systems once these are introduced in Exercise 6.10; V denotes potential energy.
- $\mathbf{F}$: a vector-valued function, typically the RHS in the generic ODE $\mathbf{x}' = \mathbf{F}(\mathbf{x})$.
- L: a Lipschitz constant (Chapter 3), a length scale (Chapter 5), or a Lyapunov function (Chapter 6).
- N: typically a large integer, but also a nilpotent matrix (Chapter 2) or, if boldface $\mathbf{N}$, the normal to a surface in $\mathbb{R}^d$.
- T: generally a constant measuring time. In Chapter 5 it sets the time scale in a nondimensionalization, and starting in Chapter 7 it denotes the period of a periodic solution. (As a superscript, T indicates the transpose of a matrix.)

Lowercase Greek letters. Most dimensionless parameters are written as lowercase Greek letters. Some specific recurring notation:

- α, β: the endpoints of the time interval over which an ODE has a solution.
- $\boldsymbol{\gamma}(t)$: a periodic solution of an ODE (starting in Chapter 7). (This is boldface, indicating a vector quantity, although you need to look closely to see this.)
- $\delta, \varepsilon, \eta$: this notation hints that constants are small.
- λ: an eigenvalue.
- μ: starting in Chapter 8, a bifurcation parameter.

- $\boldsymbol{\varphi}(t, \mathbf{b})$: the flow map, the solution of an IVP with initial condition $\mathbf{b}$.
- ϕ, ψ: auxiliary scalar functions.
- τ: scaled time in a perturbation-theory calculation (starting in Chapter 7) or the time of first return in defining the Poincaré map (also starting in Chapter 7).
- ω: besides a frequency, an adjective indicating $t \to +\infty$, as in the ω-limit of a trajectory.

Uppercase Greek letters. Some specific recurring notation:

- Γ: the orbit traced out by a periodic solution (starting in Chapter 7).
- $\boldsymbol{\Pi}$: a Poincaré map (starting in Chapter 7).
- Σ: a small piece of surface used in defining a Poincaré map.
- $\boldsymbol{\Psi}$: a mapping between Euclidean spaces (starting in Chapter 7).

Letters in different fonts. Some specific recurring notation:

- $\mathcal{C}$: the space of continuous functions of time, or $\mathcal{C}^k$, k-times differentiable functions.
- $\mathcal{K}$: a closed subset of $\mathbb{R}^d$, typically a trapping region.
- $\mathcal{M}$: a stable or unstable manifold (starting in Chapter 6).
- $\mathcal{O}, o$: the order notation, introduced in Section 4.6.4. (You need to distinguish between the formal and informal uses of $\mathcal{O}$.)
- $\Re, \Im$: real and imaginary parts of a complex number.
- $\boldsymbol{\mathfrak{T}}$: a mapping on a Banach space of functions (Chapter 3).
- $\mathcal{U}$: an *open* subset of $\mathbb{R}^d$, typically the domain on which an ODE is defined.
- $\mathfrak{U}$: In Chapter 5 a Gothic U indicates the units of a parameter or variable.
- X: sans serif letters (S and X) are used in Chapter 3 to indicate objects related to a Banach space; they are used in Chapter 5 to indicate chemicals.

A.2 Other Notations

Note 1. Many authors, us included, are sloppy in notation regarding functions. For example, suppose $f : (a, b) \to \mathbb{R}$. Informally, one may say "consider the function $f(t)$, where $a < t < b$." However, strictly speaking, the notation $f(t)$ denotes a single real number, the number that this function produces when evaluated at $t \in (a, b)$. When speaking of the *function*, one should write simply f. Nevertheless, in circumstances where such precision seems fussy, we will stick with the more common informal usage.

A dot may be used to indicate an unwritten argument of a function; thus, f and $f(\cdot)$ mean exactly the same thing. This notation is sometimes useful in dealing with functions of more than one variable. For example, $\boldsymbol{\varphi}(t, \cdot)$ indicates the function of one argument $\mathbf{b} \mapsto \boldsymbol{\varphi}(t, \mathbf{b})$, where t is fixed, which is subtly different from either $\boldsymbol{\varphi}$ (the function of both arguments) or $\boldsymbol{\varphi}(t, \mathbf{b})$ (the quantity that evaluation of $\boldsymbol{\varphi}$ produces).

Note 2. In Chapter 5, a hat over a quantity has a special meaning: it indicates that the quantity has nontrivial dimensions. For example, $\hat{t}$ indicates time measured in minutes, years, nanoseconds, or some such physical units. By contrast, if time is scaled $t = \hat{t}/T$, where T is some time scale extracted from the parameters in an ODE, then the absence of a hat over t indicates that it is dimensionless.

Note 3. Regarding derivatives, prime denotes a derivative with respect to a scalar argument, typically time as in $\mathbf{x}'(t)$; $\nabla f(\mathbf{x})$ denotes the gradient of a scalar function with respect to a vector argument; and $\mathbf{DF}(\mathbf{x})$ denotes the Jacobian matrix of derivatives of a vector-valued function with respect to all the coordinates in its vector argument. If $\mathbf{F}(\mathbf{x}, \mu)$ depends an a parameter as well as on $\mathbf{x}$, then $\mathbf{DF}$ indicates derivatives *only* with respect to $\mathbf{x}$. Similarly, for the flow map, $\mathbf{D}\boldsymbol{\varphi}(t, \mathbf{b})$ indicates derivatives with respect to $\mathbf{b}$, not t.

Miscellaneous.

- $\mathrm{Col}(\mathbf{v}_1, \ldots, \mathbf{v}_m)$ indicates the matrix with columns $\mathbf{v}_1, \ldots, \mathbf{v}_m$. If these vectors have d components, this notation indicates a $d \times m$ matrix.
- $\mathrm{Diag}(\lambda_1, \ldots, \lambda_d)$ denotes the $d \times d$ diagonal matrix with the indicated entries. This notation is also generalized to specifying blocks on the diagonal of a block-diagonal matrix.
- In Chapter 4 and elsewhere, to focus attention on one component of a d-dimensional coordinate, we write $\mathbf{x} = (x_1, \tilde{\mathbf{x}})$, where $\tilde{\mathbf{x}}$ is shorthand for the remaining coordinates $(x_2, \ldots, x_d)$.
- Norms are indicated by single bars for vectors as $|\mathbf{x}|$, double bars for matrices as $\|A\|$. Double bars are also used for norms in a function space (Chapter 3).

- Restriction of a function to a set smaller than its domain is written with a vertical bar, as for example in $\mathbf{F}|\mathcal{K}$.
- A subscript $*$ indicates a special value for a quantity, such as writing $\mathbf{b}_*$ for an equilibrium point. We also write the shorthand $\mathbf{DF}_*$ for $\mathbf{DF}(\mathbf{b}_*)$. Starting in Chapter 8, μ_* may indicate the value of μ at which bifurcation occurs.
- Complex conjugates are indicated by an overline: thus if $z = x + iy$, where $x, y \in \mathbb{R}$, then $\overline{z} = x - iy$.

A.3 Other Conventions

Abbreviations.

- The phrase "if and only if" is sometimes shortened to "iff."
- "RHS" and "LHS" indicate the "right-hand side" and "left-hand side" of an equation.

Color conventions in figures.

- *Brown* and *cyan* are used for x-nullclines and y-nullclines of a two-dimensional ODE, respectively.
- *Red* and *blue* are used for unstable and stable manifolds, respectively. *Purple* indicates that a manifold is unstable with respect to one saddle point and stable with respect to a second (or possibly the same) saddle. Purple is also used for some global center manifolds.
- *Green* is used for periodic orbits.

Organizational units. Each of the 10 chapters is divided into sections; thus, for example, Section 4.6 refers to the sixth section of Chapter 4. The sections are further divided into subsections. When referring to a specific subsection by number, we write, for example, simply Section 4.6.2 rather than the more cumbersome Subsection 4.6.2. By contrast, we distinguish between "the next section" and "the next subsection."

The phrase Exercise n, with a single number, refers to the nth exercise in the current chapter. Exercise $m.n$, with a double number, refers to an exercise in the mth chapter.

Inconsistent use of the plural. Sometimes we regard a system of ODEs as a single vector equation and call it an ODE (in the singular), but with no consistency. Don't waste your time trying to figure out what underlies our choices; we can't figure this out ourselves.

Appendix B

Notions from Advanced Calculus

In this appendix we recall selected parts of advanced calculus. We make no pretense of completeness. We assume that you have taken an analysis course that deals rigorously with calculus, and we hope to jog your memory more than anything else. In particular, you will need to consult a reference for most proofs. Whatever book you used in a course is probably the most familiar, and therefore the most convenient, reference. Alternative texts are by Rudin [68], Reed [66], and Marsden and Hoffman [51].

B.1 Basic Issues

(a) Supremum and infimum. If $E \subset \mathbb{R}$, a number b is called an *upper bound* for E if $x \leq b$ for all $x \in E$. We say that b is the *supremum* of E if b is the *least upper bound* for E in the sense that (i) b is an upper bound for E, and (ii) if $\tilde{b}$ is an upper bound for E, then $b \leq \tilde{b}$. The *infimum* of E is defined analogously as the greatest lower bound for E. The supremum and infimum of $E \subset \mathbb{R}$ are denoted by $\sup(E)$ and $\inf(E)$, respectively. If, for example, $E = \{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\} \subset \mathbb{R}$, then $\sup(E) = 1$ and $\inf(E) = 0$. Notice that in this example, $\sup(E) \in E$, whereas $\inf(E) \notin E$. When a set E contains its supremum, we typically write $\sup(E) = \max(E)$, the maximum of the set E. Likewise for infimum and minimum. If $a < b$, then the open interval (a, b) and the closed interval $[a, b]$ have the same infimum and supremum, whereas only the closed interval has a maximum and a minimum.

(b) Compactness. Suppose E is a subset of $\mathbb{R}^d$. An *open cover* of E is any collection $\{\Omega_\alpha\}_{\alpha \in I}$ of open subsets of $\mathbb{R}^d$ with the property that

$$E \subset \bigcup_{\alpha \in I} \Omega_\alpha.$$

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8

This union of sets need not be countable, and for that reason we enumerate the sets using an arbitrary index set I. The set E is called *compact* if every open cover of E has a *finite subcover*. That is, E is compact if from every open cover of E, we may extract[1] *finitely* many open sets whose union still contains E.

Applying this general definition directly is utterly hopeless. Fortunately, we have the following, infinitely more workable, criterion.

Theorem B.1.1. *(Heine–Borel) A subset $E \subset \mathbb{R}^d$ is compact if and only if E is closed and bounded.*

The same definition of compactness applies in every topological space, but in general we do not have the luxury of the Heine–Borel theorem: in every metric space, a compact set always is always closed and bounded, but *not conversely.* (Cf. Exercise 10.)

Many of the technical proofs throughout this text require us to estimate functions or their derivatives over some given subset of $\mathbb{R}^d$. For continuous functions[2] over a compact set, the following two theorems facilitate such estimates. For the record, the notation $|\mathbf{a}|$ for $\sqrt{\sum_1^d a_k^2}$, the length of a vector in $\mathbb{R}^d$, is introduced and discussed in Section 2.2.1.

Theorem B.1.2. *(Extreme value theorem) Suppose that $E \subset \mathbb{R}^{d_1}$ is compact and $\mathbf{F} : E \to \mathbb{R}^{d_2}$ is continuous. Then there exist points $\mathbf{x}_{\min}, \mathbf{x}_{\max} \in E$ such that*

$$|\mathbf{F}(\mathbf{x}_{\min})| = \inf_{\mathbf{x}\in E} |\mathbf{F}(\mathbf{x})| \qquad \text{and} \qquad |\mathbf{F}(\mathbf{x}_{\max})| = \sup_{\mathbf{x}\in E} |\mathbf{F}(\mathbf{x})|.$$

In particular, both the infimum and supremum are finite.

Incidentally, the above result for vector-valued functions may be reduced to the corresponding result for scalar-valued functions by considering the composition $g \circ \mathbf{F}$, where $g(\mathbf{y}) = |\mathbf{y}|$.

Theorem B.1.3. *Suppose that $E \subset \mathbb{R}^{d_1}$ is compact and $\mathbf{F} : E \to \mathbb{R}^{d_2}$ is continuous. Then the image $F(E)$ is a compact subset of $\mathbb{R}^{d_2}$.*

Phrased more compactly (bad pun intended, sorry), images of compact sets under continuous functions are compact.

(c) One more definition. A subset E of a topological space is called *dense* if its closure equals the entire space. For example, the rational numbers are dense in $\mathbb{R}$.

[1] It might be more honest to say "God could extract ...." Although the definition merely asserts the existence of a certain finite subcollection of sets, mathematicians seem to feel more comfortable phrasing this as though we had the superhuman power to pick out the necessary sets from an arbitrary collection.

[2] We expect that you are familiar with the concept of continuity. For a refresher, see [68].

B.2 Pointwise and Uniform Convergence

B.2.1 Sequences

(a) Definitions. Convergence of a sequence of constants is a fairly dry affair. Suppose $\{\mathbf{a}_n\}$, $n = 1, 2, 3, \dots$ is a sequence of points in $\mathbb{R}^d$. We say that $\{\mathbf{a}_n\}$ *converges to a limit* $\mathbf{L}$ if for every $\varepsilon > 0$, there exists an integer[3] $N = N(\varepsilon)$ such that $|\mathbf{a}_n - \mathbf{L}| < \varepsilon$ whenever $n \geq N$.

