An Invitation to Applied Mathematics with Differential Equations\textsuperscript{1}

Carmen Chicone\textsuperscript{2}

August 31, 2012

\textsuperscript{1}@ Carmen Chicone
\textsuperscript{2}Department of Mathematics, University of Missouri, Columbia, MO 65211.
E-mail: chiconeC@missouri.edu.
What is applied mathematics? Every answer to this question is likely to initiate a debate. My definition is the use of mathematics to solve problems or gain insight into phenomena that arise outside of mathematics. The prototypical example is the use of mathematics to solve problems in physics. Of course, the world of applied mathematics is much broader: important applications of mathematics occur in all areas of science, engineering, and technology.

The concept of this book is to introduce the reader to one aspect of applied mathematics: the use of differential equations to solve physical problems. To cover the full (ever expanding) range of applications of mathematics would require a series of books, which would include invitations to applied mathematics using the other branches of mathematics: calculus, linear algebra, differential geometry, graph theory, combinatorics, number theory, the calculus of variations, probability theory, and others. The application of statistics (especially in experimental science) is a branch of applied mathematics of great importance, but of a different character than the applied mathematics considered here.

While there are already many books and articles devoted to applications of mathematical subjects, I believe that there is room for more introductory material accessible to advanced undergraduates and beginning graduate students. If my invitation is accepted, perhaps the reader will pursue further study, find a problem in applied mathematics, and make a contribution to technology or the understanding of the physical universe.

My invitation includes a tour through a few of the historically important uses of differential equations in science and technology. The relevant mathematics is presented in context where there is no question of its importance.

A typical scenario in many research papers by mathematicians is an introduction that includes such phrases as “our subject is important in the study
of . . .”, “this problem arises in . . .”, or “our subject has many applications . . .” The authors go on to state a precise mathematical problem, they prove a theorem, and perhaps they give a mathematical example to illustrate their theorem; but all too often, they do not show how their theorem solves a problem in one of the applications that they used to advertise their work. This is not applied mathematics. The correct approach is joint work with an expert in some area of science. A physical problem is stated, a mathematical model is proposed, a prediction is made from the mathematical model—a step that might require some new mathematics including mathematical theorems, and the prediction is tested against a physical experiment.

We will discuss the basics of mathematical modeling; but, we will consider models for which the underlying science is easily accessible. The simple truth is that the construction of many important models requires a serious treatment of the corresponding science where the application is to be made. On the other hand, there are many useful models that can be understood with a basic knowledge of science. A carefully chosen selection of such models is explored in this book.

Applied mathematics requires an understanding of mathematics, creativity, hard work, and experience. The study of pure mathematics is important. As an aspiring applied mathematician, you must know at least what constitutes a mathematical proof and have a working knowledge of basic analysis. At best, you will acquire a deep understanding of several areas of mathematics including differential equations, analysis, and linear algebra. Your mathematical education should also include the study of mathematics in the context of various applications. This book is intended to provide a wealth of this valuable experience.

Columbia, Missouri

Carmen Chicone

June 2011
Acknowledgments

I thank all the people who have offered valuable suggestions for corrections and additions to this book, especially, Oksana Bihun,

Request

Please send your corrections or comments.
E-mail: chiconeC@missouri.edu
Mail: Department of Mathematics
    University of Missouri
    Columbia, MO 65211
## Contents

<table>
<thead>
<tr>
<th>Preface</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>To Do List and Notes</td>
<td>1</td>
</tr>
<tr>
<td>Differential Equations</td>
<td>5</td>
</tr>
<tr>
<td>Applied Mathematics and Mathematical Modeling</td>
<td>11</td>
</tr>
<tr>
<td>3.1 What is Applied Mathematics?</td>
<td>11</td>
</tr>
<tr>
<td>3.2 Fundamental, Constitutive and Phenomenological Models</td>
<td>13</td>
</tr>
<tr>
<td>3.3 Descriptive Models</td>
<td>15</td>
</tr>
<tr>
<td>3.4 Applied Mathematics in Practice</td>
<td>16</td>
</tr>
<tr>
<td>Conservation of Mass: Chemistry, Biology, and Thermodynamics</td>
<td>19</td>
</tr>
<tr>
<td>4.1 An Environmental Pollutant</td>
<td>19</td>
</tr>
<tr>
<td>4.2 Acid Dissociation, Buffering, and Titration with a Base</td>
<td>26</td>
</tr>
<tr>
<td>4.2.1 A Model for Dissociation</td>
<td>26</td>
</tr>
<tr>
<td>4.2.2 Titration with a Base</td>
<td>36</td>
</tr>
<tr>
<td>4.3 An Improved Titration Model</td>
<td>40</td>
</tr>
<tr>
<td>4.4 Reaction, Diffusion, and Convection</td>
<td>44</td>
</tr>
<tr>
<td>4.4.1 Fundamental and Constitutive Model Equations</td>
<td>44</td>
</tr>
<tr>
<td>4.4.2 Reaction-Diffusion in One Spatial Dimension: Heat, Mutations, and Traveling Waves</td>
<td>47</td>
</tr>
<tr>
<td>4.4.3 Reaction-Diffusion Systems: The Gray-Scott Model and Pattern Formation</td>
<td>65</td>
</tr>
<tr>
<td>4.4.4 Analysis of Reaction-Diffusion Models: Qualitative and Numerical Methods</td>
<td>68</td>
</tr>
<tr>
<td>4.5 Excitable Media: Transport of Electrical Signals on Neurons</td>
<td>129</td>
</tr>
</tbody>
</table>
Chapter 1

To Do List and Notes

What is the future of modeling from physical laws versus modeling from empirical observations? Should this subject be included.

One reason to learn mathematics is to use it to recast problems into a form where they can be effectively solved using computers.

A widely held view of applied mathematics education is to teach general principles, not specific applications. In an ideal world, general principles of mathematics will be taught in other courses. The knowledge gained in these courses is important: We cannot do applied mathematics without PDE, ODE, linear algebra, and analysis. My idea is to teach general principles in context with the applications.

There are issues in applied mathematics that should be mentioned, but would require perhaps too lengthy treatments. Two very important problems are parameter identification and the use, handling, and assimilation of experimental data.

Comments on chapters already in draft form:

General comment. More simple applications at the beginning of every chapter would be a good idea. I already tried to do this to some extent; that is, start simple and build up to more complex problems. But, it would be good to have enough simple material at the beginning of each chapter to make a course for undergrads without strong backgrounds in mathematics. I find that some of the material is difficult for undergrads at Missouri. Of course, we do not have the strongest students. But, MU may be a good indicator of the level of most students who will use this book.

Differential Equations The idea here is to give a warmup and review. Only a sketch is written. I don’t want to write a DE book in this chapter. But
perhaps I should expand this chapter to include more introductory material.

**Celestial Mechanics** This is the prototype for all applied mathematics. Something has to be included. The question is where to stop.

**Conservation of Mass** This chapter is the most complete. What if anything should be changed.

**Conservation of Momentum** This chapter is in reasonably good shape except for the section on SPH. The SPH method is wonderful stuff, but my current computing power seems insufficient to resolve the water waves for the Tiger Fountain Problem. Perhaps I should not use the method to approach the Tiger Fountain Problem; instead, I could use the method to solve a different applied problem that goes beyond the dam break problem. I am thinking about flight dynamics and the pickup truck tailgate problem: Is it more fuel efficient to drive with the tailgate up or down. This is an excellent problem in aerodynamics because you don’t need an airplane to do experiments. Should I include some finite difference methods for fluids? This is a place where we have to solve an important elliptic problem: finding the pressure. The section on elasticity is not yet complete. It is clear to me that the fundamental problem is the panel flutter problem, which is still an area of research. This problem will be introduced and a simple version of the problem will be addressed. I plan to end the chapter with a punch line.

**Pursuit:** I am not at all sure about including the chapter already written on pursuit. This is a nice problem in applied mathematics, in particular as it relates to missile guidance systems. But, maybe it is a bit too esoteric for beginning students. Should this chapter be eliminated? A second possibility, which is more attractive, is to transform the chapter into a discussion of basic flight dynamics.

The following is a list of topics with a high priority for inclusion as new chapters. The order of the list has changed several times. My intention is to list the highest priority items first.

**Biology:** There is not yet a fundamental mathematical biology: We do not yet know how to include life into our equations. But, we can and do apply the science we know from other disciplines. Pattern formation and chemical oscillations have already been included. Pattern formation is related to developmental biology and morphogenesis. The quintessential model in mathematical biology is the Hodgkin-Huxley/Fitzhugh-Nagumo model for the squid axon. It is included.

**Electricity and Magnetism:** One good problem that requires the solution of an elliptic equation is desirable. This would be an opening to the
finite element method, for example, and all the mathematics connected to elliptic theory. The problem is to introduce some models without a long introduction to the physics. Circuits are relatively easy. Linear circuits are likely already familiar to students. Nonlinear circuits serve as an intro to nonlinear dynamics. The PLL (Phase Locked Loop) might be a good model. Perhaps a more interesting topic is re-radiation. It might be possible to talk a bit about radar. But, the main difficulty is how to introduce the underlying science of electromagnetism from scratch.

**Optimization and Control** Optimization has been included in the BEM approach to free boundary problems for water waves. Control (PID feedback control) has been fully discussed in the section on controlling the temperature of a chamber. Perhaps at least one more good example from optimization should be included. Should the Euler-Lagrange equation be included? I have written a nice intro to this theory in my ODE book. Should some of this be repeated here. How about optimal control? It is difficult to be brief. It takes some work to build up to interesting problems.

**Modern Physics:** A fundamental problem is finite propagation speed. This leads to functional differential equations. Some basic quantum physics might perhaps be included, but it is not at all clear that there is a meaningful modeling problem that does not require a serious background in physics.

**Materials Science**

**Mechanical Engineering**
Ideas and To Do List
Chapter 2
Differential Equations

The reader should have a working knowledge of elementary differential equations (see [11]). A few essential results are discussed in this chapter.

The Harmonic Oscillator

The fundamental ordinary differential equation (ODE) for physics, mechanical engineering, and electrical engineering is the periodically forced harmonic oscillator

\[ m\ddot{x} + \epsilon \dot{x} + \omega^2 x = A \sin(\Omega t + \rho), \quad (2.1) \]

where (in its mechanical applications) \( x \) is a displacement, \( m \) is a mass, \( \epsilon \) is a (viscous) damping coefficient, \( \omega \) the natural frequency, and \( A \sin \Omega t \) a sinusoidal external force with amplitude \( A \), circular frequency \( \Omega \) and phase shift \( \rho \).