By contrast, convergence of sequences of *functions* gives rise to a wealth of interesting phenomena. (For most of our discussion below we consider scalar-valued functions, but the extension to vector-valued functions poses no difficulties.) Two notions of convergence for sequences of functions are relevant for this text. Suppose that $\{f_n,\ n = 1, 2, 3, \dots\}$ is a sequence of functions defined over some set $E \subset \mathbb{R}^d$. We say that $\{f_n\}$ converges

- *pointwise* on E to a function f if for every $\varepsilon > 0$ and for every $\mathbf{x} \in E$, there exists an integer $N = N(\mathbf{x}, \varepsilon)$ such that $|f_n(\mathbf{x}) - f(\mathbf{x})| < \varepsilon$ whenever $n \geq N$;
- *uniformly* on E to a function f if for every $\varepsilon > 0$, there exists an integer $N = N(\varepsilon)$ such that for all $\mathbf{x} \in E$, $|f_n(\mathbf{x}) - f(\mathbf{x})| < \varepsilon$ whenever $n \geq N$.

On a formal level, the difference between these two definitions is obvious: in pointwise convergence, the integer N can depend upon both ε and x, whereas in uniform convergence, N depends only upon ε. That is, for uniform convergence, the *same* integer N has to work for all $x \in E$. (Of course, uniform convergence implies pointwise convergence.) To make this difference more concrete, consider, for example, the functions $f_n : \mathbb{R} \to \mathbb{R}$ defined by

$$f_n(x) \ = \ \frac{2}{\pi} \arctan(nx), \qquad n = 1, 2, 3, \dots . \tag{B.1}$$

(To understand such an example, it is most effective to draw a graph. In this case, we have done it for you in Figure B.1.)

Claim B.2.1. *This sequence converges pointwise to the step function*

$$f(x) \ = \ \begin{cases} 1 & \text{if } 0 < x < \infty, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } -\infty < x < 0. \end{cases} \tag{B.2}$$

[3] An incidental, but hopefully instructive, comment: Sometimes in definitions of this sort one appends a restriction like $N > 0$. This is necessary, for example, in the usual ε, δ definition of the continuity of a function of a real variable, i.e., for every $\varepsilon > 0$, there is a $\delta > 0$ such that $|x - x_0| < \delta$ implies that $|f(x) - f(x_0)| < \varepsilon$. The issue is that if δ were chosen negative, the implication would be valid by virtue of its hypothesis never being satisfied. In the present case, no such restriction on N is needed to avoid trivialities.

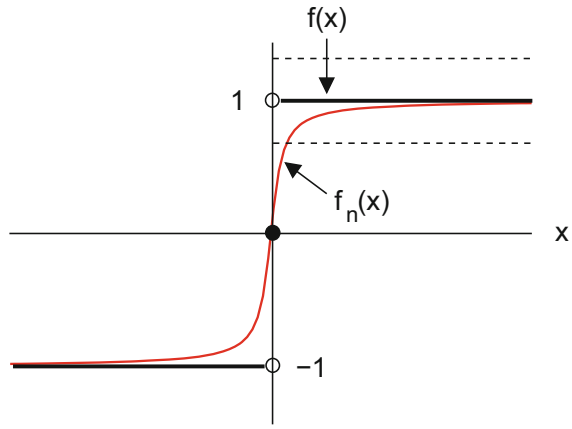

Figure B.1: *Pointwise but nonuniform convergence of the sequence* (B.1), $f_n(x) = \frac{2}{\pi}\arctan(nx)$.

Proof. Since $f_n(0) = 0$, the middle case is obvious. To attack the first case, suppose $x > 0$ and let $\varepsilon > 0$ be given. We must produce an integer $N = N(x, \varepsilon)$ for which

$$|f_n(x) - f(x)| = |(2/\pi)\arctan(nx) - 1| < \varepsilon \qquad \text{for all } n \geq N. \tag{B.3}$$

Now

$$\lim_{y\to\infty} \arctan y = \pi/2,$$

which means that there is a number $Y(\varepsilon)$ such that $|\arctan y - \pi/2| < \varepsilon$ if $y > Y(\varepsilon)$. Since $2/\pi < 1$, we may satisfy (B.3) by choosing N to be any integer such that $Nx > Y(\varepsilon)$.

Pointwise convergence in the third case, $x < 0$, may be proved effortlessly by invoking symmetry. □

The convergence of (B.1) is not uniform, and Figure B.1 illustrates why. The limit function is graphed in black, and the dotted lines show an ε-neighborhood, with $\varepsilon \approx 1/3$, of this graph over the positive real axis. If the convergence were uniform, then for sufficiently large n, the graph of $f_n(x)$, which is shown in red, would need to stay between the dotted lines. This is clearly inconsistent with the facts that $f_n(0) = 0$ and $f_n(x)$ is continuous.

(b) Regularity of limit functions. The limit function (B.2) is discontinuous even though the functions $f_n(x)$ are continuous. The next theorem, proved in [68], ensures that this cannot happen for *uniformly* convergent sequences of continuous functions.

Theorem B.2.2. *If a sequence of continuous (scalar- or vector-valued) functions* $\{f_n(\mathbf{x})\}$ *converges uniformly to* $f(\mathbf{x})$ *on a set* $E \subset \mathbb{R}^d$, *then* $f(\mathbf{x})$ *is continuous on* E.

	$\{f_n'\}$ converges	$\{f_n'\}$ diverges
$\lim_{n\to\infty} f_n(x)$ not differentiable	Exercise 4	Exercise 5
$\lim_{n\to\infty} f_n(x)$ differentiable	Exercise 6	Exercise 7

Table B.1: *Examples of uniformly convergent sequences of smooth functions with varying behavior of derivatives.*

What about derivatives of the limit of a sequence of functions? For our purposes it suffices to address this question for functions of one variable. But note: even if every function $f_n(x)$ in a sequence is differentiable and if $\{f_n\}$ is uniformly convergent, $\lim_n f_n(x)$ may or may not be differentiable, and the sequence of derivatives $\{f_n'\}$ may or may not converge. Fairly dramatic examples of each behavior are given in the exercises cited in Table B.1.

However, in studying ODEs, we need control over the derivative of the limit function. The following result for (scalar- or vector-valued) functions of one variable, which is proved in [68], will do the job. We say that a function $f(x)$ on an interval is $\mathcal{C}^1$ if it is everywhere differentiable and both f and f' are continuous on the interval. (In case of a closed interval, at the endpoints f' means the one-sided derivative.)[4]

Theorem B.2.3. *If $\{f_n\}$ is a sequence of $\mathcal{C}^1$ functions on $[a,b]$ such that both $\{f_n\}$ and $\{f_n'\}$ converge uniformly, then the limit function $f(x)$ is $\mathcal{C}^1$ on $[a,b]$ and*

$$f'(x) = \lim_{n\to\infty} f_n'(x) \qquad (a \le x \le b).$$

In point of fact, the hypotheses in the preceding theorem can be weakened slightly: it suffices to assume that $\{f_n'\}$ converges uniformly and that for one point $x_0 \in [a,b]$, the sequence (of constants) $f_n(x_0)$ converges. Some hypothesis on $\{f_n\}$ is needed, because otherwise, the functions could march off to infinity (e.g., $f_n(x) \equiv n$). In practice, these weaker hypotheses are no easier to verify.

B.2.2 Series

There is a natural one-to-one correspondence between infinite sequences and infinite series. To a series $\sum_{n=1}^{\infty} \mathbf{a}_n$ we associate the sequence of partial sums $\mathbf{S}_m = \sum_{n=1}^{m} \mathbf{a}_n$; and for a sequence $\{\mathbf{b}_n\} \subset \mathbb{R}^d$ we associate the telescoping series

$$\mathbf{b}_1 + (\mathbf{b}_2 - \mathbf{b}_1) + (\mathbf{b}_3 - \mathbf{b}_2) + (\mathbf{b}_4 - \mathbf{b}_3) + \dots,$$

[4]Alternatively, one can define a $\mathcal{C}^1$ function on $[a,b]$ by requiring that $f'(x)$, defined by the usual difference quotient for $x \in (a,b)$, extend to a continuous function on $[a,b]$.

which has $\{\mathbf{b}_n\}$ as its sequence of partial sums. Despite the overlap, it is convenient to have separately formulated results for both cases.

(a) Series of constants. If $\mathbf{a}_n \in \mathbb{R}^d$, $n = 1, 2, 3, \ldots$, we say that the infinite series $\sum_{n=1}^{\infty} \mathbf{a}_n$ *converges* if the sequence of partial sums

$$\mathbf{S}_m = \sum_{n=1}^{m} \mathbf{a}_n$$

converges as $m \to \infty$. We say that an infinite series *converges absolutely* if the series $\sum_{n=1}^{\infty} |\mathbf{a}_n|$, whose terms are nonnegative scalars, converges (to a finite limit).

Proposition B.2.4. *If the series $\sum_{n=1}^{\infty} \mathbf{a}_n$ converges absolutely, then the series converges.*

This result is proved in [68]. The implication is in one direction only; for example,

$$\sum_{n=1}^{\infty} (-1)^n / n \tag{B.4}$$

is convergent but not absolutely convergent. (Some counterintuitive behavior of such conditionally convergent series is discussed in the Pearls.)

(b) Series of functions. When dealing with a series of functions $\sum_{n=1}^{\infty} f_n(x)$, we use the correspondence between infinite sequences and infinite series to define *pointwise* and *uniform convergence* of a series. Let us record two results for sequences that carry over to series.

Corollary B.2.5. *If a series $\sum_{n=1}^{\infty} f_n(\mathbf{x})$ of continuous functions converges uniformly to $S(\mathbf{x})$ on a set $E \subset \mathbb{R}^d$, then $S(\mathbf{x})$ is continuous on E.*

Corollary B.2.6. *Suppose that $\sum_{n=1}^{\infty} f_n(x)$ and $\sum_{n=1}^{\infty} f_n'(x)$ converge uniformly on the interval $[a, b]$, where each term $f_n(x)$ is continuously differentiable. Then the sum $S(x)$ is continuously differentiable, and $S'(x)$ equals the sum of the derivatives: in symbols,*

$$\left(\sum_{n=1}^{\infty} f_n(x) \right)' = \sum_{n=1}^{\infty} f_n'(x). \tag{B.5}$$

On occasion, we will need to differentiate under an integral. We include a theorem justifying this operation here because the result is loosely analogous to Corollary B.2.6—the integral is a continuous analogue of a sum, the limit of Riemann sums.

Theorem B.2.7. *If $f(s,x)$ is continuously differentiable for (s,x) in a closed rectangle*[5] *$[a,b]\times[c,d]$, then*

$$\frac{\partial}{\partial x}\int_a^b f(s,x)ds = \int_a^b \frac{\partial f}{\partial x}(s,x)ds \qquad (c<x<d).$$

You may wonder why no analogue of the hypothesis of uniform convergence appears in the above result. We remind you that since $\partial_x f$ is continuous on a compact set, it is uniformly continuous. By contrast, for differentiation of an integral over an infinite range, a uniformity hypothesis is needed. See Section 9.7 of [51] for more details.

B.2.3 Convergence of Integrals

In profound contrast to derivatives, uniform convergence of a sequence or series of functions *does* give control of the integral of the limit (provided the range of integration is finite).

Theorem B.2.8. *Suppose that $\{f_n\}$ is a sequence of continuous functions on $[a,b]$ that converges uniformly to f on $[a,b]$. Then*

$$\int_a^b f(x)\,dx \;=\; \lim_{n\to\infty}\int_a^b f_n(x)\,dx.$$

The theorem is proved in [68]. The example in Exercise 8 shows that without uniform convergence the result may fail. There are more general results than Theorem B.2.8—the functions need not be continuous, only integrable, and a less restrictive notion of convergence suffices—but we have no need for such results.

To conclude, let's rephrase Theorem B.2.8 for series.

Corollary B.2.9. *Suppose that a sum $\sum_{n=1}^{\infty} f_n(x)$ of continuous functions on the interval $[a,b]$ converges uniformly on $[a,b]$. Then the order of integration and summation can be interchanged, i.e.,*

$$\int_a^b \left(\sum_{n=1}^{\infty} f_n(x)\right)\,dx \;=\; \sum_{n=1}^{\infty}\left(\int_a^b f_n(x)\,dx\right).$$

B.3 Selected Issues in Vector Calculus

One-variable calculus can be generalized to a vector context in two ways: to vector-valued functions of a single real variable, or to functions of several variables. The former generalization, which suffices for much of the study of ODEs, poses no great

[5] That is, f is continuous on $[a,b]\times[c,d]$ and continuously differentiable on $(a,b)\times(c,d)$, and both derivatives $\partial_s f$, $\partial_x f$ extend to continuous functions on $[a,b]\times[c,d]$.

challenges. Most results can be derived componentwise from their scalar counterparts. By contrast, calculus of functions of several variables introduces many new, subtle complications, such as divergence, curl, multiple integrals, and Green's theorem. Only a few ideas from the latter generalization, such as the chain rule, are needed for this book, and these are reviewed below.

B.3.1 Differentiability

(a) Definitions. Suppose $E \subset \mathbb{R}^{d_1}$ is open and that $\mathbf{F} : E \to \mathbb{R}^{d_2}$. We say that $\mathbf{F}$ is *differentiable* at $\mathbf{x} \in E$ if there exists a linear transformation $A : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ such that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{|\mathbf{F}(\mathbf{x}+\mathbf{h}) - \mathbf{F}(\mathbf{x}) - A\mathbf{h}|}{|\mathbf{h}|} \;=\; 0. \tag{B.6}$$

We refer to A as the *derivative* of $\mathbf{F}$ at $\mathbf{x}$, and we write $A = \mathbf{DF}(\mathbf{x})$.

Equivalently, in terms of the order notation of Section 4.6.4, $\mathbf{F}$ is differentiable at $\mathbf{x}$ if and only if there exists a linear transformation $\mathbf{DF}(\mathbf{x}) : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ such that for $\mathbf{h}$ in a neighborhood of $\mathbf{0}$,

$$\mathbf{F}(\mathbf{x}+\mathbf{h}) = \mathbf{F}(\mathbf{x}) + \mathbf{DF}(\mathbf{x})\,\mathbf{h} + \mathbf{r}(\mathbf{h}),$$

where the remainder $\mathbf{r}$ is "small" in the sense that $\mathbf{r}(\mathbf{h}) = o(|\mathbf{h}|)$.

A function $\mathbf{F} : E \subset \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ can be written out in components as

$$\mathbf{F}(x_1, x_2, \ldots, x_{d_1}) \;=\; \begin{bmatrix} F_1(x_1, x_2, \ldots, x_{d_1}) \\ F_2(x_1, x_2, \ldots, x_{d_1}) \\ \vdots \\ F_{d_2}(x_1, x_2, \ldots, x_{d_1}) \end{bmatrix}.$$

If $\mathbf{F}$ is differentiable on E, then the partial derivatives $\partial F_i/\partial x_j$, for $i = 1, 2, \ldots, d_2$ and $j = 1, 2, \ldots, d_1$, exist at all points in E. For every $\mathbf{x} \in E$, the derivative $\mathbf{DF}(\mathbf{x})$ has a compact representation as a $d_2 \times d_1$ matrix

$$\mathbf{DF}(\mathbf{x}) = \begin{bmatrix} \partial F_1/\partial x_1 & \partial F_1/\partial x_2 & \cdots & \partial F_1/\partial x_{d_1} \\ \partial F_2/\partial x_1 & \partial F_2/\partial x_2 & \cdots & \partial F_2/\partial x_{d_1} \\ \vdots & \vdots & \ddots & \vdots \\ \partial F_{d_2}/\partial x_1 & \partial F_{d_2}/\partial x_2 & \cdots & \partial F_{d_2}/\partial x_{d_1} \end{bmatrix} \tag{B.7}$$

with respect to the standard bases for $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$. This matrix, called the *Jacobian* of $\mathbf{F}$, defines a linear transformation from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$.

(b) Results about derivatives. Existence of all the partial derivatives in the matrix (B.7) is *not* enough to conclude that $\mathbf{F}$ is differentiable. Consider, for example, the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x_1, x_2) \;=\; \begin{cases} x_1 x_2/(x_1^2 + x_2^2) & \text{if } (x_1, x_2) \neq \mathbf{0}, \\ 0 & \text{if } (x_1, x_2) = \mathbf{0}\,. \end{cases} \tag{B.8}$$

Both $\partial f/\partial x_1$ and $\partial f/\partial x_2$ exist everywhere in the plane, but f is not even continuous at $(0, 0)$, let alone differentiable! (See Exercise 9.)

As the following theorem asserts, such pathologies can be avoided if the partial derivatives in the Jacobian matrix are continuous on E. The notation $\mathcal{C}^1$ designates a function whose partial derivatives are continuous on an (open) set.

Theorem B.3.1. *Suppose $E \subset \mathbb{R}^{d_1}$ is open and $\mathbf{F} : E \to \mathbb{R}^{d_2}$. If all partial derivatives $\partial F_i/\partial x_j$ in the Jacobian matrix (B.7) of $\mathbf{F}$ exist and are continuous on E, then $\mathbf{F}$ is differentiable everywhere in E.*

Compositions of differentiable functions are also differentiable, and the chain rule from single-variable calculus generalizes to a composition $\mathbf{H}(\mathbf{x}) = (\mathbf{G} \circ \mathbf{F})(\mathbf{x}) = \mathbf{G}(\mathbf{F}(\mathbf{x}))$, where $\mathbf{F}$ maps a neighborhood of a point $\mathbf{x}_0 \in \mathbb{R}^{d_1}$ into $\mathbb{R}^{d_2}$ and $\mathbf{G}$ maps a neighborhood of $\mathbf{F}(\mathbf{x}_0) \in \mathbb{R}^{d_2}$ into $\mathbb{R}^{d_3}$. The multidimensional chain rule may be written compactly

$$\mathbf{DH}(\mathbf{x}_0) \;=\; \mathbf{DG}(\mathbf{F}(\mathbf{x}_0)) \cdot \mathbf{DF}(\mathbf{x}_0), \tag{B.9}$$

where the dot indicates matrix multiplication of Jacobians. More formally, we could write $\mathbf{DG}(\mathbf{F}(\mathbf{x}_0)) \circ \mathbf{DF}(\mathbf{x}_0)$ for the composition of linear transformations.

In Chapter 4 we need the following technical result (proved in Section 6.8 of [51]) about the equality of mixed partial derivatives.

Theorem B.3.2. *Suppose $\mathbf{F} : E \to \mathbb{R}^{d_2}$ is $\mathcal{C}^1$ on the open set $E \subset \mathbb{R}^{d_1}$. If a mixed partial derivative $\partial^2\mathbf{F}/\partial x_j \partial x_k$ exists and is continuous on E, then the derivative in the opposite order also exists, and both mixed partial derivatives are equal on E.*

B.3.2 The Implicit Function Theorem

Often we want to define a real variable y as a function of another variable x implicitly through an equation

$$f(x, y) = 0. \tag{B.10}$$

The implicit function theorem gives conditions under which this is possible. Below, we will consider such issues for vector variables, but for the moment, we assume that x and y are scalars.

Theorem B.3.3. *Suppose that (the scalar-valued) function $f(x, y)$ is continuously differentiable in a neighborhood of $(a, b) \in \mathbb{R}^2$, that $f(a, b) = 0$, and that*

$$\frac{\partial f}{\partial y}(a, b) \neq 0. \tag{B.11}$$

Then there are neighborhoods $\mathcal{U}_1$ of $a \in \mathbb{R}$ and $\mathcal{U}_2$ of $b \in \mathbb{R}$ such that for every $x \in \mathcal{U}_1$, equation (B.10) has a unique solution $y = Y(x) \in \mathcal{U}_2$. The function $Y : \mathcal{U}_1 \to \mathcal{U}_2$ is continuously differentiable, and of course $Y(a) = b$.

You may find the formulation of this result oppressively technical. To see why such care is needed, observe that the equation

$$x^4 + y^4 - 1 = 0 \tag{B.12}$$

satisfies the hypotheses of the theorem with $(a, b) = (0, 1)$. Its conclusions are valid, for example, for the choices

$$\mathcal{U}_1 = \{x : |x| < 1\} \qquad \text{and} \qquad \mathcal{U}_2 = \{y : y > 0\}.$$

However, if x leaves $\mathcal{U}_1$, then (B.10) may have no solutions. Moreover, for $x \in \mathcal{U}_1$, (B.10) has two solutions, $\pm(1 - x^4)^{1/4}$; we get uniqueness only by restricting our search for solutions to $y \in \mathcal{U}_2$.

To illustrate why the hypothesis (B.11) is necessary, consider solving (B.12) for y near $(a, b) = (1, 0)$.

A formula for the derivative $Y'(x)$ may be obtained by implicit differentiation of the relation $f(x, Y(x)) \equiv 0$, which yields

$$Y'(x) = -\frac{\partial f/\partial x(x, Y(x))}{\partial f/\partial y(x, Y(x))}. \tag{B.13}$$

Note that because of (B.11), the denominator in (B.13) is nonzero at $x = a$.

Having formulated the scalar version of the implicit function theorem with complete precision, we accept more informal language in generalizing to the vector version, say

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}, \tag{B.14}$$

where $\mathbf{x} \in \mathbb{R}^{d_1}$, $\mathbf{y} \in \mathbb{R}^{d_2}$, and $\mathbf{F} : E \to \mathbb{R}^{d_2}$ with $E \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ being open. We want to solve this system for $\mathbf{y} = (y_1, y_2, \ldots, y_{d_2})$ as functions of $\mathbf{x} = (x_1, x_2, \ldots, x_{d_1})$. Note that d_2 equations are to be solved for an equal number of unknowns.

Theorem B.3.4. *(Implicit function theorem) Suppose that $\mathbf{F} : E \to \mathbb{R}^{d_2}$ is $\mathcal{C}^1$, where $E \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ is open. Given a point $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ such that $\mathbf{F}(\mathbf{a}, \mathbf{b}) = \mathbf{0}$, let J be the partial Jacobian (i.e., only derivatives with respect to $\mathbf{y}$)*

$$J = \begin{bmatrix} \partial F_1/\partial y_1 & \partial F_1/\partial y_2 & \cdots & \partial F_1/\partial y_{d_2} \\ \partial F_2/\partial y_1 & \partial F_2/\partial y_2 & \cdots & \partial F_2/\partial y_{d_2} \\ \vdots & \vdots & \ddots & \vdots \\ \partial F_{d_2}/\partial y_1 & \partial F_{d_2}/\partial y_2 & \cdots & \partial F_{d_2}/\partial y_{d_2} \end{bmatrix},$$

evaluated at $(\mathbf{a}, \mathbf{b})$. *If* J *is invertible, then* (B.14) *may be solved locally near* $(\mathbf{a}, \mathbf{b})$ *for* $\mathbf{y}$ *as a function of* $\mathbf{x}$. *The solution is unique, and it is a* $\mathcal{C}^1$ *function of* x.

It is only too easy to forget whether it's the matrix of derivatives of $\mathbf{F}$ with respect to $\mathbf{x}$ or with respect to $\mathbf{y}$ that needs to be nonsingular. If you find yourself stranded on the proverbial desert island and need to sort this out, observe that the hypotheses imply that the matrix of derivatives of $\mathbf{F}$ with respect to $\mathbf{y}$ must be square, a $d_2 \times d_2$ matrix. By contrast, if $d_1 \neq d_2$, the matrix of derivatives with respect to $\mathbf{x}$ will not be square. Of course, only a square matrix can be invertible.

B.3.3 Surfaces and Manifolds

(a) Basic definitions. Informally, a *surface* in $\mathbb{R}^d$ may be defined as a subset such that near each of its points it is specified by the zero set of a smooth function. More precisely, if $S \subset \mathbb{R}^d$ is a $\mathcal{C}^1$-surface,[6] then for every $\mathbf{x}_0 \in S$ there exist a neighborhood $\mathcal{V}$ of $\mathbf{x}_0$ and a $\mathcal{C}^1$ function $\phi : \mathcal{V} \to \mathbb{R}$ with $\nabla\phi(\mathbf{x}_0) \neq \mathbf{0}$ such that

$$S \cap \mathcal{V} = \{\mathbf{x} \in \mathcal{V} : \phi(\mathbf{x}) = 0\}. \tag{B.15}$$

Given such a point $\mathbf{x}_0 \in S$, at least one partial derivative, say $\partial\phi/\partial\mathbf{x}_j(\mathbf{x}_0)$, is nonzero, so it follows from the implicit function theorem that the equation $\phi(\mathbf{x}) = 0$ may be solved locally for x_j as a function of the other coordinates. This property may be used as an alternative definition of a surface: *near every point, the set is a graph in which one coordinate can be expressed as a function of the others.*

Here's a trivial example: the set

$$S = \{(x, y) \in \mathbb{R}^2 : x^4 + y^4 - 1 = 0\} \tag{B.16}$$

is a surface in $\mathbb{R}^2$. (Since $d = 2$, it might be more natural to call S a curve.) At most points of (B.16) we can solve for either x or y as a function of the other variable; but at $(0, \pm 1)$, we can solve only for y as a function of x, and at $(\pm 1, 0)$, only for x as a function of y.

[6] This definition describes a surface *without boundary*. By contrast, we would call a set like the closed upper hemisphere, $\{(x, y, z) : x^2 + y^2 + z^2 = 1,\ z \geq 0\}$, a surface with boundary. The definition above does not describe points in the set that lie in the plane $\{z = 0\}$, i.e., the boundary.

Since this curve is defined globally by the zero set of a single function, there is little reason to bother with all the neighborhoods in the above definition. However, for a complicated curve such as in Figure B.2, you can see that being allowed to choose different defining functions at different points gives a more flexible concept.

Incidentally, in the neighborhood $\mathcal{V}$ of (B.15), the vector $\nabla\phi$ provides a local normal to S, generally not a unit normal.

Although manifolds can be defined abstractly [76], the following pedestrian definition is sufficient for purposes of this book: a k-dimensional *manifold* is a subset of $\mathbb{R}^d$ such that near each of its points, $d-k$ of the coordinates may be expressed as functions of the remaining k coordinates. Thus, a surface in $\mathbb{R}^d$ is a $(d-1)$-dimensional manifold, and a curve is a 1-dimensional manifold. The torus, represented as

$$\left\{\mathbf{x} \in \mathbb{R}^4 : x_1^2 + x_2^2 = 1,\ x_3^2 + x_4^2 = 1\right\},$$

is a two-dimensional manifold.