Exercise 2.1. Use Newton’s second law of motion to derive the harmonic oscillator model for a damped mass-spring system under the influence of gravity and an external sinusoidal force.

Exercise 2.2. (a) Solve the harmonic oscillator equation for the case \( A = 0 \) and \( \epsilon = 0 \) and describe the qualitative behavior of the corresponding solutions. (b) Solve the harmonic oscillator equation for the case \( A = 0 \) and describe the qualitative behavior of the corresponding solutions. (c) Solve the initial value problem

\[ \ddot{x} + \frac{1}{2} \dot{x} + 9x = 2 \sin t, \quad x(0) = 1, \quad \dot{x}(0) = 0 \]

and determine the long-term behavior of the displacement \( x \).
Exercise 2.3. Suppose that incoming waves oscillating at $\omega$ cycles per second are to be detected by the motion of a mass-spring system with fixed mass $m$ and damping coefficient $\epsilon$. How should the spring be chosen so that the response of the mass to the incoming wave has the greatest amplitude?

Exponential and Logistic Growth

The fundamental ODE for chemistry and biology is the exponential growth equation

$$\dot{Q} = rQ,$$

where $Q$ is the quantity of some substance and $r$ is a growth rate.

The logistic growth model is

$$\dot{Q} = rQ\left(1 - \frac{Q}{k}\right),$$

where $r$ and $k$ are parameters.

Exercise 2.4. (a) Solve the initial value problem

$$\dot{Q} = -2Q, \quad Q(0) = 3.$$

(b) Solve the initial value problem

$$\dot{Q} = -2Q + 8, \quad Q(0) = 3.$$

(c) Solve the initial value problem

$$\dot{Q} = \frac{2}{k}Q + 8, \quad Q(1) = 3.$$

(d) Solve the initial value problem

$$\dot{Q} = 2Q\left(1 - \frac{Q}{10}\right), \quad Q(0) = 5$$

and determine the fate of the solution (that is, determine the behavior of the solution as $t$ grows without bound).

Exercise 2.5. What is the monthly payment on a loan of $Q$ dollars for $T$ years in case the interest is compounded continuously at an annual rate $r$?
Differential Equations

Linear Systems

Linear systems occur in all areas of applied mathematics.

Exercise 2.6. (a) Solve the initial value problem
\[
\dot{x} = 2x - y, \quad \dot{y} = x + 2y, \quad x(0) = 1, \quad y(0) = -1.
\]
(b) Solve the initial value problem
\[
\dot{x} = x + 4y, \quad \dot{y} = 2x + y, \quad x(0) = 1, \quad y(0) = -1.
\]
(c) Find the general solution of the system
\[
\dot{x} = 4x - 2y + 4te^{6t}, \quad \dot{y} = 3x - 3y - 5t.
\]
Hint: The key concept here is the variation of parameters formula (see Appendix A.15).

Linear PDE

Applied mathematics is sometimes equated with the study of partial differential equations. Skip this section if you are unfamiliar with partial differential equations. This subject will be discussed in this book.

Exercise 2.7. (a) Solve the initial boundary value problem
\[
ut = u_{xx}, \quad u(t, 0) = 1, \quad u(t, 1) = 1, \quad u(0, x) = 1 - \sin \pi x.
\]
The PDE is called the heat equation. (b) Determine T such that \(u(T, 1/2) = 3/4\).

Exercise 2.8. (a) Suppose that \(f : \mathbb{R} \to \mathbb{R}\) is a twice continuously differentiable function. Solve the initial value problem
\[
\ddot{u} = u_{xx}, \quad u(0, x) = f(x), \quad u_t(0, x) = 0.
\]
The PDE is called the wave equation. (b) Show that the function \(f\) given by
\[
f(x) = -(10x - 1)^3(10x + 1)^3\text{ for } x \text{ in the interval } (-1/10, 1/10) \text{ and zero otherwise}
\]
is twice continuously differentiable. (c) Using \(f\) defined in (b), find the smallest time \(t > 0\) when \(u(t, 20) = 1/4\). (A numerical approximation correct within 1% is an acceptable answer.)

Exercise 2.9. Let \(\Omega\) denote the open set bounded by the unit square in the plane and let \(\bar{\Omega}\) denote its closer. The unit square has vertices at \((0, 0), (1, 0), (1, 1),\) and \((0, 1)\). Find the value of \(u : \bar{\Omega} \to \mathbb{R}\) at the point \((1/2, 1/2)\) in case \(u\) is harmonic (that is, \(u_{xx} + u_{yy} = 0\) in \(\Omega\)) and \(u(x, y) = 1\) everywhere on the boundary of \(\Omega\). The PDE is called Laplace’s equation.
Exercise 2.10. (a) Find a nonconstant solution of the PDE
\[ \sin^2 \phi \frac{\partial^2 f}{\partial \phi^2} + \frac{\partial^2 f}{\partial \theta^2} + \sin \phi \cos \phi \frac{\partial f}{\partial \phi} = 0 \]
that is periodic in each variable separately and with the additional property that it is continuous at \( \phi = 0 \). Hint: Use separation of variables. (b) The PDE has a geometric interpretation (see [33]). The Laplace-Beltrami operator is a generalization of the Laplacian to Riemannian manifolds. The usual Riemannian metric on a sphere of radius \( R \) is given in spherical coordinates by
\[ R^2 (d\phi^2 + \sin^2 \phi d\theta^2). \]
It is a covariant 2-tensor, which—in the present case—can be written more precisely as
\[ Ed\phi \otimes d\phi + Fd\phi \otimes d\theta + Fd\theta \otimes d\phi + Gd\phi \otimes d\phi. \]
The Laplace-Beltrami operator \( \Delta \) is given by
\[
\Delta f = \frac{\partial}{\partial \phi} \left( \frac{1}{\sqrt{EG - F^2}} \left( G \frac{\partial f}{\partial \phi} - F \frac{\partial f}{\partial \theta} \right) \right) + \frac{\partial}{\partial \theta} \left( \frac{1}{\sqrt{EG - F^2}} \left( E \frac{\partial f}{\partial \theta} - F \frac{\partial f}{\partial \phi} \right) \right).
\]
Our PDE is equivalent to the equation \( \Delta f = 0 \). Prove the last statement.

Nonlinear ODE

Newton’s second law states that a particle influenced by forces moves so that its mass times its acceleration equals the sum of the forces. In case the force is conservative, this law leads to the classical differential equation
\[ m\ddot{x} = -\nabla U(x), \]
where \( U \) is the potential energy. The kinetic energy is \( \frac{1}{2}m\dot{x}^2 \). Newton’s law of universal gravitation states that the gravitational force on a body due to a second body is directly proportional to the product of their masses and inversely proportional to the square of the distance between them. The direction of the force is toward the center of mass of the second body.

Exercise 2.11. (a) Show that the (total) energy (the sum of the potential and kinetic energies) is constant on solutions of Newton’s equation.
(b) Determine the total energy for Duffing’s equation
\[ \ddot{x} - x + x^3 = 0 \]
and draw its phase portrait. Describe the qualitative behavior of the system.
(c) Determine the total energy for the mathematical pendulum
\[ \ddot{\theta} + \sin \theta = 0 \]
Differential Equations

and draw its phase portrait. Describe the qualitative behavior of the system.

(d) Draw the phase portrait for (the alternate form of) Duffing’s equation

\[ \ddot{x} + x - x^3 = 0. \]

Describe the qualitative behavior of the solution with initial condition \( x(0) = 0 \) and \( \dot{x}(0) = 7/10 \).

(e) Determine the qualitative behavior of the solutions of the differential equation

\[ \ddot{x} + \frac{1}{10} \dot{x} - x + x^3 = 0. \]

(f) Suppose that two point masses fall from rest in the gravitational field due to a third mass, which might be the earth. And, suppose they move on the same line through the center of the third mass. Does the distance between them remain constant as they fall? If not, describe the relative distance as a function of time. (Hint: A numerical approximation of the solution might be necessary.)

(g) Determine the fate of the solution of the initial value problem

\[ \ddot{x} + \frac{1}{10} \dot{x} - x + x^3 = 0, \quad x(0) = \frac{1}{2}, \quad \dot{x}(0) = 0. \]

(h) Draw the phase-line portrait (for \( x \geq 0 \)) of the differential equation

\[ \dot{x} = x - x\sqrt{x} \]

and find the general solution. Does your phase portrait agree with your solution? Does your solution include all orbits depicted in your phase portrait?

(i) Determine the fate of the solution of the initial value problem

\[ \ddot{x} + \frac{1}{10} \dot{x} - x + x^3 = 0, \quad x(0) = 0, \quad \dot{x}(0) = 1. \]

Hint: This problem may not be solvable with paper and pencil. You might have to resort to numerics.

Numerics

We will study some numerical methods for solving differential equations in this book. Before we do so, go back to your ODE book and find the section on numerical methods. Write a code to implement some of the numerical methods discussed there, or find some software that you can use to solve a system of differential equations. Key words here are Euler Method, Improved Euler Method, Runge-Kutta Method, Adams Method, etc.
Exercise 2.12. (a) Consider the system

\[ \dot{x} = 1, \quad \dot{y} = axy, \]

where \( a \) is a parameter. Solve this system with initial data \( x(0) = y(0) = -1 \), and show that the exact value of the solution at \( t = 2 \) is \( (x, y) = (1, -1) \) independent of \( a \). (b) Generalize the result of part (a); that is, given \( x(0) < 0 \), there is a time \( T > 0 \), which is independent of \( a \), such that at \( t = T \) the solution starting at \( (x(0), y(0)) \) is at the point with coordinates \( (-x(0), y(0)) \).

Exercise 2.13. Use various numerical methods to solve the system of ODEs in Exercise 2.12 at least for the parameter values \( a = 1, 10^2, \) and \( a = 10^4 \) with \( x(0) \leq -1 \). Do your computer codes produce correct results? Discuss your experiments (compare, Exercise 4.61).

Exercise 2.14. Consider the system of differential equations

\[ \begin{align*}
\dot{x}_1 &= -k_1 x_1, \\
\dot{x}_i &= k_{i-1} x_{i-1} - k_i x_i, \quad i = 2, 3, 4, \ldots, n-1, \\
\dot{x}_n &= k_{n-1} x_{n-1}.
\end{align*} \]

It arises in situations that may schematically be described by a process \( X_1 \to X_2 \to X_3 \to \cdots \to X_n \) where the amount or concentration \( x_i \) of some substance in a region \( X_i \) (perhaps a tank) is determined by the amount of the substance coming into \( X_i \) from \( X_{i-1} \) minus the amount going out. The parameter \( k_i \) is the rate constant for the amount of substance leaving \( X_i \). Suppose initial data \( x_i(0) = \xi_i \) is also given. (a) Show that the system can be solved explicitly. (b) Let \( n = 10 \), \( k_i = i/1000 \), and \( \xi_i = 1 - 10i/101 \). Determine \( x_{10} \) at time \( t = 2000 \). Compare the exact solution with approximations using numerical methods for ODEs.
Chapter 3

Applied Mathematics and Mathematical Modeling

3.1 What is Applied Mathematics?

Applied mathematics is the use of mathematics to solve problems or gain insight into phenomena that arise outside of mathematics.