(b) Piecewise smooth bounding surfaces. In this text, surfaces arise in Chapter 4 as boundaries of trapping regions. We also consider regions whose boundary is a *piecewise* $\mathcal{C}^1$ surface, such as, for example, the set

$$\{(x,y) : x^4 + y^4 \le 1\} \sim \{(x,y) : x < 0 \text{ and } y < 0\}, \tag{B.17}$$

which we would describe as three-quarters of a misshapen pie. *(Draw it!)* To avoid unrewarding technicalities, we refrain from defining this concept in dimensions higher than two. A closed set $\mathcal{K} \subset \mathbb{R}^2$ has a *piecewise* $\mathcal{C}^1$ *boundary* if for every $\mathbf{x}_0 \in \partial\mathcal{K}$, there is a neighborhood $\mathcal{V}$ of $\mathbf{x}_0$ for which one of the following applies:

1. There is a $\mathcal{C}^1$ function $\phi : \mathcal{V} \to \mathbb{R}$ with $\nabla\phi(\mathbf{x}_0) \neq \mathbf{0}$ such that
$$\mathcal{K} \cap \mathcal{V} = \{\mathbf{x} \in \mathcal{V} : \phi(\mathbf{x}) \ge 0\}.$$
2. There are $\mathcal{C}^1$ functions $\phi_1 : \mathcal{V} \to \mathbb{R}$ and $\phi_2 : \mathcal{V} \to \mathbb{R}$ with $\nabla\phi_1(\mathbf{x}_0)$ and $\nabla\phi_2(\mathbf{x}_0)$ linearly independent such that *either*
$$\mathcal{K} \cap \mathcal{V} = \{\mathbf{x} \in \mathcal{V} : \phi_1(\mathbf{x}) \ge 0\} \cap \{\mathbf{x} \in \mathcal{V} : \phi_2(\mathbf{x}) \ge 0\} \tag{B.18}$$
or
$$\mathcal{K} \cap \mathcal{V} = \{\mathbf{x} \in \mathcal{V} : \phi_1(\mathbf{x}) \ge 0\} \cup \{\mathbf{x} \in \mathcal{V} : \phi_2(\mathbf{x}) \ge 0\}. \tag{B.19}$$

We call points described by Alternative 1 *regular* points; by Alternative 2, *corner* points. Regarding corner points of the set (B.17), the "internal" corners $(0,-1)$ and $(-1,0)$ are described by intersection, as in (B.18), and the "external" corner

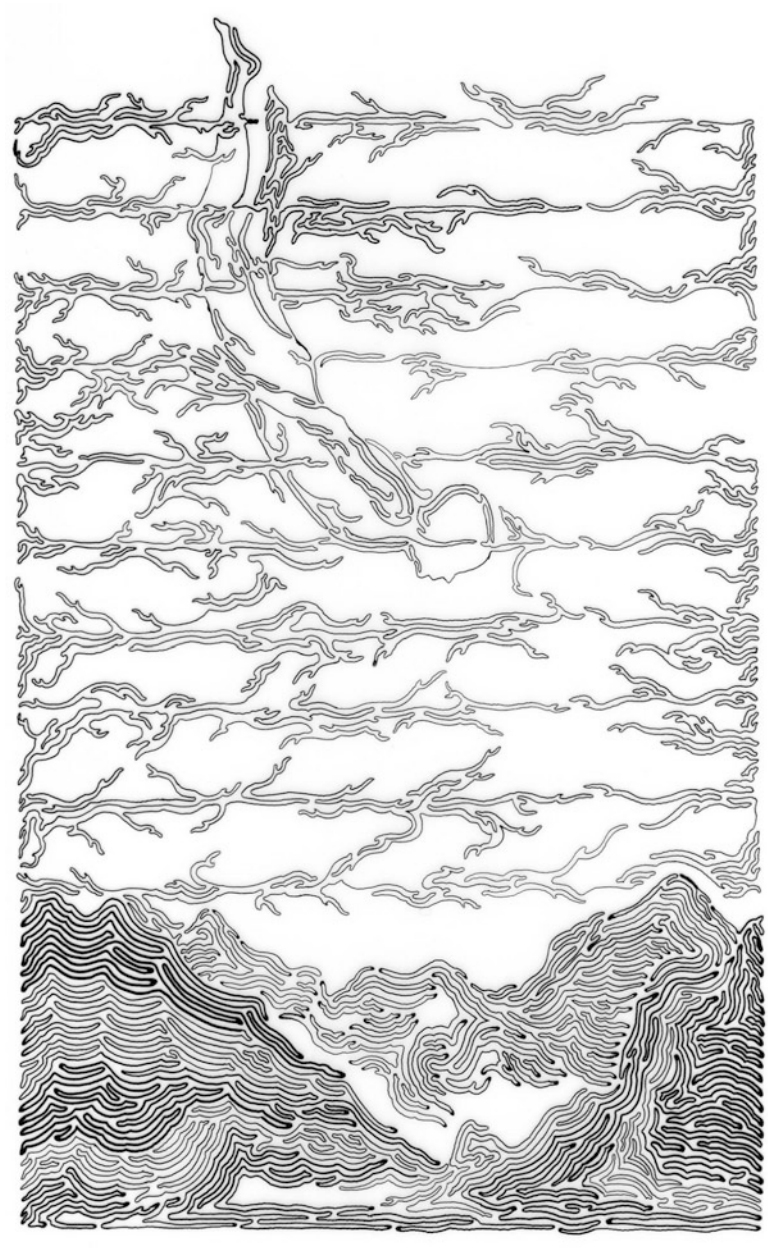

Figure B.2: *Example of a complicated simple closed curve. (Artist: Fiona Ross. "When we could be diving for pearls," 9 3/4 " x 6", 2011, Micron ink on Denril paper. Reprinted with permission.)*

$(0,0)$, by union, as in (B.19). For each of the corner points you should identify an appropriate neighborhood and functions ϕ_k that provide the desired representation.

What is the problem with defining a piecewise smooth boundary in $\mathbb{R}^d$? As a first attempt, you might imagine letting $\mathcal{K} \cap \mathcal{V}$ be defined locally by up to d functions, with more possibilities for unions and intersections of the sets $\{\mathbf{x} \in \mathcal{V} : \phi_k(\mathbf{x}) \geq 0\}$ than we care to enumerate. Such a definition would miss natural examples like a cone,

$$\{(x,y,z) : 0 \leq z \leq 1,\ \sqrt{x^2+y^2} \leq z\}.$$

We leave these depths unplumbed.

Even in two dimensions, our definition doesn't include all the cases you might expect. For example,

$$\{(x,y) : 0 \leq x \leq 1,\ 0 \leq y \leq x^2\}$$

is excluded because the gradients of $\phi_1(x,y) = y$ and $\phi_2(x,y) = x^2 - y$ are linearly dependent at the origin. This loss is not burdensome, and our definition has the virtue of simplifying the proof of Theorem 4.2.3. (Cf. the hint following Exercise 4.2.)

B.4 Exercises

B.4.1 Core Exercises

1. (a) Find the limit of the sequence of functions $f_n : [0,2] \to \mathbb{R}$,

$$f_n(x) = \frac{x^n}{1+x^n}, \qquad n = 1, 2, \ldots .$$

   (b) Does this series converge uniformly, or only pointwise?

2. Consider the sequence of functions

$$g_n(x) = \frac{2}{\pi} \arctan[nx - n^{2/3}];$$

these are translates of (B.1). Show that

$$\lim_{n\to\infty} g_n(x) = \begin{cases} 1 & \text{if } 0 < x < \infty, \\ -1 & \text{if } -\infty < x \leq 0, \end{cases}$$

and for all x,

$$\lim_{n\to\infty} g_n'(x) = 0,$$

both limits being pointwise.

3. *Introduction:* The point of this exercise is to construct a $\mathcal{C}^1$ monotone increasing function on $[0,\infty)$ that has a finite limit as $x \to \infty$ but whose derivative does not converge. (What you proved in Exercise 4.9 shows that such behavior is impossible for a function that satisfies an ODE.) For $x \in \mathbb{R}$, let

$$\phi(x) = \begin{cases} x^2(1-x)^2 & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise;} \end{cases}$$

let

$$f(x) = \sum_{n=1}^{\infty} n\phi(n^3(x-n)).$$

Draw the graph of the general term in the series. This graph is a little "blip" that as $n \to \infty$, gets higher and much thinner as it marches off to infinity.

(a) Show that the series for $f(x)$ converges pointwise and $f(x)$ is $\mathcal{C}^1$, nonnegative, and

$$\liminf_{x\to\infty} f(x) = 0, \qquad \limsup_{x\to\infty} f(x) = \infty.$$

(If you need a refresher on the definitions of lim inf and lim sup, see Section 1.5 of [51].)

(b) Show that the function

$$g(x) = \int_0^x f(s)ds \qquad (0 \le x < \infty)$$

is monotone increasing and $\lim_{x\to\infty} g(x) < \infty$.

Remark: Of course $g'(x) = f(x)$ does not converge as $x \to \infty$.

4. (a) Show that as $n \to \infty$, the functions $\sqrt{x^2 + n^{-1}}$ converge uniformly to $|x|$ for $x \in \mathbb{R}$, and the derivatives $(d/dx)\sqrt{x^2 + n^{-1}}$ converge pointwise to a step function.

Discussion: Of course $\lim_n \sqrt{x^2 + n^{-1}} = |x|$ is not differentiable at the origin.

With the addition of a bit of analysis, we can make this behavior more dramatic (see Part (e) below). Let $\{q_k : k = 1, 2, \ldots\}$ be an enumeration of the rational numbers in the interval $[0, 1]$. For $n = 1, 2, \ldots$ and $x \in \mathbb{R}$, let

$$f_n(x) = \sum_{k=1}^{\infty} \frac{\sqrt{(x-q_k)^2 + n^{-1}}}{k^2}, \tag{B.20}$$

i.e., an infinite linear combination of translates of $\sqrt{x^2 + n^{-1}}$.

(b) Show that for each n, the series (B.20) converges uniformly for x in every bounded interval and that its derived series

$$f_n'(x) = \sum_{k=1}^{\infty} \frac{x - q_k}{k^2\sqrt{(x-q_k)^2 + n^{-1}}}$$

converges uniformly for $x \in \mathbb{R}$.

Remark: By Corollary B.2.6, each function $f_n(x)$ in the sequence is continuously differentiable.

(c) Show that as $n \to \infty$, the sequence $\{f_n\}$ converges uniformly to

$$\sum_{k=1}^{\infty} \frac{|x - q_k|}{k^2}. \tag{B.21}$$

Remark: By Theorem B.2.2, $\lim_n f_n(x)$ is continuous. In fact, it is easily seen that (B.21) is Lipschitz continuous. However, this function is not differentiable at any rational $q_j \in [0, 1]$.

(d) Prove that $\{f_n'\}$ converges pointwise.

Hint: This is trivial if $x \notin [0, 1]$. If $x \in [0, 1]$ is irrational, then

$$\lim_{n\to\infty} f_n'(x) = \sum_{k=1}^{\infty} \frac{\operatorname{sign}(x - q_k)}{k^2},$$

while if $x = q_j \in [0, 1]$, we obtain the same series for $\lim_n f_n'(q_j)$ except with the term with $k = j$ omitted.

Challenge: Can you modify this example to get a function that is nondifferentiable at *all* rational numbers, not just those in $[0, 1]$?

5. *Introduction:* The function defined for $x \in \mathbb{R}$ by

$$\sum_{k=1}^{\infty} a^k \cos(2^k x), \tag{B.22}$$

where $1/2 < a < 1$, is continuous but nowhere differentiable. Weierstrass constructed the first such function in 1892; the proof that (B.22) has this behavior is due to Hardy [35]. Note that the terms of the derived series, $-(2a)^k \sin(2^k x)$, grow exponentially fast. As you might expect, this means that the derived series diverges badly. We don't ask you to establish this behavior rigorously, merely to examine numerical evidence. (Incidentally, (B.22) is an example of an important class of series known as *Fourier series*. Reference [4] provides an accessible introduction.)

(a) Show that the series (B.22) converges uniformly (so that the limit is continuous).

(b) Choose a value of a, say $a = 3/4$. Use the computer to graph the partial sums $1 \le k \le n$ over the interval $0 \le x \le \pi$ for $n = 1, 3, 5, 7, 9$, and observe how irregular the graphs become as n increases.

6. *Introduction:* Let $\{q_k : k = 1, 2, \ldots\}$ be an enumeration of the rational numbers. For $n = 1, 2, \ldots$ and $x \in \mathbb{R}$, let

$$f_n(x) = \frac{1}{n} \sum_{k=1}^{\infty} \frac{\arctan n(x - q_k)}{k^2}. \tag{B.23}$$

   (a) Show that for each n, the series (B.23) and its derived series

$$f_n'(x) = \sum_{k=1}^{\infty} \frac{1}{k^2 \left[1 + n^2(x - q_k)^2\right]} \tag{B.24}$$

   converge uniformly for $x \in \mathbb{R}$. Conclude that each function $f_n(x)$ is continuously differentiable.

   (b) Show that as $n \to \infty$, the sequence $\{f_n\}$ converges uniformly to zero.

   (c) Show that $\{f_n'\}$ converges pointwise.

   Hint: For each x, the sequence $\{f_n'(x)\}$ is a decreasing sequence of positive numbers, since each term in (B.24) has this behavior.

   (d) Show that at the jth rational number q_j,

$$\lim_{n\to\infty} f_n'(q_j) \geq \frac{1}{j^2} > 0.$$

   Discussion: In other words, although $\{f_n'\}$ converges, it does not converge to the derivative of $\lim_{n\to\infty} f_n(x)$.

7. In (B.23), define a sequence $g_n(x)$ by changing the leading coefficient n^{-1} to $n^{-1/2}$. Show that the sequence $\{g_n\}$ so defined still converges uniformly to zero, but the derived sequence $\{g_n'\}$ diverges; in particular, for every rational number q_j,

$$\lim_{n\to\infty} g_n'(q_j) = \infty.$$

   Remark: Incidentally, if the leading coefficient n^{-1} in (B.23) is removed altogether, the resulting sequence of functions converges pointwise to an increasing function that is discontinuous at every rational number.

8. *Introduction:* For $x \in \mathbb{R}$, let

$$f(x) = \begin{cases} x & \text{if } 0 \leq x \leq 1, \\ 2 - x & \text{if } 1 < x \leq 2, \\ 0 & \text{otherwise}, \end{cases}$$

   and define a sequence of (continuous) functions $f_n : [-1, 1] \to \mathbb{R}$ by $f_n(x) = nf(nx)$. Our standard advice: to get intuition about what's going on, graph $f_n(x)$.

(a) Show that $\{f_n\}$ converges pointwise to zero.

(b) Evaluate the integrals to conclude that

$$\lim_{n\to\infty} \int_{-1}^{1} f_n(x)\,dx = 1 \neq \int_{-1}^{1} \lim_{n\to\infty} f_n(x)\,dx. \tag{B.25}$$

9. (a) Show that both partial derivatives of the function (B.8) exist everywhere in the plane.

(b) Show that this function is not continuous at $(0,0)$.

Hint: This function vanishes at the origin, while along the line $\{x = y\}$, it equals $1/2$. Use this information to contradict the definition of continuity at the origin "for every $\varepsilon > 0 \ldots$," say with $\varepsilon = 1/3$.

B.4.2 PHD Exercises

10. Show that the unit ball in the set of continuous functions, say

$$\mathsf{S} = \{f \in \mathcal{C}([0,1]) : \|f\| \leq 1\},$$

is not compact.

Hint: First prove (or recall or see Theorem 3.1.3 in [51]) that if S is a compact subset of a metric space and $\{f_n\}$ is a sequence of elements of S, then $\{f_n\}$ has a convergent subsequence. Then consider the sequence

$$f_n(t) = \cos(2^n \pi t), \qquad n = 0, 1, 2, \ldots .$$

Argue that if $m < n$, then $\|f_m - f_n\| \geq |f_m(2^{-m}) - f_n(2^{-m})| = 2$. Deduce that no subsequence of $\{f_n\}$ can be Cauchy.

11. *Introduction:* The following problem, which doesn't have a lot to do with ODEs, is intended to remind you how to change variables in a multiple integral. Consider a $\mathcal{C}^1$ change of coordinates, say $\mathbf{x} = \mathbf{\Phi}(\mathbf{y})$. Given a region $R \subset \mathbb{R}^2$ and an integrable function f on the image $\mathbf{\Phi}(R)$, the change-of-variables formula (Theorem 9.3.1 in [51]) asserts that

$$\iint_{\Phi(R)} f(\mathbf{x}) dx_1 dx_2 = \iint_R f \circ \Phi(\mathbf{y}) J(\mathbf{y}) dy_1 dy_2, \tag{B.26}$$

where

$$J(\mathbf{y}) = \left| \det \begin{bmatrix} \partial\Phi_1/\partial y_1 & \partial\Phi_1/\partial y_2 \\ \partial\Phi_2/\partial y_1 & \partial\Phi_2/\partial y_2 \end{bmatrix} \right|.$$

(We confess that we still find this formula confusing and need to rethink it every time we want to use it.)

Just for practice, show that if $\Phi(r,\theta) = (r\cos\theta, r\sin\theta)$ is the transformation from polar to Euclidean coordinates, then $J(r,\theta) = r$.

Remark: Thus, in this case, if $R = \{(r,\theta) : 0 \le r \le a,\ 0 \le \theta \le 2\pi\}$, then (B.26) gives the usual expression for evaluating an integral over the disk using polar coordinates.

12. *Introduction:* Usually, to define the derivative with the difference quotient (B.6), one assumes that the function $\mathbf{F}$ is defined on an open set. In this way, $\mathbf{F}(\mathbf{x}+\mathbf{h})$ is defined for all $\mathbf{h}$ in a neighborhood of zero. Sometimes, however, it is convenient to consider a differentiable function defined on a closed set. It is trivial to do this for functions of one variable on a closed interval. Depending on the set, the issue may be problematic in several dimensions, but not for functions defined on a closed region with a smooth boundary. If the boundary is smooth, a $\mathcal{C}^1$ extension in a larger open set may be defined by reflection across the boundary. This exercise illustrates this construction in a special case in which the notation is simple.

Suppose that $f : \mathcal{K} \to \mathbb{R}$ is continuous, where $\mathcal{K} = \{x^2 + y^2 \le 1\}$ is the closed unit disk, and that f is continuously differentiable on Int $\mathcal{K}$. Show that if $\partial f/\partial x$ and $\partial f/\partial y$ extend to continuous functions on the closed disk, then the definition (in polar coordinates)

$$\tilde{f}(r,\theta) = \begin{cases} f(r,\theta) & \text{if } r \le 1, \\ 2f(1,\theta) - f(2-r,\theta) & \text{if } 1 < r < 3/2, \end{cases}$$

provides a $\mathcal{C}^1$ extension of f to a neighborhood of $\mathcal{K}$.

13. Given an arbitrary sequence $\{a_k\}$ of real numbers, show that there is a $\mathcal{C}^\infty$ function on the line whose derivatives satisfy $g^{(k)}(0) = a_k$.

Hint: To begin, construct a $\mathcal{C}^\infty$ function $\phi(x)$ such that $\phi(x) \equiv 1$ in a neighborhood of zero and $\phi(x) = 0$ if $|x| \ge 1$. Do this in stages as follows (see Figure B.3): if $f(x)$ is defined by (B.32) below, let

$$\begin{aligned} \phi_1(x) &= f(1-x^2), \\ \phi_2(x) &= \left\{\textstyle\int_{-\infty}^{\infty} \phi_1(s)ds\right\}^{-1} \int_{-\infty}^{x} \phi_1(s)ds, \\ \phi(x) &= \phi_2(2-4x^2). \end{aligned}$$

For $k = 0, 1, \ldots$, record the estimate for derivatives of order k or less,

$$M_k = \max_{x\in\mathbb{R}} \max_{0\le\ell\le k} \phi^{(\ell)}(x).$$

Now let $c_n = \max\{1, |a_n|\}$ and consider the series

$$g(x) = \sum_{n=0}^{\infty} a_n \phi(c_n x) \frac{x^n}{n!}. \tag{B.27}$$

The crux of the proof is to show that this series may be differentiated, arbitrarily many times, term by term. Since $\phi(c_n x) \equiv 1$ near the origin, a simple calculation then shows that $g^{(k)}(0) = a_k$.

Claim. *For every nonnegative integer k, the series*

$$\sum_{n=k+1}^{\infty} |a_n| \left| \left(\frac{d}{dx}\right)^k \left[\phi(c_n x)\frac{x^n}{n!}\right] \right| \tag{B.28}$$

converges uniformly for $x \in \mathbb{R}$.

Remarks: The claim permits the application of Theorem B.2.6 to show that $g(x)$ is $\mathcal{C}^k$ for all k. Note that the first $k+1$ terms in the series for $g(x)$ have been discarded in (B.28). This is convenient for the estimates below and of course has no bearing on convergence.

Outline for the proof of the claim: Iterate Leibniz's rule to conclude that

$$\left(\frac{d}{dx}\right)^k \left[\phi(c_n x)\frac{x^n}{n!}\right] = \sum_{j=0}^{k} \binom{k}{j} c_n^{k-j}\, \phi^{(k-j)}(c_n x)\, \frac{x^{n-j}}{(n-j)!}, \tag{B.29}$$

where $\binom{k}{j}$ is a binomial coefficient. Observe that since $\phi(c_n x)$ vanishes if $|x| \geq 1/c_n$, it follows that

$$c_n^{k-j} \left|\phi^{(k-j)}(c_n x) x^{n-j}\right| \leq c_n^{k-j} \frac{M_k}{c_n^{n-j}} = \frac{M_k}{c_n^{n-k}} \leq \frac{M_k}{c_n}.$$

Moreover, if $j \leq k$, then $1/(n-j)! \leq 1/(n-k)!$. Deduce from (B.29) that

$$\sum_{n=k+1}^{\infty} |a_n| \left| \left(\frac{d}{dx}\right)^k \left[\phi(c_n x)\frac{x^n}{n!}\right] \right| \leq 2^k M_k \sum_{n=k+1}^{\infty} \frac{|a_n|}{c_n} \frac{1}{(n-k)!}.$$

Since $|a_n| \leq c_n$, the sum (B.28) is bounded independently of x, as needed.

B.5 Pearls of Wisdom

In (2.26) we formed separate convergent series from the even-numbered and odd-numbered terms of a series. Such a splitting would be disastrous for the alternating series (B.4), since both

$$\sum_{n=1}^{N} \frac{1}{2n-1} \quad \text{and} \quad \sum_{n=1}^{N} \frac{1}{2n} \tag{B.30}$$

tend to infinity as $N \to \infty$.

Every rearrangement of a conditionally convergent series is problematic. For example, consider a rearrangement of (B.4),

$$\begin{aligned} 1 \; &- \tfrac{1}{2} \quad + \left(\tfrac{1}{3} + \tfrac{1}{5} + \tfrac{1}{7} + \ldots + \tfrac{1}{2n_1-1}\right) \; - \tfrac{1}{4} \\ &+ \left(\tfrac{1}{2n_1+1} + \tfrac{1}{2n_1+3} + \ldots + \tfrac{1}{2n_2-1}\right) \; - \tfrac{1}{6} \\ &\left(\tfrac{1}{2n_2+1} + \tfrac{1}{2n_2+3} + \ldots + \tfrac{1}{2n_3-1}\right) \; - \tfrac{1}{8} + \; \ldots \; . \end{aligned}$$

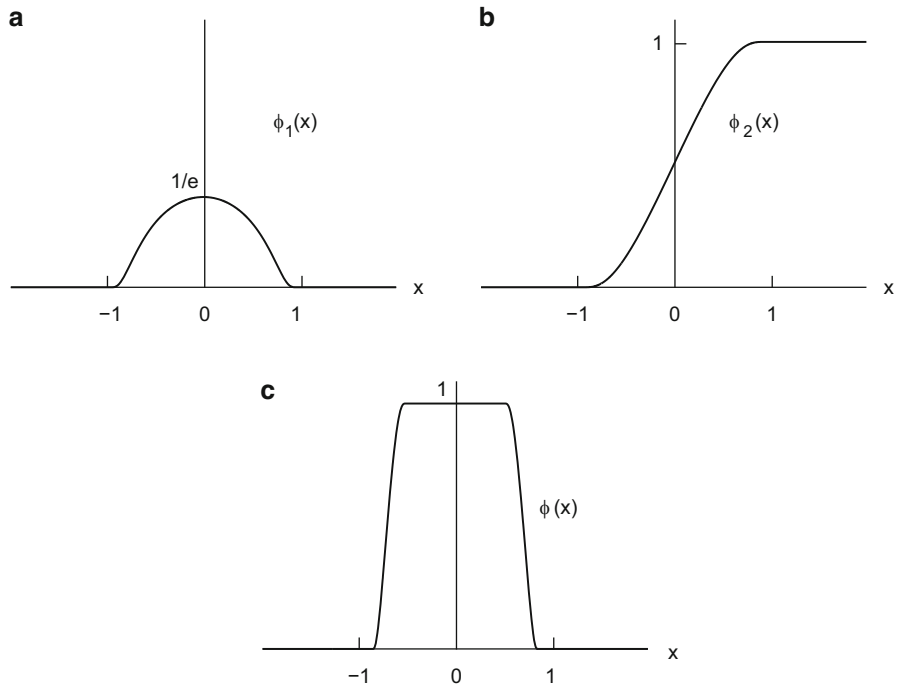

Figure B.3: *Three $\mathcal{C}^\infty$ functions used in the construction outlined in Exercise 13.*

For each block of positive terms, say

$$\left(\frac{1}{2n_k+1}+\frac{1}{2n_2+3}+\ldots+\frac{1}{2n_{k+1}-1}\right),$$

stuff enough terms into it so these terms add up to at least 1; this is possible, regardless of how many terms may have been assigned to previous blocks, because (B.30) still diverges even after any finite number of terms has been deleted. Then the partial sums of the rearranged series tend to infinity.

By rearranging the terms with a little more finesse, you can make the partial sums converge to any real number. *(Try it!)*

If $f(x)$ is a $\mathcal{C}^\infty$ function on the line, then for every $x_0 \in \mathbb{R}$, one can consider the formal power series

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n. \tag{B.31}$$

The series need not converge for any $x \neq x_0$. Indeed, in Exercise 13 it is shown that for every sequence $\{a_n\}$ of real numbers there is a $\mathcal{C}^\infty$ function such that at x_0,

$$f^{(n)}(x_0) = a_n.$$

Even when (B.31) does converge, its sum may not equal $f(x)$. A counterexample for this is

$$f(x) = \begin{cases} e^{-1/x} & x > 0, \\ 0 & x \leq 0\,; \end{cases} \tag{B.32}$$

this function is $\mathcal{C}^\infty$, and all its derivatives vanish at the origin, so (B.31) with $x_0 = 0$ reduces to a series of all zeros, but this doesn't sum to $f(x)$ for $x > 0$.

Convergence of (B.31) gives rise to the following definition. (Let's consider only functions of one variable.) A function is called *real analytic*[7] on (a, b) if for every $x_0 \in (a, b)$ and for all x in some neighborhood of x_0, the series (B.31) converges and its sum equals $f(x)$. Such functions are incredibly rigid. For example, if f is real analytic on (a, b) and if f vanishes on any subinterval $\mathcal{I} \subset (a, b)$, then $f(x) \equiv 0$ on the entire interval (a, b).

[7]The name "real analytic" relates to the property that a complex variable may be inserted into the convergent power series (B.31) to obtain a function that is analytic in the complex sense. See also [11].

Appendix C

Notions from Linear Algebra

C.1 How to Work with Jordan Normal Forms

In a linear algebra text, one expects the author to prove that an arbitrary square matrix is similar to a Jordan canonical form, and this proof is a messy affair.[1] Here we assume that you have seen the definitions and the statement of the theorem but not necessarily followed the complete proof. We accept that a Jordan normal form exists, and we ask, more simply, how to find it. We break this task into two subquestions, focusing more on examples than theory: given a matrix A,

(a) What is the Jordan normal form of A?

(b) What similarity transformation produces the normal form?

(a) Calculating the normal form: The first step in determining the normal form of A is to find its eigenvalues. Of course, finding eigenvalues analytically is an intractable problem in general, and one is quickly driven to the computer. We work with hand-picked examples in which the eigenvalues are readily determined.

Example 1:

$$A = \begin{bmatrix} 5 & -2 \\ 2 & 1 \end{bmatrix}.$$

It is readily computed that $\det(A - \lambda I) = (\lambda - 3)^2$. Even if your command of the theory is a little shaky, probably you know that there are two possible Jordan forms for A,

[1] We like the treatment in Appendix B of Strang [79]. Other widely used references include Hoffman and Kunze [42] and Meyer [55].