Recall that a prime number is an integer larger than one whose only divisors and itself and one. For example, 2, 3, 5, and 7 are the first four prime numbers. How many integers are prime numbers? This question arises inside mathematics.

Recall Newton’s second law of motion: the rate of change of linear momentum of a particle is the sum of the forces acting on it; and, Newton’s law of universal gravitation: the magnitude of the gravitational force on a mass produced by a second mass is directly proportional, with a universal constant of proportionality, to the product of the masses and inversely proportional to the square of the distance between their centers of mass; the direction of the force is along the line connecting their centers toward the second mass. These laws describe the relative motion of two masses each influenced only by the gravitational force of the other. The problem of determining the motion of two masses—the Newtonian two-body problem—is the prototype for applied mathematics; it is the subject of Chapter ??.

The two-body problem arises from celestial mechanics, or more precisely the theory of gravity, which is believed to be a fundamental force. Using Newton’s theory, we may build a mathematical model: Let $m_1$ and $m_2$ be
point masses in three-dimensional Euclidean space moving according to Newton’s second law of motion and his law of universal gravitation. Denote their positions in space by the position vectors $R_1$ and $R_2$, define the vector $R = R_2 - R_1$, the distance between their centers $r = |R|$, and let $G_0$ denote the universal gravitational constant. The equations of motion of the two bodies are

$$m_1 \ddot{R}_1 = \frac{G_0 m_1 m_2}{r^3} R, \quad m_2 \ddot{R}_2 = -\frac{G_0 m_1 m_2}{r^3} R.$$

This mathematical model consists of a pair of second-order ordinary differential equations, which is typical in classical mechanics. We may now make predictions about two-body motion with no further reference to physics or observations of nature by making mathematical deductions from these equations of motion. When this model was first proposed—not exactly in such compact language, Newton showed by mathematical deduction that these equations of motion predicted Kepler’s three laws of planetary motion, which were derived directly from observations of the motion of the planets in the solar system. For example, Kepler stated that each planet moves in a plane on an elliptical orbit with the sun at one focus of the ellipse. Kepler described the motions of the planets; Newton explained their motion. Kepler’s laws applied only to the motions of the observable planets; Newton’s laws applied equally well on the earth. For example, Galileo’s law of falling bodies is a direct consequence of Newton’s laws. These astounding successes verified that Newton’s laws are (close approximations of) fundamental laws of nature and that mathematical deductions from fundamental laws are predictive. While these events were preceded by the development and important applications of algebra, geometry and probability, the development of calculus and Newton’s laws are at the foundation of applied mathematics.

Exactly why mathematical deductions from physical laws are predictive of natural phenomena is a deep philosophical question; but, this fact is bedrock. The rationality and determinism of nature lie at the heart of the scientific method, the power of mathematical modeling, and applied mathematics.

The effectiveness of mathematics applied to science legitimizes the entire mathematical enterprise. While there are certainly other compelling arguments for the value of pure mathematics (that is, addressing the questions that arise within mathematics), the potential applicability to nature of mathematics that might have been developed without attention to this possibility is certainly an excellent reason to study mathematics.
The predictive power of Newton's laws cannot be overestimated. But, as you might know, these laws are not correct. The true nature of gravity is much more complicated and Newton's second law is not valid for masses whose relative velocities approach the speed of light. An easy thought experiment should convince you that the law of universal gravitation is not a perfect model of gravity. Simply note that the gravitational force is felt instantaneously with a change in distance between two masses. If the sun starts to oscillate, the motion of the earth would be effected immediately. By the same reasoning, a message could be sent instantaneously anywhere in the universe: shake the sun for a second to represent a one and pause the shaking for a second to represent a zero. Newton's law of universal gravitation predicts instantaneous action at a distance. Of course, Newton was well aware of this fact. While the theory of gravity was modified by Albert Einstein and will likely be modified in the future to conform more closely with observations, Newton's model of motion due to gravity is predictive up to the precision of most practical measurements as long as the relative velocity of the masses is much less than the speed of light. It is the prototypical example of an excellent predictive model that is routinely used in many important applications; for instance, the planning of space missions. As far as we know, there are no perfect mathematical models of reality. Fortunately, utility does not require perfection.

The prime objective of applied mathematics is to develop, analyze, and use mathematical models to make predictions, test hypotheses, and explain natural phenomena.

### 3.2 Fundamental, Constitutive and Phenomenological Models

Newton's two-body model is a fundamental model; it is derived from two fundamental laws of nature: the law of conservation of momentum and the law of universal gravitation. As suggested in the last section, the adjective "fundamental" means that the hypotheses are laws within Newtonian theory. There is also a two-body problem in Einsteinian theory, which has not yet been solved.

Most useful models are not fundamental. A familiar example is the usual model for the motion of a mass attached to the free end of a spring. Let $m$ de-
note the mass and $x$ the displacement of the spring from its equilibrium position. Newton’s second law (which is fundamental) states that $m\frac{d^2x}{dt^2} = F$, where $F$ is the sum of the forces on the mass. While the total force may contain a gravitational summand, the most important summand is the restoring force of the spring. At a fundamental level this force is electromagnetic; it is due to the atomic structure of the spring. In practice, the restoring force is not modeled using the Lorentz force law and Maxwell’s equations of electromagnetism; instead, the model is constructed from the constitutive (also called a phenomenological) Hooke’s law: the magnitude of the restoring force of the spring is proportional to its displacement from equilibrium and acts in the direction opposite the displacement. Hooke’s law is not a fundamental law of nature. It leads to the spring model $m\frac{d^2x}{dt^2} = -kx$, where $k$ is the constant of proportionality. This model is used extensively in physics and engineering. The spring equation is also called the harmonic oscillator.

We might imagine the nature of a fundamental model for spring motion. At least it would involve a coupled system of partial differential equations to account for the electromagnetic force. A correctly constructed model of this type would surely have the potential to yield more accurate predictions of spring motion. But, the added complexity of a fundamental model would certainly require sophisticated (perhaps yet unknown) mathematics or extensive numerical computations (perhaps beyond the limits of existing computers) to make predictions. To use Hooke’s law we must determine the spring constant $k$ before we can apply the spring equation to a physical spring, at least we must identify $k$ if we wish to make quantitative predictions. A fundamental model is likely to contain several constants that must be determined, some of which might not be easily measured. While a fundamental model would be more accurate, it would likely not be as useful as the spring equation. There is also the basic problem of formulating a fundamental model. At present, no one knows how to construct such a model for a spring. Indeed, modern elasticity theory is based on constitutive laws, Hooke’s law being the fundamental example.

Except for theoretical physics, which remains the prototype discipline for applied mathematics, constitutive models are ubiquitous in science. For many situations of practical interest, no one knows how to construct a fundamental model. In other cases, where a fundamental model might be constructed, constitutive models are usually preferred because they are simpler, provide insight, and often are sufficiently close representations of reality to provide predictions that agree with experiments up to current experi-
Figure 3.1: The data in this table was produced by observation of the distance from its origin of the diffusion front of a quantity of red ink deposited in a trough of water.

### Table 3.1

<table>
<thead>
<tr>
<th>time (sec)</th>
<th>distance (In)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.33</td>
<td>5.0</td>
</tr>
<tr>
<td>11.6</td>
<td>5.25</td>
</tr>
<tr>
<td>22.0</td>
<td>6.5</td>
</tr>
<tr>
<td>33.0</td>
<td>7.25</td>
</tr>
<tr>
<td>45.0</td>
<td>9.0</td>
</tr>
<tr>
<td>56.0</td>
<td>10.038</td>
</tr>
<tr>
<td>60.33</td>
<td>11.75</td>
</tr>
</tbody>
</table>

Experimental accuracy. The simplest model that provides insight and consistency with experiments is usually the best model (at least until a better model is invented).

The hallmark of a fundamental or constitutive model is that it is derived from some fundamental or constitutive laws. Scientists say they understand a phenomenon when they can create fundamental or constitutive models whose predictions always agree with experiments that measure some aspects of the phenomenon. In other words, understanding in this sense means that the phenomenon can be predicted using a more basic theory that applies more generally. Models derived from Newton’s law of motion, the law of universal gravitation, or Hooke’s law are prime examples.

### 3.3 Descriptive Models

Kepler’s laws are prototypical descriptive models: the planets move on ellipses, the radial vector from the sun to a planet sweeps out equal areas of the ellipse in equal time, and the square of the period of a planet is proportional to cube of the semi-major axis of its elliptical orbit. His statements are not derived from fundamental or constitutive laws.

A typical descriptive model is an equation that is chosen to fit experimental data. For example, Kepler’s law concerning the period of a planet’s motion was obtained by fitting to data recorded by the astronomer Tycho Brahe.

Table 3.1 lists experimental data for a crudely constructed experiment on the diffusion of ink in pure water. A 14.5 Inch trough (3.0 Inches wide)
was filled to a depth of 0.75 Inches with water and left undisturbed for a period of time to diminish the strength of convection currents. Red ink was deposited in the water near one end of the trough and allowed to diffuse. Measurements of the changing position of the diffusion front were recorded as a function of elapsed time. The data is plotted in Figure 3.2. The line in the figure is the graph of the linear function \( f \) given by \( f(x) = 3.84 + 0.126x \); it is a descriptive model of the measured phenomenon.

The model \( f \) can be used to make predictions. For instance, we may predict that the diffusion front will be 12.7 Inches from the origin after 70 seconds of elapsed time. While this might be an excellent prediction, the model tells us nothing about why the diffusion front moves in the observed fashion. A constitutive model for this diffusion experiment is suggested in Exercise 4.26.

### 3.4 Applied Mathematics in Practice

In an ideal world, scientists and engineers would create models and applied mathematicians would derive predictions from these models. But, in practice, the boundaries between scientists, engineers, and mathematicians are blurred. Perhaps one person assumes all three roles. The main feature of quality applied mathematics is its relevance to the subject area of application. Often the best way to achieve quality results is interdisciplinary collaboration.
A major difficulty to overcome for aspiring applied mathematicians (and textbook authors) is the necessity of learning enough of some scientific discipline outside of mathematics to aid in the development of useful models and the formulation of research questions that address important science. Poor quality applied mathematics is often a result of insufficient knowledge about the scientific area of application; mathematics is developed and theorems are proved that do not answer questions posed by scientists working in the area of application. Applied mathematicians should at least be aware of the important questions in the science they seek to advance.