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8

$$J_1 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} \qquad \text{and} \qquad J_2 = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}.$$

If A were similar to $J_1 = 3I$, then every vector in $\mathbb{R}^2$ would be an eigenvector. However, $\mathbf{v} \in \mathbb{R}^2$ is an eigenvector iff $(A - 3I)\mathbf{v} = \mathbf{0}$, and not every vector satisfies this equation. Thus, J_2 must be the normal form for A. Indeed, in hindsight we may see that if a 2×2 matrix has equal eigenvalues but is not equal to a multiple of the identity, then its Jordan normal form must be a 2×2 block.

Higher-dimensional examples in which there are several double eigenvalues, but none of higher multiplicity, do not pose any additional difficulties, as we illustrate in Exercise 1. Let us turn our attention to eigenvalues of multiplicity three.

Example 2: Consider

$$A_1 = \begin{bmatrix} a & 1 & 1 \\ 0 & a & 0 \\ 0 & 0 & a \end{bmatrix}, \qquad A_2 = \begin{bmatrix} a & 1 & 1 \\ 0 & a & 1 \\ 0 & 0 & a \end{bmatrix}, \qquad A_3 = \begin{bmatrix} a & 0 & 1 \\ 0 & a & 1 \\ 0 & 0 & a \end{bmatrix}.$$

By inspection, $\lambda = a$ is the only eigenvalue of each matrix A_j. Thus the possible normal forms for A_j are

$$J_1 = \begin{bmatrix} a & & \\ & a & \\ & & a \end{bmatrix}, \qquad J_2 = \begin{bmatrix} a & 1 & \\ 0 & a & \\ & & a \end{bmatrix}, \qquad J_3 = \begin{bmatrix} a & 1 & 0 \\ 0 & a & 1 \\ 0 & 0 & a \end{bmatrix},$$

where to facilitate visualization, entries that are zero but lie outside of any Jordan block are left blank. We distinguish between cases by examining the dimension of the eigenspaces. These dimensions may be computed most easily by applying the "rank-plus-nullity" theorem (see Section 2.4 of Strang [79]), which gives us

$$\dim \ker(J_j - aI) = 3 - \operatorname{rank}(J_j - aI).$$

Thus J_1, J_2, J_3 have eigenspaces of dimension $3, 2, 1$, respectively. Proceeding similarly, we find that A_1, A_2, A_3 have eigenspaces of dimension $2, 1, 2$, respectively. Since the dimension of eigenspaces is preserved under similarity transformations, we conclude that A_1, A_2, A_3 have Jordan forms J_2, J_3, J_2, respectively.

Example 3:

$$A_1 = \begin{bmatrix} a & 0 & 0 & 1 \\ 0 & a & 0 & 1 \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix}, \qquad A_2 = \begin{bmatrix} a & 0 & 1 & 0 \\ 0 & a & 0 & 1 \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix}, \qquad A_3 = \begin{bmatrix} a & 1 & 0 & 0 \\ 0 & a & 0 & 1 \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix}.$$

The possible Jordan forms are

$$J_1 = \begin{bmatrix} a & & & \\ & a & & \\ & & a & \\ & & & a \end{bmatrix}, \qquad J_2 = \begin{bmatrix} a & 1 & & \\ 0 & a & & \\ & & a & \\ & & & a \end{bmatrix}, \qquad J_3 = \begin{bmatrix} a & 1 & 0 & \\ 0 & a & 1 & \\ 0 & 0 & a & \\ & & & a \end{bmatrix},$$

$$J_4 = \begin{bmatrix} a & 1 & 0 & 0 \\ 0 & a & 1 & 0 \\ 0 & 0 & a & 1 \\ 0 & 0 & 0 & a \end{bmatrix}, \qquad J_5 = \begin{bmatrix} a & 1 & & \\ 0 & a & & \\ & & a & 1 \\ & & 0 & a \end{bmatrix}.$$

Proceeding as above, we compute that J_1, J_2, J_3, J_4, J_5 have eigenspaces of dimension $4, 3, 2, 1, 2$, respectively. We can see potential trouble here in that J_3 and J_5 both have two-dimensional eigenspaces. Now A_1, A_2, A_3 have eigenspaces of dimension $3, 2, 2$ respectively. Thus, we may conclude that A_1 has J_2 as its normal form, but for A_2 and A_3, the dimension of the eigenspace does not distinguish between J_3 and J_5. To proceed, we turn to generalized eigenvectors: a vector $\mathbf{v} \in \mathbb{R}^d$ is called a *generalized eigenvector* of a matrix A with eigenvalue λ if for some power p,

$$(A - \lambda I)^p \mathbf{v} = \mathbf{0}.$$

Choosing $p = 2$, we compute that $(J_3 - aI)^2$ has a three-dimensional null space; and $(J_5 - aI)^2$, four-dimensional. On the other hand, $(A_2 - aI)^2$ has a four-dimensional null space, and $(A_3 - aI)^2$, three-dimensional. Thus the normal forms for A_2, A_3 are J_5, J_3, respectively.

We have not considered examples with complex eigenvalues or with several different multiple eigenvalues, but the above examples should be adequate preparation for these complications.

(b) Calculating the similarity transformation: For what matrix S does $S^{-1}AS$ produce the Jordan form of A? If A is diagonalizable, every matrix whose columns are eigenvectors of A will work (cf. Proposition 2.3.2). When the Jordan form for A is nondiagonal, we shall see that the columns of S should be appropriate *generalized* eigenvectors of A.

Recall Example 1, where

$$A = \begin{bmatrix} 5 & -2 \\ 2 & 1 \end{bmatrix}, \qquad \text{with } J = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$$

its Jordan form. Observe that with respect to the standard basis $\mathbf{e}_1, \mathbf{e}_2$ for $\mathbb{R}^2$, the matrix J satisfies

$$(J - 3I)\mathbf{e}_1 = \mathbf{0} \qquad (J - 3I)\mathbf{e}_2 = \mathbf{e}_1.$$

To match this behavior for A, we need to find vectors $\mathbf{v}_1, \mathbf{v}_2$ such that

$$(A - 3I)\mathbf{v}_1 = \mathbf{0} \qquad (A - 3I)\mathbf{v}_2 = \mathbf{v}_1,$$

and then the matrix $S = \mathrm{Col}(\mathbf{v}_1, \mathbf{v}_2)$ will achieve the required transformation. (Note that $(A - 3I)^2\mathbf{v}_2 = \mathbf{0}$, so $\mathbf{v}_2$ is a generalized eigenvector.) One possible choice is

$$S = \begin{bmatrix} 1 & 1/2 \\ 1 & 0 \end{bmatrix}. \tag{C.1}$$

We ask you, now or in Exercise 2, to check that this matrix performs the desired task. Incidentally, $\mathbf{v}_2$ could be replaced by $\mathbf{v}_2$ plus any multiple of $\mathbf{v}_1$. When A has multiple eigenvalues, there is generally more latitude in the choice of S than in the case of distinct eigenvalues.

More subtle issues may arise in cases of higher multiplicity. For instance, let A be the first of the three matrices considered in Example 2, and let J be its Jordan form, the second of the candidates. Observe that J satisfies

$$(J - aI)\mathbf{e}_1 = \mathbf{0}, \qquad (J - aI)\mathbf{e}_2 = \mathbf{e}_1, \qquad (J - aI)\mathbf{e}_3 = \mathbf{0}.$$

Thus we need to find vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ such that

$$(A - aI)\mathbf{v}_1 = \mathbf{0}, \qquad (A - aI)\mathbf{v}_2 = \mathbf{v}_1, \qquad (A - aI)\mathbf{v}_3 = \mathbf{0} \tag{C.2}$$

and then let $S = \mathrm{Col}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$. Note that $\mathbf{v}_1$ and $\mathbf{v}_3$ are eigenvectors of A, but $\mathbf{v}_1$ must be chosen with care in order that the middle equation in (C.2), which is inhomogeneous, have a solution. Now the eigenspace of A is spanned by

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}. \tag{C.3}$$

Suppose $\mathbf{v}_1$ is a linear combination of these vectors with coefficients α, β. Writing out the middle equation in (C.2), we have

$$\begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \\ -\beta \end{bmatrix}.$$

To have a solution, we need $\beta = 0$. Thus, choosing $\alpha = 1$, we see that

$$S = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \end{bmatrix} \tag{C.4}$$

is one of the possible similarity matrices that transforms A to its Jordan form. In Exercise 2(c) we ask you to determine the most general such similarity matrix.

We reckon that if you understand the above examples, you will be able to handle any matrix that is likely to come your way.

C.2 The Real Canonical Form of a Matrix

If a real matrix has complex eigenvalues, use of the Jordan normal form requires introducing matrices with complex entries, which may be inconvenient. The *real canonical form* avoids this. In the simplest case with complex eigenvalues, the real canonical form has the structure

$$\Gamma = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}. \tag{C.5}$$

In general, the real canonical form of a matrix consists of square blocks along the diagonal,

$$C = \begin{bmatrix} B_1 & 0 & 0 & \dots & 0 \\ 0 & B_2 & 0 & \dots & 0 \\ 0 & 0 & B_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & B_M \end{bmatrix}, \tag{C.6}$$

where each matrix B_m may be either a Jordan block (for real eigenvalues) or a block of the form

$$B = \begin{bmatrix} \Gamma & I & 0 & \dots & 0 & 0 \\ 0 & \Gamma & I & \dots & 0 & 0 \\ 0 & 0 & \Gamma & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \Gamma & I \\ 0 & 0 & 0 & \dots & 0 & \Gamma \end{bmatrix}, \tag{C.7}$$

where Γ has the structure (C.5), I is the 2×2 identity matrix, and 0 is the 2×2 zero matrix. (This includes the possibility that B is just the 2×2 matrix Γ.) Note that the only eigenvalues of (C.7) are $\lambda = a \pm ib$, each of which has only one eigenvector.

The real canonical form provides a proof of the following proposition. In Exercise 8 we guide you through the not-very-inspiring proof.

Proposition C.2.1. *Let A be a $d \times d$ matrix with real entries. For every $\varepsilon > 0$, there is a matrix B similar to A such that for all $\mathbf{x} \in \mathbb{R}^d$,*

$$|\langle \mathbf{x}, B\mathbf{x} \rangle - \langle \mathbf{x}, \Lambda \mathbf{x} \rangle| \leq \varepsilon |\mathbf{x}|^2, \tag{C.8}$$

where Λ is the diagonal matrix whose entries are the real parts of eigenvalues of A,

$$\Lambda = \mathrm{Diag}(\Re\lambda_1, \ldots, \Re\lambda_d).$$

C.3 Eigenvalues as Continuous Functions of Matrix Entries

Near simple eigenvalues, the eigenvalues of a matrix depend continuously on its entries, even differentiably. Specifically, we have the following result.

Proposition C.3.1. *Let λ_1 be a simple eigenvalue of a $d \times d$ matrix A_0. There exist an $\varepsilon > 0$ and a neighborhood $\mathcal{U}$ of A_0 in $\mathbb{R}^{d^2}$ such that every matrix $A \in \mathcal{U}$ has exactly one eigenvalue in the disk $\{z \in \mathbb{C} : |z - \lambda_1| < \varepsilon\}$. This eigenvalue, which is real if λ_1 is real, is a differentiable function of A.*

Proof. We prove this result by applying the implicit function theorem (over $\mathbb{R}$ or $\mathbb{C}$ as appropriate) to solve for λ in the equation for eigenvalues,

$$f(\lambda, A) = \det(A - \lambda I) = 0.$$

If $A = A_0$, this function can be written as a product of eigenvalues,

$$f(\lambda, A_0) = (\lambda_1 - \lambda)(\lambda_2 - \lambda)\ldots(\lambda_d - \lambda).$$

Differentiation of this product with respect to λ gives d terms, but only one of them is nonzero at $\lambda = \lambda_1$, i.e.,

$$\frac{\partial f}{\partial \lambda}(\lambda_1, A_0) = -(\lambda_2 - \lambda_1)\ldots(\lambda_d - \lambda_1).$$

Since λ_1 is a simple eigenvalue, none of these factors vanish, and the result follows. □

In the context of the above proposition, the following formula for the derivative $d\lambda/d\varepsilon$ may be of interest. Let $\mathbf{v}$ be an eigenvector of A with eigenvalue λ_1 and let $\mathbf{w}$ be the vector orthogonal to range $(A - \lambda_1 I)$ such that $\langle \mathbf{w}, \mathbf{v} \rangle = 1$. *(Why is this possible? Why, once $\mathbf{v}$ is chosen, is $\mathbf{w}$ defined uniquely?)* Then

$$d\lambda/d\varepsilon(0) = \langle \mathbf{w}, B\mathbf{v} \rangle.$$

A proof of this formula can be found on the web site.

The dependence of eigenvalues on the matrix is problematic near multiple eigenvalues. The simple example

$$A(\alpha) = \begin{bmatrix} 0 & 1 \\ \alpha & 0 \end{bmatrix},$$

where α varies near zero, provides the first indication of trouble: its eigenvalues $\pm\sqrt{\alpha}$ may be either real or complex, and although continuous, they certainly are not differentiable with respect to α at the origin. If complex entries are allowed or if the dimension is higher, even continuity cannot be guaranteed. For example, consider the matrix

$$A(\alpha,\beta) = \begin{bmatrix} 0 & 1 \\ \alpha+i\beta & 0 \end{bmatrix},$$

which has eigenvalues $\lambda_\pm = \pm\sqrt{\alpha+i\beta}$. In polar coordinates defined by $\alpha+i\beta = \rho\, e^{i\phi}$, we have $\lambda_\pm = \sqrt{\rho}\, e^{\pm i\phi/2}$. Since the exponentials $e^{\pm i\phi/2}$ are multiple-valued, eigenvalues cannot be defined as continuous functions near zero. The same difficulty arises in the 4×4 matrix (with real entries)

$$A(\alpha,\beta) = \begin{bmatrix} 0 & I \\ \Gamma(\alpha,\beta) & 0 \end{bmatrix},$$

where 0 is the 2×2 zero matrix, I is the 2×2 identity matrix, and Γ is the matrix in real canonical form

$$\Gamma(\alpha,\beta) = \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix}.$$

Although at a multiple eigenvalue, one cannot define individual eigenvalues continuously, nonetheless the *set* of eigenvalues does vary continuously, in the sense of the following proposition.