Fortunately, the apprentice can learn the tools of applied mathematics in context by studying the mathematics required to understand and appreciate known applications of mathematics to science. Doing useful applied mathematics, journeyman status, requires a deep understanding of mathematics, understanding of the area of application, skill in computation, and a strong motivation to advance scientific knowledge. Newton was a master: he produced a fundamental model and developed the mathematics (calculus) that made it useful.
Chapter 4

Conservation of Mass: Chemistry, Biology, and Thermodynamics

4.1 An Environmental Pollutant

Consider a region in the natural environment where a water borne pollutant enters and leaves by stream flow, rainfall, and evaporation. A plant species absorbs this pollutant and returns a portion of it to the environment after death. A herbivore species absorbs the pollutant by eating the plants and drinking the water, and it return a portion of the pollutant to the environment in excrement and after death. We are given (by ecologists) the initial concentrations of the pollutant in the ambient environment, the plants, and the animals together with the rates at which the pollutant is transported by the water flow, the plants and the animals. A basic problem is to determine the concentrations of the pollutant in the water, plants, and animals as a function of time and the parameters in the system.

Figure 4.1 depicts the transport pathways for our pollutant, where the state variable $x_1$ denotes the pollutant concentration in the environment, $x_2$ its concentration in the plants, and $x_3$ its concentration in the herbivores. These three state variables all have dimensions of mass per volume (in some consistent units of measurement). The pollutant enters the environment at the rate $A$ (with dimensions of mass per volume per time) and leaves at the rate $e x_1$ (where $e$ has dimensions of inverse time). Likewise the remaining
rate constants $a$, $b$, $c$, $d$, and $f$ all have dimensions of inverse time.

The scenario just described is typical. We will derive a model from the law of conservation of mass: the rate of change of the amount of a substance (which is not created or destroyed) in some volume is its rate in (through the boundary) minus its rate out; or, to reiterate,

\[
\text{time rate of change of amount of substance} = \text{rate in} - \text{rate out}.
\]

This law is easy to apply by taking the units of measurement into account. Often we are given the volume $V$ of a container, compartment, or region measured in some units of volume, the concentration $x$ of some substance in the compartment measured in amount (mass) per volume, and rates $k$ of inflow and $\ell$ of outflow for some medium carrying the substance measured in units of volume per time. The rate of change $dx/dt$ in this case, where $t$ denotes time measured in some consistent unit, has units of amount per volume per time. The rate of outflow $\ell$ times the concentration $x$ has units of amount per time. Thus, we must multiply the time-derivative by the total volume to achieve consistent units in the differential equation

\[
\frac{d(Vx)}{dt} = \text{rate in} - \ell x,
\]

where the inflow rate must have units of amount per time. In case $V$ is constant, we may divide both sides of the equation by $V$. The quantity $\ell/V$
Conservation of Mass

has units of inverse time. In this case, it is called the rate constant for the
outflow of the substance.

For the environmental pollutant, the rate constants are supposed to be
given; the volumes of the compartments (the environment, herbivores, and
plants) are in the background but not mentioned. Using conservation of
mass, the transport equations are

\[ \begin{align*}
\dot{x}_1 &= A + ax_2 + bx_3 - (c + d + e)x_1, \\
\dot{x}_2 &= cx_1 - (a + f)x_2, \\
\dot{x}_3 &= dx_1 + fx_2 - bx_3.
\end{align*} \tag{4.1} \]

This type of model, which is ubiquitous in mathematical biology, is called a
compartment model.

The parameters in the model (4.1) would be measured by the field work
of ecologists. When the parameters are determined, the model can be used
to make quantitative predictions of the future concentrations of the pollutant
by solving the system of ODEs (4.1).

Some of the rates might be difficult to measure directly (for example, the
rate constant \( f \) that determines the transport of the pollutant from plants to
the herbivore). These rates might be estimated using a two step procedure:
The pollutant concentrations are measured in the field over some suitable
period of time and the parameters are chosen in the model to match these
measurements. Exactly how to choose the parameters based on available
data is the parameter estimation problem; it is one of the most important
and difficult issues in mathematical modeling.

Our model may also be used to make qualitative predictions. For example,
we might make a mathematical deduction from the form of the equations
that determines the long-term substance concentrations for some range of
parameter values. We might also use our model to predict the outcome
of some intervention in the environment. For example, suppose that some
portion of the plants are harvested by humans and removed from the region
under study. The effect of this change can be predicted by solving our model
equations with appropriate assumptions. The ability to make predictions
without conducting new field studies is one of the most important motivations
for developing a mathematical model. Another reason to develop a model
is to test hypotheses about the underlying physical process. For example, a
model that leads to predictions which do not agree with data obtained from
field observations must be based on at least one false hypothesis.
The linear system (4.1) can be solved explicitly. But, since the eigenvalues of the system matrix are roots of a cubic polynomial, explicit formulas for the time evolution of the state variables \( x_1 \), \( x_2 \), and \( x_3 \) are complicated. This is to be expected. Differential equations with simple explicit solutions are rare. Most differential equations do not have explicit solutions and most explicit solutions are too complicated to be useful. For this reason, model validation and predictions are usually obtained by a combination of qualitative methods, analytic approximations, and numerical approximations. The most valuable information is usually obtained by qualitative analysis.

Physically meaningful concentrations are all nonnegative. Hence, our model should have the property that nonnegative initial concentrations remain nonnegative as time evolves. Geometrically, the positive octant of the three-dimensional state space for the variables \( u \), \( v \), and \( w \) should be positively invariant (that is, a solution starting in this set should stay in the set for all positive time or at least as long as the solution exists). To check the invariance, it suffices to show that the vector field given by the right-hand side of the system of differential equations is tangent to the boundary of the positive octant or points into the positive octant along its boundary, which consists of the nonnegative parts of the coordinate planes. In other words, positive invariance follows because \( \dot{x}_3 \geq 0 \) whenever \( x_3 = 0 \), \( \dot{x}_2 \geq 0 \) whenever \( x_2 = 0 \), and \( \dot{x}_1 \geq 0 \) whenever \( x_1 = 0 \).

What happens to the concentrations of the pollutant after a long time? Our expectation is that the system will evolve to a steady state (that is, a zero of the vector field that defines the ODE). To determine the steady state, we simply solve the system of algebraic equations obtained by setting the time derivatives of the system equal to zero. A computation shows that if \( be(a + f) \neq 0 \), then the solution of the system

\[
\begin{align*}
A + ax_2 + bx_3 - (c + d + e)x_1 &= 0, \\
cx_1 - (a + f)x_2 &= 0, \\
dx_1 + fx_2 - bx_3 &= 0
\end{align*}
\]

(4.2)
is

\[
\begin{align*}
x_1 &= \frac{A}{e}, & x_2 &= \frac{Ac}{e(a + f)}, & x_3 &= \frac{A(ad + (c + d)f)}{be(a + f)}.
\end{align*}
\]

If \( be(a + f) > 0 \), then there is a steady state in the positive first octant. If \( b = 0 \), \( e = 0 \), or \( (a + f) = 0 \), then there is no steady state. This is reasonable on physical grounds. If the pollutant is not returned to the environment
by the herbivores, then we would expect the pollutant concentrations in the herbivores to increase. If the pollutant does not leave the environment, then the concentration of the pollutant in our closed system will increase. If the plants do not transfer the pollutant to the environment or the herbivores, we would expect their pollutant concentration to increase.

Suppose that there is a steady state. Do the concentrations of the pollutant approach their steady state values as time passes? This basic question is answered for our linear system by determining the stability of the corresponding rest point in our dynamical system. The stability of a steady state can usually be determined from the signs of the real parts of the eigenvalues of the system matrix of the linearization of the system at the steady state (see Appendix A.17). In particular, if all real parts of these eigenvalues are negative, then the steady state is asymptotically stable.

The system matrix at our steady state is

\[
\begin{pmatrix}
-c - d - e & a & b \\
 0 & c & -(a + f) \\
 d & f - b
\end{pmatrix}
\] (4.3)

To show the asymptotic stability of the rest point, it suffices to prove that all the real parts of the eigenvalues are negative. This follows by a direct computation using the Routh-Hurwitz criterion (see Appendix A.17 and Exercise 4.1). Of course, this also agrees with physical intuition. Since our system is linear, we can make a strong statement concerning stability: if \( e \neq 0 \), then all initial concentrations evolve to the same steady state concentrations.

Our qualitative analysis makes numerical computation unnecessary for answering some questions about our system, at least in those cases where we can reasonably assume the system is in steady state. For example, if the rate at which the pollutant leaves the environment is decreased by 50%, then we can conclude that the pollutant concentration in the plants will be doubled. On the other hand, we may resort to numerical computation if we wish to predict some transient behavior of our system (see Exercise 4.2).

**Exercise 4.1.** (a) Prove that if \( a, b, c, d, e, \) and \( f \) are all positive, then all the roots of the characteristic polynomial of the matrix (4.3) have negative real parts. (b) Show that the same result is true if \( a, b, e, \) and \( f \) are positive and the constants \( c \) and \( d \) are nonnegative.
Exercise 4.2. Suppose that the parameters in the model (4.1) are

\[ A = 10, \quad a = 1, \quad b = 1/100, \quad c = 3, \quad d = 1, \quad e = 1, \quad f = 4. \]

See Section 4.4.4 if you are not familiar with numerical methods for ODEs. (a) For the initial concentrations (at time \( t = 0 \)) \( x_1 = 0, \ x_2 = 0, \) and \( x_3 = 0, \) determine the time \( t = T \) such that \( x_3(T) = 1/2. \) (b) Suppose that the pollutant enters the environment periodically, \( A = 10 + 5 \sin(2\pi t) \) instead of the constant rate \( A = 10. \) Argue that the state concentrations fluctuate periodically and determine the amplitude of this fluctuation in the plant species.