Proposition C.3.2. *Let λ_* be an eigenvalue of a $d\times d$ matrix A_0 of multiplicity k. For every sufficiently small $\varepsilon>0$, there is a neighborhood $\mathcal{U}$ of A_0 in $\mathbb{R}^{d^2}$ such that every matrix $A\in\mathcal{U}$ has exactly k eigenvalues in the disk $\{|z-\lambda_*|<\varepsilon\}$.*

Remarks: (i) In the proposition, ε must be less than the minimum separation between λ_* and the other eigenvalues of A_0. *(ii)* As $\varepsilon\to 0$, the maximum diameter of $\mathcal{U}$ scales like ε^k.

One elegant proof of this proposition involves some complex function theory. This subject is not a prerequisite for this text, so we refer you to Section 86 of [11].

In the language of Section 2.4, the following corollary of Proposition C.3.2 shows that if the origin is a sink for the linear system $\mathbf{x}' = A\mathbf{x}$, then it is also a sink for small perturbations of the system.

Corollary C.3.3. *If all the eigenvalues of A lie in the left half-plane $\{\Re\lambda<0\}$, then there is a neighborhood $\mathcal{U}$ of A in $\mathbb{R}^{d^2}$ such that the eigenvalues of every matrix $B\in\mathcal{U}$ also lie in the left half-plane.*

Continuity issues are a little different for symmetric matrices: All eigenvalues of a symmetric matrix are real, and one may define individual eigenvalues continuously by ordering them; i.e., we may define $\lambda_1(A)$ to be the smallest eigenvalue of A, $\lambda_2(A)$ to be the next smallest eigenvalue, etc. (Multiple eigenvalues do not matter

for these definitions.) However, even though with this convention the eigenvalues are continuous, they need not be differentiable. This is demonstrated by the matrix

$$A(\alpha,\beta) = \begin{bmatrix} \alpha & \beta \\ \beta & -\alpha \end{bmatrix},$$

which has eigenvalues $\lambda_1 = -\sqrt{\alpha^2+\beta^2}$ and $\lambda_2 = +\sqrt{\alpha^2+\beta^2}$.

C.4 The Routh–Hurwitz Criterion

It is astonishingly easy to determine whether a polynomial with real coefficients has all its zeros in the left half-plane. For example, for the two polynomials

$$\begin{aligned} Q_1(\lambda) &= \lambda^4 + 2\lambda^3 + 3\lambda^2 + 2\lambda + 1, \\ Q_2(\lambda) &= \lambda^5 + 2\lambda^4 + 3\lambda^3 + 3\lambda^2 + 2\lambda + 1, \end{aligned}$$

the calculations in Table C.1 show that $Q_1(\lambda)$ has all its zeros in $\{\Re\lambda < 0\}$, while $Q_2(\lambda)$ has at least one zero in $\{\Re\lambda \geq 0\}$. Let us explain these calculations in the context of a general polynomial

$$P(\lambda) = \lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \ldots + c_{n-1}\lambda + c_n.$$

The algorithm is slightly different, depending on whether n is even or odd. Reflecting this difference, we define $\nu = [n/2]$, where $[\cdot]$ is the greatest-integer function, i.e., $\nu = n/2$ if n is even and $\nu = (n-1)/2$ if n is odd. The algorithm forms an $(n+1)\times(\nu+1)$ matrix A as follows. The first two rows of A contain the coefficients of even and odd powers of λ:

$$\begin{array}{lccccc} a_{1l}: & 1 & c_2 & c_4 & \ldots & \\ a_{2l}: & c_1 & c_3 & c_5 & \ldots & . \end{array}$$

(If n is even, then 0 is inserted as the last entry of the second row, as in the table on the left.) Subsequent rows, $3, 4, \ldots, n+1$, are calculated inductively from products of entries from the two preceding rows that, apart from a sign, resemble a 2×2 determinant:

$$a_{k+1,l} = a_{k,l}a_{k-1,l+1} - a_{k,l+1}a_{k-1,l}. \tag{C.9}$$

In calculating the last column ($l = \nu+1$), entries $a_{k,\nu+2}$ or $a_{k-1,\nu+2}$ outside the appropriate range are assumed to be zero, as has been done in Table C.1; thus $a_{k,\nu+1} = 0$ for $k \geq 3$. Then we have the following theorem.

Theorem C.4.1. *All the zeros of P lie in the open left half-plane iff all entries in the first column of the above matrix are positive.*

If the calculation produces a zero row, as in the table on the right, then the calculation is stopped and we conclude that there is at least one zero in closed right half-plane. Indeed, note that $Q_2(\pm i) = 0$.

	1	2	3
1	1	3	1
2	2	2	0
3	4	2	0
4	4	0	0
5	8	0	0

	1	2	3
1	1	3	2
2	2	3	1
3	3	3	0
4	3	3	0
5	0	0	0
6	–	–	–

Table C.1: *The matrices* $\{a_{kl}\}$ *in the Routh–Hurwitz calculations for* $Q_1(\lambda) = \lambda^4 + 2\lambda^3 + 3\lambda^2 + 2\lambda + 1$ *(left table) and* $Q_2(\lambda) = \lambda^5 + 2\lambda^4 + 3\lambda^3 + 3\lambda^2 + 2\lambda + 1$ *(right table). Values of* k *from* 1 *to* $n+1$ *appear in the first column of each table; values for* l *from* 1 *to* $\nu + 1$ *appear in the top row. The two rows* $\{a_{kl} : k = 1, 2\}$*, which come directly from the coefficients of the polynomial, are separated from later rows that come from the calculation indicated in* (C.9).

This theorem is proved in Section 4.2 of [22]. Although the proof requires careful reading, it is not terribly difficult, just clever. In cases where some of the zeros of $P(\lambda)$ lie in the right half-plane, it is usually possible to deduce how many zeros lie there.

Turning to matrices and their eigenvalues, if A has real entries, then in principle, one could calculate the characteristic polynomial of A and apply the Routh–Hurwitz criterion to it to determine whether the eigenvalues of A lie in the left half-plane. However, calculating the characteristic polynomial of even a moderately large matrix by hand is not a pleasant task. (One could, of course, resort to symbolic computations to obtain the characteristic polynomial, but if the computer is involved, one might as well compute eigenvalues directly.) For 3×3 matrices, Proposition 2.4.6 provides a viable alternative to calculating the characteristic polynomial. Let us now apply the Routh–Hurwitz theorem to prove this result.

Proof of Proposition 2.4.6. Let A be a 3×3 matrix with characteristic polynomial

$$\det(A - \lambda I) = -[\lambda^3 + c_1\lambda^2 + c_2\lambda + c_3].$$

The application of the Routh–Hurwitz criterion to this polynomial is shown in Table C.2. Thus the roots of this polynomial are all in the left half-plane iff

$$\text{(a) } c_1 > 0, \quad \text{(b) } c_1c_2 - c_3 > 0, \quad \text{(c) } c_3 > 0. \tag{C.10}$$

	1	2
1	1	c_2
2	c_1	c_3
3	$c_1c_2 - c_3$	0
4	$c_3(c_1c_2 - c_3)$	0

Table C.2: *The matrix* $\{a_{kl}\}$ *in the Routh–Hurwitz calculations for the general cubic* $\lambda^3 + c_1\lambda^2 + c_2\lambda + c_3$.

The coefficients c_j are related to the eigenvalues of A through

$$\begin{aligned} c_1 &= -(\lambda_1 + \lambda_2 + \lambda_3), \\ c_2 &= \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1, \\ c_3 &= -\lambda_1\lambda_2\lambda_3. \end{aligned}$$

Thus, it is apparent that (C.10a) and (c) are equivalent to Conditions (i) and (iii) of Proposition 2.4.6, and the equivalence of (C.10b) with Condition (ii) follows on observing that

$$c_2 = \frac{1}{2}[(\mathrm{tr}A)^2 - \mathrm{tr}(A^2)].$$

□

C.5 Exercises

C.5.1 Core Exercises

1. (a) The following matrix has eigenvalues $2, 2, -4, -4$. Find its Jordan normal form and the similarity matrix that transforms A to normal form.

$$A = \begin{bmatrix} 0 & 1 & 4 & 1 \\ -1 & -2 & -1 & 2 \\ 4 & 1 & 0 & 1 \\ -1 & 2 & -1 & -2 \end{bmatrix}$$

   (b) The following matrix has eigenvalues $i, i, -i, -i$. Find its Jordan normal form and the corresponding similarity matrix.

$$A = \begin{bmatrix} 0 & -2 & 1 & 0 \\ 2 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Remark: See Exercise 7 for the real canonical form of this matrix.

2. (a) Show that S given by (C.1) transforms the matrix in Example 1 into normal form.

   (b) Show that S given by (C.4) transforms the matrix A_1 in Example 2 into normal form.

   (c) Find the most general S that transforms the matrix A_1 in Example 2 into normal form.

   Hint: Express each column of S as a linear combination of the vectors in a basis consisting of the two eigenvectors (C.3) and any linearly independent third vector, say $\mathbf{e}_2$. The answer involves a total of five independent parameters, restricted by a couple of inequalities to guarantee that S is nonsingular.

   (d) Find similarity matrices that transform the matrices A_2 and A_3 in Example 2 into normal form.

3. Find the Jordan normal form for the matrix

$$\begin{bmatrix} a & 0 & 1 & 0 \\ 0 & a & 1 & 1 \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix}.$$

4. *Definition:* A matrix (or linear transformation) such that $A^2 = A$ is called a *projection*.

   (a) Show that if A is a projection, then so is $I - A$.

   (b) Show that if A is a projection, then

$$\text{range } A = \ker(I - A).$$

   (c) Show that if A is a d-dimensional projection, then

$$\mathbb{R}^d = \ker A \oplus \text{range } A.$$

   (d) Show that if $|\mathbf{w}| = 1$, then the formula

$$A\mathbf{x} = \langle \mathbf{w}, \mathbf{x} \rangle \mathbf{w}$$

   defines a projection.

5. Write out the derivation of Corollary C.3.3 from Proposition C.3.2.

6. Check the Routh–Hurwitz criterion on some polynomials whose zeros you know, such as

$$\begin{aligned} (\lambda + 1)^5 &= \lambda^5 + 5\lambda^4 + 10\lambda^3 + 10\lambda^2 + 5\lambda + 1, \\ (\lambda + 1)^4(\lambda - 1) &= \lambda^5 + 3\lambda^4 + 2\lambda^3 - 2\lambda^2 - 3\lambda - 1 \end{aligned}$$

or

$$\lambda^4 + \lambda^3 + \lambda^2 + \lambda + 1.$$

Remark: The third polynomial, called the *cyclotomic* polynomial of degree 4, is the quotient $(\lambda^5 - 1)/(\lambda - 1)$, from which you can locate its zeros.

C.5.2 PHD Exercises

7. The real canonical form for the matrix in Exercise 1(b) has the block form

$$\begin{bmatrix} \Gamma & I \\ 0 & \Gamma \end{bmatrix},$$

where Γ is a 2×2 matrix in real canonical form, equation (C.5). Determine Γ and find the similarity matrix S that transforms A into real canonical form.

8. Prove Proposition C.2.1.

Hint: The proof is based on the simple observation that if Γ is the 2×2 real canonical form (C.5), then

$$\langle \mathbf{x}, \Gamma \mathbf{x} \rangle = a \langle \mathbf{x}, \mathbf{x} \rangle, \qquad \mathbf{x} \in \mathbb{R}^2,$$

but this simplicity is obscured by technicalities. To set up the necessary technical framework, let S be the similarity transformation that reduces A to real canonical form: $S^{-1}AS = C$. Split off the nilpotent part N of C, say $C = C_0 + N$, where

$$C_0 = \mathrm{Diag}(\Gamma_1, \ldots, \Gamma_K, \lambda_{2K+1}, \ldots, \lambda_d).$$

That is, each Γ_k, where $k = 1, \ldots, K$, is a 2×2 matrix of the form (C.5), and each λ_k is a real eigenvalue of A. Choose a scaling matrix T, as in Exercise 3(d) in Chapter 2, such that $T^{-1}CT = C_0 + \varepsilon N$. Complete the proof by letting $B = (ST)^{-1}A(ST)$ and observing that for all $\mathbf{x} \in \mathbb{R}^d$,

$$\langle \mathbf{x}, C_0 \mathbf{x} \rangle = \langle \mathbf{x}, \Lambda \mathbf{x} \rangle.$$

Remarks Regarding Additional References

So much has been published about ODEs that a complete bibliography is out of the question. Most of the following references are here because we actually cited them. We tended to cite secondary sources because (i) we thought they might be more readable for the beginning student and (ii) in this way, the total number of distinct reference sources could be kept smaller.

Here are some ODE books that have been important to us.

- Braun [10] and Strogatz [81] are unique books at a more elementary level than this one. Braun covers the standard material in a first course in ODEs, but he supports the theory with interesting applications peppered with quirky details that make them a joy to read. Strogatz covers an amazing range of applications and sneaks a lot of mathematical insight into very readable text.

- Hirsch–Smale [40], Meiss [54], and Perko [63] are general introductions to ODEs at more or less the same level as this book. But how can we recommend the competition?