Exercise 4.3. Consider the arrangement of tanks and pipes depicted in Figure 4.2. In this case, liquid is pumped at the indicated rates. Two solutions enter the first tank, which has liquid volume \( V_I, \) with concentrations \( \alpha \) lbs/gal and \( \beta \) lbs/gal and flow rates of \( a \) gal/min and \( b \) gal/min (respectively) and the solution in the tank (which is assumed to be stirred so that the system is mixed instantly) enters and leaves the first tank at the rates \( h \) and \( c \) gal/min (respectively), and it leaves the second tank, which has liquid volume \( V_{II}, \) at the rate \( k \) gal/min. The liquid volume of each tank is constant. We are also given the initial amounts of the chemical in each tank. (a) Determine the amounts of the solutes in each of the tanks as a function of time. (b) What happens to the concentrations of the solutes in the long run?
Exercise 4.4. Imagine three tanks each containing three different substances in solution. The tanks are connected by equal sized pipes to form a closed loop; that is, tank-1 feeds tank-2 and is fed by tank-3, tank-2 feeds tank-3 and is fed by tank-1, and tank-3 feeds tank-1 and is fed by tank-2. There are pumps in the pipes that maintain the circulation in the indicated direction at a fixed flow rate and the contents of the tanks are stirred so that the mixtures are homogeneous. (a) Make a model for the concentrations of the first substance in the three tanks and use it to predict concentrations as functions of time. What does the model predict? (Hint: Use pencil and paper to determine the steady states.) (b) Consider three bins each containing a mixture of red, blue, and green balls. The same number of balls (less than the minimum number of balls in a bin) is chosen at random from each bin and redistributed. Those chosen from the first bin are moved to the second, those from the second bin are moved to the third, and those from the third bin are moved to the first. This process may be repeated. Write a computer simulation of the process and track the concentration of red balls in each bin. (c) A simple model of the process in part (b) is given by the differential equations model of part (a). How well does the model predict the concentrations of red balls? Discuss the assumptions of the model and their validity for the redistribution process. (Hint: Make sure the initial states are the same for the process and the ODE model.) (d) Suppose a fixed amount of the first substance is created continuously in tank-2 and the same amount is destroyed in tank-1. What happens in the long run? Compare with a bin simulation. Note: The bin simulation is an example of a Markov process, for which there is a well-developed theory that is beyond the scope of this book.

Exercise 4.5. [PID-Controller] (a) Imagine a tank partially filled with water. A pipe feeds water to the tank at a variable flow rate, and there is also a drain pipe with a computer controlled variable flow valve connected to a sensor in the tank that measures the tank’s volume. The valve opens exactly enough to let water drain from the tank at a rate proportional to the volume of the tank. The program allows the user to set one number: the constant of proportionality. Write a model for this physical problem. Be sure to define all the variables in your model. (b) Suppose the inflow rate is constant. How should the proportionality constant in the control mechanism be set to keep the tank near a constant desired volume? (c) Suppose the inflow rate is periodic. To be definite, take the flow rate to be sinusoidal and known exactly, how should the constant of proportionality be set for the controller to best keep the tank at a constant desired volume. Part of the problem is to define “best.” Explain your choice. The abbreviation PID stands for proportional-integral-derivative. This exercise is about a proportional controller. It is a feed-back controller where the feedback is proportional to the volume as a function of time. Adding into the feedback a constant times the integral of the volume change would create a PI control. Adding into the feedback a constant
times the derivative of the volume function would create a PD control. A feed-back control given by a linear combination of the volume as a function of time, the integral of the volume, and the derivative of the volume creates a PID control. (d) Find the constants of proportionality for the PID control of the tank volume that tunes the control to best keep the tank volume constant for the case of constant inflow and for periodic inflow.

4.2 Acid Dissociation, Buffering, and Titration with a Base

Dissociation is a basic chemical process that is amenable to a mathematical description. We will discuss the dissociation of acetic acid in water, titration with sodium hydroxide, and buffering.

4.2.1 A Model for Dissociation

In a beaker of water, some water molecules ionize; that is, a water molecule may dissociate producing a proton and an ion consisting of an oxygen atom with one hydrogen attached. In chemical symbols

\[ H_2O \rightleftharpoons H^+ + OH^- \]

where the arrows indicate that the reaction occurs in both directions: water dissociates and the ions may recombine to form water.

Similarly, acetic acid dissociates in water

\[ CH_3COOH + H_2O \rightleftharpoons CH_3COO^- + H^+; \]

or, in shorthand,

\[ AcH \rightleftharpoons Ac^- + H^+ \]

with \( Ac^- \) used here to denote the acetate ion \( CH_3COO^- \). (Note: acetate is also denoted by \( AcO^- \).)

A simple chemistry experiment is to put some acetic acid in water. What happens?

There are four chemical species of interest: acetic acid, acetate, protons, and OH ions:

\[ AcH, \quad Ac^-, \quad H^+, \quad OH^- \]
The $OH^-$ ion is produced from the dissociation of water and is lost by association. The rate of change of the concentration $[OH^-]$ is the difference of some constant $W_f$ (forward water rate constant) times the concentration of water and some rate constant $W_f$ times the product of the proton and hydroxide ions. The latter statement is a model that comes from chemical kinetic theory. For simplicity, we may view the product of the concentrations as a measure of the number of times the ions meet each other in the solution. The principle of mass action states that the rate of reaction of two chemical species is proportional to the product of their concentrations. These assumptions lead to the equations in the following system:

\[
[OH^-]' = W_f[H_2O] - W_b[H^+][OH^-],
\]

\[
[AcH]' = a_b[H^+][Ac^-] - a_f[AcH][H_2O],
\]

\[
[Ac^-]' = a_f[AcH][H_2O] - a_b[H^+][Ac^-],
\]

\[
[H^+]' = a_f[AcH][H_2O] - a_b[H^+][Ac^-] + W_f[H_2O] - W_b[H^+][OH^-].
\]

A complete model would also contain an equation for the rate of change of water concentration. But, we may ignore this equation in our model for the dissociation of the acid dissolved in a relatively large amount of water. There is lots of water compared with the other species and it is known that water does not dissociate easily. Its concentration will remain constant for the measurements we might consider in our experiment. This is a good example of a simplifying assumption in mathematical modeling. Under our assumption, the model is less realistic but perhaps more useful. To implement this idea, we define the parameters (quantities that for our model do not change with time)

\[
w_f := W_f[H_2O], \quad a_f := A_f[H_2O], \quad w_b = W_b, \quad a_b = A_b
\]

and rewrite the model in the form

\[
[OH^-]' = w_f - w_b[H^+][OH^-],
\]

\[
[AcH]' = a_b[H^+][Ac^-] - a_f[AcH],
\]

\[
[Ac^-]' = a_f[AcH] - a_b[H^+][Ac^-],
\]

\[
[H^+]' = a_f[AcH] - a_b[H^+][Ac^-] + w_f - w_b[H^+][OH^-].
\]

With our simplification, we have a system of four first-order ordinary differential equations in four unknowns. By the existence theory for ordinary
differential equations, there is a unique solution of the system corresponding to each choice of initial concentrations (see Appendix A.3).

Mathematicians generally do not like all the extra decorations on the species names, and they do not care to be reminded about what the symbols represent. Let us follow this convention by defining

\[ w = [OH^-], \quad x = [AcH], \quad y = [Ac^-], \quad z = [H^+] \]

and

\[ a = a_f, \quad b = a_b, \quad c = w_f, \quad d = w_b \]

so that our model has the more aesthetically pleasing form

\[
\begin{align*}
    w' &= c - dwz, \\
    x' &= byz - ax, \\
    y' &= ax - byz, \\
    z' &= ax - byz + c - dwz.
\end{align*}
\]

(4.4)

The fundamental mathematical question is to describe the behavior of this system (4.4) for arbitrary choices of the parameters and arbitrary initial data. As applied mathematicians, we should perhaps not be so general; we should restrict attention to the physically relevant parameter values and initial data. Chemists should supply this data.

Unfortunately, the rate constants in our model are difficult to measure directly. Instead, the basic experimental measurements are made at equilibrium and involve ratios of concentrations. For example, the dissociation constant for water is

\[ K_w := \frac{[OH^-][H^+]}{[H_2O]} \approx 10^{-14}. \]

Its value implies that almost no water molecules dissociate. At steady state, the concentrations of our species are not changing; therefore, we must have

\[ W_f[H_2O] - W_b[H^+][OH^-] = 0. \]

A rearrangement of this equality is

\[ K_w = \frac{W_f}{W_b}. \]

(4.5)
We are given a ratio of rate constants, not the rate constants themselves. A reasonable choice of the rate constants might be
\[ c = w_f = W_f[H_2O] = 10^{-14}, \quad d = w_b = W_b = 1. \]

For acetic acid
\[ \mathbb{K}_a = \frac{[Ac^-][AcH]}{[H^+]} \approx 10^{-4.75}. \]

The number
\[ p\mathbb{K}_a := -\log_{10} \mathbb{K}_a \approx 4.75 \]
is also useful. There is no obvious way to choose the forward and backward rate constants. Thus, some chemical intuition is required to make a reasonable choice, for example,
\[ a = a_f = A_f[H_2O] = 10^5, \quad b = A_b = A_f[H_2O]/\mathbb{K}_a \]
so that again \( \mathbb{K}_a = A_f/A_b. \)

We are using parameters in our dynamic model derived from experiments at equilibrium. Is this justified? Perhaps this is a good time to emphasize that no mathematical model is a perfect representation of reality. If the model confirms other experimental data, we can be reasonably certain that the underlying chemistry (including rate constants, dissociation equations, first-order reactions, and constant water concentration assumption) is correct. We could then make predictions of the outcomes of new experiments with some confidence that the predictions will agree with nature. Of course, the final word in science is determined by experiment. On the other hand, we have very strong reasons to believe the universe is rational: if we start with fundamental laws and make logical deductions (for example, by applying mathematics), the conclusions will agree with new experiments. This is one of the main reasons why mathematics is so important in science.

Rest points—solutions of our model that do not change with time—correspond to states in physical equilibrium. The rest points are exactly the solutions of the equations
\[ c - dwz = 0, \quad byz - ax = 0, \quad ax - byz = 0, \quad ax - byz + c - dwz = 0. \]

Simple algebraic manipulation shows that this nonlinear system has infinitely many solutions that are given by the relations
\[ x = \frac{byz}{a}, \quad w = \frac{c}{dz}. \]
that is, there is a two-dimensional surface of rest points, which may be parametrized by the two variables \( y \) and \( z \). These rest points cannot be asymptotically stable (see Appendix A.17) for a simple reason. Each rest point lies on an entire surface of rest points. Nearby rest points are not attracted; they stay fixed. But, we may still ask about the destination of initial states that are not rest points. Are they attracted to rest points as time approaches infinity? To obtain some insight, let us linearize as usual and see what happens.

The Jacobian matrix of the vector field is given by

\[
\begin{pmatrix}
-dz & 0 & 0 & -dw \\
0 & -a & bz & by \\
0 & a & -bz & -by \\
-dz & a & -bz & -by - dw
\end{pmatrix}.
\]

At a rest point \( w = c/(dz) \); thus, we would like to find the eigenvalues of the matrix

\[
J := \begin{pmatrix}
-dz & 0 & 0 & -\frac{c}{z} \\
0 & -a & bz & by \\
0 & a & -bz & -by \\
-dz & a & -bz & -by - \frac{c}{z}
\end{pmatrix}.
\]

There are several ways to proceed. Perhaps the most instructive method uses basic geometry and linear algebra. The rest points lie on a two-dimensional surface of rest points in four dimensional space. The function given by

\[
(y, z) \mapsto \left( \frac{c}{dz}, \frac{byz}{a} : y, z \right)
\]

parameterizes the surface. If we hold \( z \) fixed and let \( y \) move, the function traces out a curve on the surface of rest points whose velocity vector must be tangent to the surface. Likewise, we may hold \( y \) fixed and consider the tangent vector along the corresponding curve in the surface. These vectors (obtained by differentiation with respect to \( y \) and \( z \), respectively) are

\[
\begin{pmatrix}
bz/a \\
1 \\
0
\end{pmatrix}, \quad \begin{pmatrix}
-\frac{c}{dz} \\
bz/a \\
0 \\
1
\end{pmatrix}.
\]

We should expect that the linearization of our system of differential equations approximates the dynamics of the nonlinear system. In the directions of
the two-velocity vectors the solutions should not move; thus, these vectors should be eigenvectors of \( J \) with zero eigenvalues. This fact is easily verified by simply multiplying each vector by \( J \). Thus, we know that there are two zero eigenvalues. There are exactly two more eigenvalues. These have to be roots of the characteristic polynomial \( p \) of the matrix \( (p(\lambda) = \det(J - \lambda I)) \). It is a polynomial of degree four. But, since it has two zero roots, it will be divisible by \( \lambda^2 \) and the desired roots are simply solutions of a quadratic polynomial. Under our hypotheses, all the coefficients of this polynomial, given by

\[
p(\lambda) = ac + bcz + adz^2 + bdyz^2 + bdz^3 + (c + az + byz + bz^2 + dz^2)\lambda + z\lambda^2,
\]

are positive. Thus, the real parts of its roots are negative (see Exercise 4.8).

We have not proved that every solution converges to a rest point, but this conclusion is supported by the linearization. In fact, every physically relevant solution of our model system is asymptotic to exactly one steady state as time grows without bound. This result requires some phase plane analysis and hypotheses derived from the physical problem. The parameters \( a, b, c \) and \( d \), as computed previously, vary greatly in size. Most importantly, \( c \) is very small compared to the other parameters and \( d < b \). This is our first assumption. Note also that the concentration of acetate \( y \), which is formed by dissociation, cannot exceed the initial concentration of acetic acid \( x_0 \); that is, \( y \leq x_0 \). This is our second assumption. We will of course assume nonnegative concentrations and positive parameters.

The set of nonnegative states \( \mathcal{N} \) is positively invariant for system (4.4). For instance, on the coordinate hyperplane \( \mathcal{H} \) corresponding to the coordinates \( x, y, \) and \( z \) (that is the set of all states whose first coordinate \( w \) is zero), the tangent vectors to solutions of the differential equation at a point on \( \mathcal{H} \) are given by the transpose of a vector of the form

\[
(c, byz - ax, ax - byz, ax - byz + c).
\]

The first coordinate of such a vector is not zero. Thus, this vector is not in the tangent space of the hyperplane \( \mathcal{H} \); the vector points into the region \( \mathcal{N} \). In other words, it is impossible for a solution that starts in \( \mathcal{N} \) to exit this set through the hyperplane \( \mathcal{H} \). A similar argument (which might require noting that on the boundary of \( \mathcal{N} \) the coordinates are always nonnegative) applied to each of the coordinate hyperplanes bounding \( \mathcal{N} \) can be used to show the same result: the vector field corresponding to the system of differential equations
Conservation of Mass

points into the region $\mathcal{N}$ on its boundary. Thus, this region is positively invariant, as it should be for our mathematical model. Concentrations cannot become negative. Inspection of system (4.4) suggests that we add the second and third equations. This yields the identity $x' + y' = 0$. Hence, there must be a constant $c_1$ such that $x + y = c_1$. Similarly, there is a constant $c_2$ such that $z - y - w = c_2$. In other words, the functions $(w, x, y, z) \mapsto x + y$ and $(w, x, y, z) \mapsto z - y - w$ stay constant along all solutions. Such a function is called a first integral of the system. In our case, we consider only a small set of initial conditions: acid is added to water at the start of the experiment. Thus, at time $t = 0$, the water has partially dissociated but the acid is all associated and the initial data is taken to be such that

$$w(0) > 0, \quad w(0) = z(0), \quad x(0) = x_0 > 0, \quad y(0) = 0.$$  

Using this information, we have that $c_1 = x(0)$ and $c_2 = 0$; therefore, our evolving states satisfy the relations

$$x(t) = x(0) - y(t), \quad w(t) = z(t) - y(t). \quad (4.9)$$

Clearly, we may reduce our system of four differential equations to

$$y' = a(x(0) - y) - byz, \quad z' = c + a(x(0) - y) - z(by + d(z - y)) \quad (4.10)$$

and, if desired, recover the states $x$ and $w$ from the relations (4.9).

We have proved that the evolving states, with physical initial values, must remain positive. Repeating a similar argument for system (4.10) and using our assumption that $y \leq x(0)$, it follows that the region bounded by the horizontal $y$-axis, the vertical axis, and the line with equation $y = x(0)$ is positively invariant. For $z$ sufficiently large, $z' < 0$ in this strip because the dominant term is $-dz^2$. Thus, we have proved that system (4.10) has a positively invariant rectangle $\mathcal{R}$. Moreover, we can arrange the choice of the upper boundary of $\mathcal{R}$ so that all solutions starting above it eventually enter $\mathcal{R}$. We may therefore restrict attention to this rectangle.

We would like to show that all solutions are asymptotic to a rest point in $\mathcal{R}$.

Are there rest points in $\mathcal{R}$? The answer is obtained by solving the system of equations

$$a(x(0) - y) - byz = 0, \quad c + a(x(0) - y) - z(by + d(z - y)) = 0.$$
The first equation may be rearranged to
\[ z = \frac{a(x(0) - y)}{by}. \] (4.11)

The graph of the corresponding function (in \( \mathbb{R} \)) contains the point with coordinates \((y, z) = (x(0), 0)\) (which is the lower right corner of the rectangle), is asymptotic to the vertical axis, and has negative slope. The second equation may be rearranged to the form
\[ y = \frac{c + ax(0) - dz^2}{a + bz - dz}. \] (4.12)

The denominator is not zero in \( \mathbb{R} \) because we have assumed \( b > d \). We may consider substitution of this relation into the equation (4.11), which leads to a cubic equation for the unknown \( y \) coordinate of the rest points we seek. For \( c = 0 \), this cubic can be solved explicitly. One root is \( y = x(0) \), one root is negative, and the third root (which is also the positive root of the equation \((a(x(0) - y)/(by) - y = 0)\) lies in \( \mathbb{R} \), as it must because the graph of the relation (4.12) is also the graph of a function with respect to \( z \), it crosses the vertical axis at \((0, \sqrt{ax(0)/d})\), and it crosses the horizontal axis at \( x(0) \). Moreover, the slope of this graph at the intersection point of the relations (4.11) and (4.12) in \( \mathbb{R} \) is positive. For \( c \) sufficiently small, the same configuration must be true with the new graph of the relation (4.12) crossing the horizontal axis at \( c/a + x(0) \), which lies to the right of \( x_0 \). In other words, there is exactly one rest point in the rectangle. To show this is true for the specific values of the parameters of our problem takes more work; we must do better than the statement “for \( c \) sufficiently small.”

Our rest point for \( c = 0 \) is the intersection of the graphs of the relations
\[ \frac{a(x(0) - y)}{by} - y = 0, \quad z = \frac{a(x(0) - y)}{by}. \]

In particular, for \( c = 0 \), our rest point lies on the line \( y = z \). The linearization at the rest point has system matrix
\[
\begin{pmatrix}
-a - by & -by \\
-a - by + dy & -by - dy
\end{pmatrix}.
\]

Its trace is negative and its determinant is positive. Using the characteristic equation for \( 2 \times 2 \)-matrices, which is given by \( \lambda^2 - \text{tr}(A)\lambda + \det(A) = 0 \), it
Conservation of Mass

follows that both eigenvalues have negative real parts. Thus, the rest point is asymptotically stable.

Unfortunately, we can conclude only that nearby solutions are attracted to the asymptotically stable rest point. There might be limit cycles surrounding the rest point that prevent solutions from reaching the rest point as time grows without bound. By the Poincaré-Bendixson theorem (see Appendix A.18), this is the only possible obstruction.

For $c = 0$, the line $y = z$ is invariant and the unique asymptotically stable rest point lies on this line. It is not possible for a limit cycle to exist. It must surround the rest point, but it cannot cross this invariant line. Thus, the phase portrait for $c = 0$ is simple, there is exactly one hyperbolic rest point in (the closed and bounded set) $\mathcal{R}$, this rest point is globally attracting, and the corresponding vector field points into the region $\mathcal{R}$ on the boundary of $R$. This is a (structurally) stable configuration. For sufficiently small $c$, the perturbed phase portrait has the same properties. This result (called the Structural Stability Theorem) is intuitively clear, but its proof is beyond the scope of this book.

Fortunately, there is another simpler proof that there are no periodic orbits in the set $\mathcal{R}$. Note first that the divergence of the vector field $X$ that determines the model equation (4.10) is

$$-a - bz - 2dz - (b - d)y. \quad (4.13)$$

Under our assumption that $b > d$ and taking into account that both $y$ and $z$ are positive in $\mathcal{R}$, it follows that the divergence of the vector field is negative everywhere in this set.