- Guckenheimer–Holmes [33] and Hastings–McLeod [39] are natural sequels to this book. Guckenheimer–Holmes develops the rigorous analysis of dynamical systems, giving many of the proofs that we felt were too technical for this text. Hastings–McLeod studies a rare selection of nonstandard advanced problems in ODEs that reflects their good taste.

Incidentally, Robinson [67] has a lovely three-page prologue that gives a succinct historical overview of ODEs; although it was written more than a decade ago, we highly recommend it. And Abraham-Shaw [1], the mathematical equivalent of a graphic novel, is a classic—do not miss it!

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8

Bibliography

[1] R. H. Abraham and C. D. Shaw, *Dynamics: The geometry of behavior, parts 1–4: Bifurcation behavior*, Aerial Press, Santa Cruz, CA, 1988.

[2] U. Alon, *An introduction to systems biology*, Chapman and Hall/CRC, Boca Raton, 2006.

[3] M. A. Armstrong, *Groups and symmetry*, Springer-Verlag, New York, 1988.

[4] G. Bachman, L. Narici, and E. Beckenstein, *Fourier and wavelet analysis*, Springer-Verlag, New York, 2000.

[5] A. K. Bajaj and P. R. Sethna, *Flow induced bifurcations to three-dimensional oscillatory motions in continuous tubes*, SIAM Journal on Applied Mathematics **44** (1984), 270–286.

[6] G. Batchelor, *The life and legacy of G. I. Taylor*, Cambridge University Press, Cambridge, 1996.

[7] F. Beer, E. R. Johnston Jr, J. DeWolf, and D. Mazurek, *Mechanics of materials, 7th edition*, McGraw-Hill, New York, 2014.

[8] P. Bergé, Y. Pomeau, and C. Vidal, *Order within chaos: Towards a deterministic approach to turbulence*, Hermann, Paris, 1984.

[9] G. Birkhoff and G.-C. Rota, *Ordinary differential equations, 4th edition*, Wiley, New York, 1989.

[10] M. Braun, *Differential equations and their applications: An introduction to applied mathematics, 4th edition*, Springer, New York, 1993.

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8

[11] J. W. Brown and R. V. Churchill, *Complex variables and applications, 8th edition*, McGraw-Hill, New York, 2009.

[12] J. Carr, *Applications of centre manifold theory*, Springer-Verlag, New York, 1982.

[13] M. L. Cartwright, *van der Pol's equation for relaxation oscillations. In: Contributions to the theory of nonlinear oscillations, Volume 2*, Princeton University Press, Princeton, 1952.

[14] S-N. Chow and J. K. Hale, *Methods of bifurcation theory*, Springer, New York, 1982.

[15] E. A. Coddington and N. Levinson, *Theory of ordinary differential equations*, McGraw-Hill, New York, 1955.

[16] M. G. Crandall and P. H. Rabinowitz, *The Hopf bifurcation theorem in infinite dimensions*, Archive for Rational Mechanics and Analysis **67** (1977), 53–72.

[17] R. L. Devaney, *An introduction to chaotic dynamical systems, 2nd edition*, Addison-Wesley, Reading, 1989.

[18] P. G. Drazin, *Nonlinear systems*, Cambridge University Press, Cambridge, 1992.

[19] L. Edelstein-Keshet, *Mathematical models in biology*, Society for Industrial and Applied Mathematics, Philadelphia, 2005.

[20] S. P. Ellner and J. Guckenheimer, *Dynamic models in biology*, Princeton University Press, Princeton, NJ, 2006.

[21] M. B. Elowitz and S. Leibler, *A synthetic oscillatory network of transcriptional regulators*, Nature **403** (2000), 335–338.

[22] S. Engelberg, *A mathematical introduction to control theory*, Imperial College Press, London, 2005.

[23] G. B. Ermentrout and D. H. Terman, *Mathematical foundations of neuroscience*, Springer, New York, 2010.

[24] T. Erneux, *Applied delay differential equations*, Springer, New York, 2009.

[25] L. Euler, *(additamentum 1 de curvis elasticis) in: Methodus inveniendi lineas curvas maximi minimivi propietate gaudentes*, (1744).

[26] L. C. Evans, *An introduction to stochastic differential equations*, American Mathematical Society, Providence, 2013.

[27] R. FitzHugh, *Impulses and physiological states in theoretical models of nerve membrane*, Biophysical Journal **1** (1961), 445–466.

[28] G. B. Thomas, Jr. *Calculus and analytic geometry: Alternate edition*, Addison-Wesley, 1972.

[29] T. S. Gardner, C. R. Cantor, and J. J. Collins, *Construction of a genetic toggle switch in* Escherichia coli, Nature **403** (2000), 339–342.

[30] J. Gleick, *Chaos: Making a new science*, Viking, New York, 1987.

[31] M. Golubitsky and D. G. Schaeffer, *Singularities and groups in bifurcation theory, volume I*, Springer-Verlag, New York, 1985.

[32] M. Golubitsky, I. Stewart, and D. G. Schaeffer, *Singularities and groups in bifurcation theory, volume II*, Springer-Verlag, New York, 1988.

[33] J. Guckenheimer and P. Holmes, *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, Springer-Verlag, New York, 1983.

[34] G. M. Hall and D. J. Gauthier, *Experimental control of cardiac muscle alternans*, Physical Review Letters **88** (2002), 198102.

[35] G. H. Hardy, *Weierstrass's nondifferentiable function*, Transactions of the American Mathematical Society **17** (1916), 301–325.

[36] P. Hartman, *On local homeomorphisms of Euclidean spaces*, Boletín de la Sociedad Matemática Mexicana **5** (1960), 220–241.

[37] ———, *Ordinary differential equations*, John Wiley and Sons, New York, 1964.

[38] B. D. Hassard, N. D. Kazarinoff, and Y.-H. Wan, *Theory and applications of Hopf bifurcation*, Cambridge University Press, New York, 1981.

[39] S. P. Hastings and J. B. McLeod, *Classical methods in ordinary differential equations with applications to boundary value problems*, American Mathematical Society, Providence, 2012.

[40] M. W. Hirsch and S. Smale, *Differential equations, dynamical systems, and linear algebra*, Academic Press, New York, 1974.

[41] A. L. Hodgkin and A. F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, Journal of Physiology, London **117** (1952), 500–544.

[42] K. Hoffman and R. Kunze, *Linear algebra, second edition*, Prentice Hall, Englewood Cliffs, 1971.

[43] M. H. Holmes, *Introduction to perturbation methods, 2nd edition*, Springer, New York, 2013.

[44] J. H. Hubbard and B. H. West, *Differential equations: A dynamical systems approach*, Springer-Verlag, New York, 1991.

[45] A. Iserles, *A first course in the numerical analysis of differential equations, second edition*, Cambridge University Press, Cambridge, 2009.

[46] C. K. R. T. Jones, *Geometric singular perturbation theory, in: Dynamical systems*, Springer, Berlin, 1995.

[47] J. P. Keener and J. Sneyd, *Mathematical physiology I: Cellular physiology, 2nd edition*, Springer, New York, 2009.

[48] F. C. Klebaner, *Introduction to stochastic calculus with applications, 2nd edition*, Imperial College Press, London, 2005.

[49] G. F. Lawler, *Introduction to stochastic processes, 2nd edition*, Chapman and Hall/CRC, Boca Raton, 2006.

[50] E. N. Lorenz, *Deterministic nonperiodic flow*, Journal of the Atmospheric Sciences **20** (1963), 130–141.

[51] J. E. Marsden and M. J. Hoffman, *Elementary classical analysis, second edition*, W. H. Freeman, New York, 1993.

[52] J. E. Marsden and M. McCracken, *The Hopf bifurcation and its applications*, Springer-Verlag, New York, 1976.

[53] T. Matsumoto, *A chaotic attractor from Chua's circuit*, IEEE Transactions on Circuits and Systems (1984), 1054–1058.

[54] J. D. Meiss, *Differential dynamical systems*, SIAM, Philadelphia, 2007.

[55] C. D. Meyer, *Matrix analysis and applied linear algebra*, SIAM, Philadelphia, 2000.

[56] C. C. Mitchell and D. G. Schaeffer, *A two-current model for the dynamics of cardiac membrane*, Bulletin of Mathematical Biology **65** (2003), 767–793.

[57] J. D. Murray, *Mathematical biology. I: An introduction*, Springer, New York, 2002.

[58] ———, *Mathematical biology. II: Spatial models and biomedical applications*, Springer, New York, 2003.

[59] J. Nagumo, S. Arimoto, and S. Yoshizawa, *An active pulse transmission line simulating nerve axon*, Proceedings of the IRE **50** (1964), 2061–2070.

[60] H. F. Nijhout and M. C. Reed, *Homeostasis and dynamic stability of the phenotype link robustness and plasticity*, Integrative and Comparative Biology (2014). doi:10.1093/icb/icu010

[61] J. Nolasco and R. Dahlen, *A graphic method for the study of alternation in cardiac action potentials*, J. Appl. Physiol. **25** (1968), 191–196.

[62] M. A. Nowak and R. M. May, *Mathematical principles of immunology and virology*, Oxford University Press, Oxford, 2000.

[63] L. Perko, *Differential equations and dynamical systems, 3rd edition*, Springer-Verlag, New York, 2001.

[64] H. Poincaré, *Analysis situs*, Journal de l'École Polytechnique **1** (1895), 1–121.

[65] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical mathematics, 2nd edition.*, Springer, Berlin, 2007.

[66] M. C. Reed, *Fundamental ideas of analysis*, Wiley, New York, 1998.

[67] R. C. Robinson, *An introduction to dynamical systems: Continuous and discrete*, Pearson Prentice Hall, Upper Saddle River, 2004.

[68] W. Rudin, *Principles of mathematical analysis, 3rd edition.*, McGraw-Hill, New York, 1976.

[69] D. Ruelle and F. Takens, *On the nature of turbulence*, Communications in Mathematical Physics **20** (1971), 167–192.

[70] L. Segel and L. Edelstein-Keshet, *A primer on mathematical models in biology*, Society for Industrial and Applied Mathematics, Philadelphia, 2013.

[71] E. E. Sel'kov, *Self-oscillations in glycolysis*, European Journal of Biochemistry **4** (1968), 79–86.

[72] R. A. Serway and J. W. Jewett, Jr. *Principles of physics: A calculus-based text, 5th ed.*, Cengage, 2012.

[73] R. Shakarchi, *Problems and solutions for undergraduate analysis*, Springer, New York, 1998.

[74] M. Shearer and R. Levy, *Partial differential equations: An introduction to theory and applications*, Princeton University Press, Princeton, 2015.

[75] C. Sparrow, *The Lorenz equations: Bifurcations, chaos, and strange attractors*, Springer-Verlag, New York, 1982.

[76] M. Spivak, *Calculus on manifolds*, W. A. Benjamin, New York.

[77] M. Stein, *Loads and deformations of buckled rectangular plates*, NASA Technical Report R-40 (1959).

[78] R. Stott, *Darwin and the barnacle*, Faber and Faber Limited, London, 2003.

[79] G. Strang, *Linear algebra and its applications, 4th edition*, Cengage, 2006.

[80] W. A. Strauss, *Partial differential equations: An introduction, 2nd edition*, Wiley, New York, 2007.

[81] S. H. Strogatz, *Nonlinear dynamics and chaos*, Addison-Wesley, 1994.

[82] S. H. Strogatz, D. M. Abrams, A. McRobie, B. Eckhardt, and E. Ott, *Crowd synchrony on the Millennium Bridge*, Nature **438** (2005), 43–44.

[83] S. H. Strogatz and I. Stewart, *Coupled oscillators and biological synchronization*, Scientific American **269** (1993), 102–109.

[84] J. Szarski, *Differential inequalities*, Polish Scientific Publishers, Warsaw, 1965.

[85] C. H. Taubes, *Modelling differential equations in biology*, Prentice Hall, Upper Saddle River, 2001.

[86] G. I. Taylor, *The formation of a blast wave by a very intense explosion (parts 1 and 2)*, Proceedings of the Royal Society of London **201** (1950), 159–186.

[87] J. M. T. Thompson and G. W. Hunt, *A general theory of elastic stability*, Wiley, 1973.

[88] S. T. Thornton and J. B. Marion, *Classical dynamics of particles and systems*, Thomson Brooks/Cole, 2004.

[89] A. Uppal, W.H. Ray, and A.B. Poore, *The classification of the dynamic behavior of continuous stirred tank reactors—influence of reactor residence time*, Chemical Engineering Science **31** (1976), 205–214.

[90] B. van der Pol, *On "relaxation-oscillations"*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 7 **2** (1926), 978–992.

[91] M. Viana, *What's new on Lorenz strange attractors?*, The Mathematical Intelligencer **22** (2000), 6–19.

[92] L. N. Virgin, *Vibration of axially loaded structures*, Cambridge University Press, Cambridge, 2007.

[93] R. Weinstock, *Calculus of variations with applications to physics and engineering*, Dover, New York, 1974.

[94] S. Wiggins, *On the detection and dynamical consequences of orbits homoclinic to hyperbolic periodic orbits and normally hyperbolic invariant tori in a class of ordinary differential equations*, SIAM Journal on Applied Mathematics **48** (1988), 262–285.

[95] ———, *Introduction to applied nonlinear dynamical systems and chaos, 2nd edition*, Springer, New York, 2003.

[96] H. R. Wilson, *Spikes, decisions, and actions: The dynamical foundations of neuroscience*, Oxford University Press, Oxford, 1999.

[97] A. T. Winfree, *The geometry of biological time, 2nd edition*, Springer, New York, 2001.

[98] T. Witelski and M. Bowen, *Methods of mathematical modelling: Continuous systems and differential equations*, Springer, New York, 2015.

Index

D.G. Schaeffer, J.W. Cain, *Ordinary Differential Equations: Basics and Beyond*, Texts in Applied Mathematics, DOI 10.1007/978-1-4939-6389-8