The vector field $X$ has two components $P$ and $Q$; indeed,

$$X = (P(y, z), Q(y, z)), \quad (y', x') = (P(y, z), Q(y, z)), \quad (P(y, z), Q(y, z)) = \left(a(x(0) - y) - byz, c + a(x(0) - y) - z(by + d(z - y))\right).$$

Suppose there is a periodic orbit that is the boundary of the set $\Omega$ contained in $\mathcal{R}$. Every periodic orbit in the plane is the boundary of a closed and bounded set that is a region in the plane with no holes. Green’s theorem states that

$$\int_\Omega P_y(y, z) + Q_z(y, z) \, dydz = \int_{\partial\Omega} P(y, z) \, dz - Q(y, z) \, dy.$$
The area integral must be negative because \( P_y(y,z) + Q_z(y,z) \) is the divergence of \( X \). The line integral vanishes. To see this, simply parameterize the boundary \( \partial \Omega \) by a periodic solution of the differential equation \((y(t), z(t))\). Let \( T > 0 \) be its period. We then have that

\[
\int_{\partial \Omega} P(y,z) \, dz - Q(y,z) \, dy = \int_0^T y'(t) z'(t) - z'(t)y'(t) \, dt = 0.
\]

This contradiction proves that there are no periodic orbits in \( \mathcal{R} \). The corresponding general result: *If a vector field has a periodic integral curve in the plane and the vector field is defined on the entire planar region bounded by this curve, then the divergence of the vector field must change sign in the region bounded by the curve.* is called Bendixon’s theorem.

We have outlined a proof of the desired result: for physical initial data and system parameters, every solution of system (4.4) is asymptotic to a rest point that depends on the initial data. In fact, the choice of \( x(0) \) determines the steady state (cf. Exercise 4.10).

**Exercise 4.6.** Consider the prototype reaction \( A + B \to C \) with forward rate constant \( k \). (a) Determine the steady state behavior of the concentration of \( C \). (b) Find an explicit formula for the concentration of \( C \) as a function of time. (c) What is the steady state behavior in case the reverse reaction \( C \to A + B \) with rate constant \( \ell \) is included?

**Exercise 4.7.** [Michaelis-Menton Enzyme Kinetics] An enzyme is catalyst for a biochemical reaction. In a typical situation there is a substrate \( S \), an enzyme \( E \), and a product \( P \). In the presence of the enzyme (perhaps a protein) biochemical elements that compose the substrate combine to form the product; in symbols, \( S + E \to E + P \). This is an example of a situation where the principle of mass action leads to a prediction that is inconsistent with experimental evidence.

(a) By applying the principle of mass action, show that the rate of change of the concentration of the product grows (without bound) as the concentration of substrate is increased and enzyme concentration is held constant.

A better model was proposed in 1913 by Lenor Michaelis and Maud Menton. They hypothesized the existence of an intermediate substance \( I \) included in the chain of reactions

\[
S + E \rightleftharpoons I \to E + P.
\]

(b) Suppose the forward and backward rate constants for the first reaction are \( k_+ \) and \( k_- \) and the forward rate constant for the section reaction is \( \ell \). Write the rate
equations for the complete reaction. Hint: There are four species and hence four rate equations.

We wish to know the rate of change of the product concentration as a function of the substrate concentration and the initial enzyme concentration.

(c) Under the further assumption that the substrate concentration is (nearly) constant, determine the rate of change of the product concentration as a function of the substrate concentration and the initial enzyme concentration.

In 1925, G. E. Briggs and J. B. S. Haldane proposed a modification of the Michaelis-Menten substrate concentration assumption: The intermediate product concentration is (nearly) constant.

(d) Determine the rate of change of the product concentration as a function of the substrate concentration and the initial enzyme concentration under the Briggs-Haldane assumption.

(e) Compare and contrast the assumptions and results of parts (c) and (d).

Exercise 4.8. Prove that the roots of the polynomial (4.8) (for the dissociation model) have negative real parts. Are the roots real?

Exercise 4.9. Prove that every solution of the system (4.4) starting in the physical space where initial concentrations are not negative converges to a unique rest point. A possible strategy is to first reduce the first-order system to a lower dimensional system by using conserved quantities, for example, \( x' + y' = 0 \).

Exercise 4.10. (a) Specify a method to obtain the steady state of a solution of acetic acid and water (where all the pure acetic acid is assumed to be added all at once to water) that is already in steady state as predicted by the model (4.4) without solving the system of differential equations. Your result should determine a function of the initial concentration of acetic acid at the instant it is added to the water. (b) Using a numerical method, approximate the length of time (in seconds) for the solution to reach steady state after the addition of one mole of acetic acid? According to the model, the solution never reaches steady state. For this exercise, assume that steady state has been reached once the final concentrations are within 0.1% of their theoretical steady state values. (c) Discuss how the time to reach steady state is related to the amount of added acetic acid. Does the time to steady state increase or decrease with an increase in the amount of acetic acid? Discuss your answer. Does your result agree with physical intuition?

4.2.2 Titration with a Base

Imagine the following experiment: A beaker contains a solution of acetic acid and water that is made by combining a known amount of acetic acid with
a known volume of water. A known concentration of sodium hydroxide in water, stored in a graduated pipe ( buret), is released slowly into the beaker, the solution is stirred continuously, and the pH of the solution is measured at specified time intervals. We will model the pH in the beaker as a function of time.

What is pH? Recall that the pH of a solution is a measure of its acidity. In fact, it is defined to be the negative logarithm, base ten, of the concentration of protons \( H^+ \) (in units of mol / l) or

\[
pH := - \log_{10}[H^+].
\]

For sodium hydroxide in water, we have

\[
NaOH \rightleftharpoons Na^+ + OH^-.
\]

The forward dissociation constant is very large compared with the association constant. Almost all of the sodium hydroxide dissociates to sodium and \( OH^- \) ions. The sodium does not react with acetic acid. Thus, we are only interested in the \( OH^- \) ions. A solution of \( A \) mol / l of \( NaOH \) has \( A \) mol / l of \( OH^- \) for “all practical purposes.” Let us assume that all of the \( NaOH \) dissociates. The solution enters at a rate of \( B \) ml / sec. We must not mix units. Thus, we should express this concentration as \( B/1000 \) l / sec. It follows that \( OH^- \) ions enters the acetic acid solution at the rate of

\[
AB \text{ mol / sec}.
\]

The concentration of this ion changes during the addition of \( NaOH \) because the volume is changing from 100 ml via the function

\[
\text{vol} = \frac{1}{10} + \frac{B}{1000} t.
\]

A model of the ion concentration entering the solution (measured in mol / l) as a function of time is

\[
\frac{AB}{\left(\frac{1}{10} + \frac{B}{1000} t\right)^2}.
\]

(4.14)

Using the model (4.14) for addition of \( OH^- \) (during the addition of the
Conservation of Mass

<table>
<thead>
<tr>
<th>sec</th>
<th>pH</th>
<th>sec</th>
<th>pH</th>
<th>sec</th>
<th>pH</th>
<th>sec</th>
<th>pH</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.567</td>
<td>16</td>
<td>3.590</td>
<td>32</td>
<td>4.196</td>
<td>48</td>
<td>4.490</td>
</tr>
<tr>
<td>2</td>
<td>2.563</td>
<td>18</td>
<td>3.738</td>
<td>34</td>
<td>4.240</td>
<td>50</td>
<td>4.514</td>
</tr>
<tr>
<td>4</td>
<td>2.563</td>
<td>20</td>
<td>3.852</td>
<td>36</td>
<td>4.283</td>
<td>52</td>
<td>4.540</td>
</tr>
<tr>
<td>6</td>
<td>2.563</td>
<td>22</td>
<td>3.923</td>
<td>38</td>
<td>4.327</td>
<td>54</td>
<td>4.564</td>
</tr>
<tr>
<td>8</td>
<td>2.689</td>
<td>24</td>
<td>3.979</td>
<td>40</td>
<td>4.363</td>
<td>56</td>
<td>4.590</td>
</tr>
<tr>
<td>10</td>
<td>3.038</td>
<td>26</td>
<td>4.042</td>
<td>42</td>
<td>4.396</td>
<td>58</td>
<td>4.607</td>
</tr>
<tr>
<td>12</td>
<td>3.279</td>
<td>28</td>
<td>4.094</td>
<td>44</td>
<td>4.436</td>
<td>60</td>
<td>4.632</td>
</tr>
<tr>
<td>14</td>
<td>3.418</td>
<td>30</td>
<td>4.148</td>
<td>46</td>
<td>4.462</td>
<td>62</td>
<td>4.656</td>
</tr>
</tbody>
</table>

Figure 4.3: The data in this table is produced by a pH meter while a one molar sodium hydroxide solution is added to a one molar solution of acetic acid and water.

NaOH solution), our basic model (4.4) is modified to

\[
\begin{align*}
    w' &= c - dwz + \frac{AB}{(\frac{1}{10} + \frac{B}{1000}t)}, \\
    x' &= byz - ax, \\
    y' &= ax - byz, \\
    z' &= ax - byz + c - dwz.
\end{align*}
\]

(4.15)

The flow rate \( B \) does not remain constant. The flow is turned on after the acetic acid solution is in steady state and turned off when the measurement device is stopped or the reservoir of NaOH is empty. The model can be used to make predictions via numerical integration.

An experiment was performed where a solution of one mol/l of NaOH is added to a solution of one mol/l of acetic acid at the rate of 0.5 ml/sec. The beaker containing the solution was stirred continuously and a pH sensor recorded the pH of the solution every two seconds for 74 sec (see Table 4.3). The flow of NaOH was started at 5 sec. The results of this experiment and a simulation using the model (4.15) are depicted in Figure 4.4. The agreement is excellent (compare, Exercise 4.11).

We note that the solution pH does not rise significantly during the titration if not too much base is added. The presence acid buffers the solution to the addition of the strong base. Buffering is an essential process in the living organisms. The pH of essential fluids, for example blood, must remain within some narrow range for certain biochemical processes to work properly. While our model is viable for small concentrations of the base, sodium hydroxide, it does not seem to predict the correct titration curve (pH versus the amount
Figure 4.4: Plots of pH versus time in seconds for the addition of NaOH in acetic acid. The continuous graph is from the mathematical model (4.15); the discrete graph is from the data in Table 4.3.

of added base) over a larger range of additional $OH^-$. A better model is the subject of the next section.

**Exercise 4.11.** (a) The results of experiment and simulation using the mathematical model (4.15) depicted in Figure 4.4 show excellent agreement. On the other hand, transient in the solutions of the model are very short. Good results can also be obtained by a static model based on the assumption that the mixtures are instantaneously in steady state. Discuss this alternative. Hint: This approach is used in textbooks on basic chemistry. (b) The accuracy of predictions from the model should depend on the accuracy of its parameters. How sensitive is the result to the given parameters. (c) Which choice of parameters best fits the experimental data?

**Exercise 4.12.** Prove the generalization of Bendixson’s theorem due to Dulac: If a vector field has a periodic integral curve in the plane and the vector field is defined on the entire planar region bounded by this curve, then every positive function multiple of the vector field must have its divergence change sign in the region bounded by the curve. This result is useful when a given vector field $X$ does not have fixed sign in a region to be tested for the existence of periodic orbits. If there is a positive function $f$ defined on the region, and the new vector field $fX$ has fixed sign, then there are no periodic orbits in the region.
4.3 An Improved Titration Model

The titration of a weak acid by a strong base (for example, titration of acetic acid by sodium hydroxide) stays in the buffer region, where the change in pH is not too great for small additions of the base until some critical amount of base is added. At this point a sharp rise in pH occurs. The pH levels off as more base is added. This general behavior is typical and is easily observed in experiments. Appropriate mathematical models build from rate equations, which are discussed in this section, predict the observed behavior.

Recall that we have introduced the chemical reactions for the dissociation of water

\[ H_2O \rightleftharpoons H^+ + OH^- \]

and acetic acid

\[ CH_3COOH + H_2O \rightleftharpoons CH_3COO^- + H^+ \]

or, in shorthand,

\[ AcH + H_2O \rightleftharpoons Ac^- + H^+ \]

and sodium hydroxide

\[ NaOH \rightleftharpoons Na^+ + OH^- \]

The water concentration is considered to be very large compared with the concentrations of the other species. To account for the additional OH\(^{-}\) ions from the dissociated sodium hydroxide, one additional reaction should be included:

\[ CH_3COOH + OH^- \rightleftharpoons CH_3COO^- + H_2O. \tag{4.16} \]

In the previous section, we ignored the rate constants in this equation and simply changed the OH\(^{-}\) concentration in the mixture.

Using the notation of system (4.4), incorporating the water concentration into the forward rate constant for the dissociation of water and the backward rate constant \(e\) for reaction (4.16), and using \(f\) for the forward rate constant in the reaction (4.16), we have the rate equations

\[
\begin{align*}
w' &= c - dwz + ey - fxw, \\
x' &= byz - ax + ey - fxw, \\
y' &= ax - byz - ey + fxw, \\
z' &= ax - byz + c - dwz. \tag{4.17}
\end{align*}
\]
We do not know the exact values of the rate constants $e$ and $f$. To determine their ratio, let us recall that

$$K_w = \frac{W_f}{W_b} \approx 10^{-14}$$

(see, equation (4.5)). At steady state for the system (4.17),

$$W_f[H_2O] - W_b[H^+][OH^-] = C_f[OH^-][AcH] - C_b[H_2O][Ac^-]$$

A rearrangement of the formula is

$$C_fK_w - C_bK_a = \frac{[H^+]}{[AcH]}(W_f - w_bK_w) = 0.$$ 

Thus, we have

$$\frac{e}{f} = \frac{C_f}{C_b} = \frac{K_a}{K_w} \approx 10^9.$$ 

Using the rate constants

$$a = 10^5, \quad b = \frac{a}{10^{-4.75}}, \quad c = 10^{-14}, \quad d = 1, \quad e = 10^{-9}, \quad f = \frac{e}{10^{-9}},$$

the titration curve depicted in Figure 4.5 is computed by first integrating system (4.17) forward with initial acetic acid concentration 0.1 mol / l until an equilibrium pH is reached, which is approximately 2.88. The equilibrium values of the states are used as initial conditions for the states during the
titration simulation (see Exercise 4.14). An initial $OH^-$ concentration is set and the states are evolved forward in time (by numerical integration) until (an approximation of) equilibrium is reached. The corresponding pH is then plotted as a function of the initial $OH^-$. The figure shows a buffering region with relatively unchanged pH for low concentrations of the base, a region of rapid pH increase, and a region of high pH for the addition of high concentrations of the base.

The steep change in pH in Figure 4.5 is not an obvious prediction from the model. Why does this happen? To explain this phenomenon without numerical approximations, let us consider the steady states in more detail.

The dynamical system (4.17) has two first-integrals:

\[ x + y = 0, \quad z - y - w \]

(see the discussion on page 32.) Using the chemistry at the initial time—which is after the acid solution is in equilibrium and at the instant the base is added, we have that $y = x(0) - x$ and $w = z - y + w(0)$. We may reduce our model to the two-dimensional system

\[ \dot{y} = a(x(0) - y) - byz - cy + f(x(0) - y)(z - y + w(0)), \]
\[ \dot{z} = a(x(0) - y) - byz + c - dz(z - y + w(0)). \]

Using the parameter values

\[ a = 10^5, \quad b = 10^{10}, \quad c = 10^{-14}, \quad d = 1, \quad e = 10^{-9}, \quad f = 1. \]

The equations for the steady states—the right-hand sides of the differential equations set to zero—have some small and some large coefficients. Inspection of these coefficients suggests defining $\mu$, a new parameter that has value $10^{-5}$ in our model, so that the coefficients are expressed in the form

\[ a = 1/\mu, \quad b = 1/\mu^2, \quad c = 10\mu^3, \quad d = 1, \quad e = 10\mu^2, \quad f = 1. \]

We may treat $\mu$ as an order parameter (which orders the relative sizes of the coefficients) and reduces the number of parameters. The equations for the steady states are equivalent to

\[ -10yz + (1 - 10y)\mu + (w(0) - y - 10w(0)y + 10y^2 + z - 10yz)\mu^2 - 100y\mu^4 = 0, \]
\[ -10yz + (1 - 10y)\mu - (10w(0)z - 10yz + 10z^2)\mu^2 + 100\mu^5 = 0. \]

(4.20)
We may solve for the state variable $y$ in the second equation, where it appears linearly, and substituting this solution into the first equation. A minor miracle occurs: the equation for $z$ factors and the desired steady state is the positive root of the cubic polynomial

$$p(z) := -10w(0)z^2 - 10z^3 + z\mu - 10w(0)z\mu - 10z^2\mu + 100z\mu^3 + 100\mu^4. \quad (4.21)$$

The pH is $-\log_{10} z$. The derivative of this expression with respect to $w(0)$ is

$$-\frac{1}{z} \frac{dz}{dw(0)} = \frac{10(z + \mu)}{20w(0)z + 30z^2 + (10w(0) - 1)\mu + 20z\mu - 100\mu^3}.$$

The rate of change of the pH with respect to $w(0)$ is large near values of $z$ that are nearly roots of both the cubic polynomial $p$ and its derivative with respect to $z$. If we ignore terms of order at least four in $\mu$, then $z = 0$ is such a point provided that $w(0) = 1/10$. Thus, for $z$ near zero (as it must be when the base and the acid are both close to being neutralized), we expect the pH to have a large derivative for $w(0)$ near 1/10, which explains the numerical result depicted in Figure 4.5 (compare, Exercise 4.16).

The kinetic theory expressed in the models discussed here predicts phenomena that are in complete agreement with physical experiments. Thus, we may have a high level of confidence that the underlying theory is correct. To make such a statement about a class of mathematical models is satisfying; the validation (and invalidation) of models is of course a strong motivation for doing applied mathematics.

**Exercise 4.13.** Redraw Figure 4.4 using the improved titration model. Is there a difference?

**Exercise 4.14.** Redraw Figure 4.5 using an alternative method: do not integrate the differential equations; solve for the steady states. Which method is more efficient?

**Exercise 4.15.** Redraw Figure 4.5 using the cubic polynomial (4.21) to solve for $z$ as a function of $w(0)$.

**Exercise 4.16.** (a) The roots of the cubic polynomial $p$ in equation (4.21) may be determined explicitly using Cardano’s formula. Can it be used to explain the shape of titration curves? (b) It is possible to determine the discriminant of a cubic. It is zero exactly when the cubic has a multiple root. The discriminant is
also equal to the resultant of the cubic and its first derivative. Find the lowest order approximation in powers of \( \mu \) for the discriminant of \( p \) and show that it vanishes for \( w(0) = 1/10 \). Hint: Learn the meaning of the concepts mentioned in the problem: discriminant and resultant.

**Exercise 4.17.** A titration of 0.1 molar acetic acid with sodium hydroxide is discussed in the section. The point of mutual neutralization occurs for the addition of approximately 0.1 mole of the base. Repeat the analysis, using the same methods, for titration of 0.3 molar acetic acid with sodium hydroxide.

**Problem 4.18.** The titration curve depicted in Figure 4.5 is obtained by a static model; that is, we set the new \( \text{OH} \) concentration, integrate to steady state, determine the pH and iterate this process. A dynamic model would predict results for a titration where the sodium hydroxide solution is poured continuously into the acetic acid solution. Model (4.15) is a first approximation of such a model. Created a model that can be used to predict titration curves consistent with Figure 4.5.

### 4.4 Reaction, Diffusion, and Convection

Conservation of mass is a fundamental physical law used to model many physical processes.

#### 4.4.1 Fundamental and Constitutive Model Equations

Let us suppose that some substance (considered to have mass) is distributed in \( \mathbb{R}^n \) and let \( \Omega \) denote a bounded region in \( \mathbb{R}^n \) with boundary \( \partial \Omega \) and outer normal \( \eta \). The density of the substance is represented by a function \( u : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R} \), where \( u(x,t) \) is the numerical value of the density in some units of measurement at the site with coordinate \( x \in \mathbb{R}^n \) at time \( t \).

The time rate of change of the amount of the substance in \( \Omega \) is given by the negative flux of the substance through the boundary of \( \Omega \) plus the amount of the substance generated in \( \Omega \); that is,

\[
\frac{d}{dt} \int_{\Omega} u \, d\mathcal{V} = -\int_{\partial \Omega} X \cdot \eta \, dS + \int_{\Omega} f \, d\mathcal{V},
\]

where \( X \) is the velocity vector field (sometimes called the diffusion flux) of the substance (with units of mass per area per time); \( d\mathcal{V} \) is the volume element; \( dS \) is the surface element; and the vector field \( \eta \) is the outer unit normal.
field on the boundary of $\Omega$; and $f$, a function of density, position and time, represents the amount of the substance generated in $\Omega$ per volume per time. The minus sign on the flux term is required because we are measuring the rate of change of the amount of substance in $\Omega$. If, for example, the flow is all out of $\Omega$, then $X \cdot \eta \geq 0$ and the minus sign is required because the rate of change of the amount of substance in $\Omega$ must be negative.

Using the divergence theorem (also called Gauss’s theorem) to rewrite the flux term and by interchanging the time derivative with the integral of the density, we have the relation

$$\int_{\Omega} u_t \, dV = - \int_{\Omega} \text{div} \, X \, dV + \int_{\Omega} f \, dV.$$ 

Moreover, because the region $\Omega$ is arbitrary in this integral identity, it follows that

$$u_t = - \text{div} \, X + f. \quad (4.22)$$

To obtain a useful dynamical equation for $u$ from equation (4.22), we need a constitutive relation between the density $u$ of the substance and the flow field $X$. It is not at all clear how to derive this relationship from the fundamental laws of physics. Thus, we have an excellent example of an important problem where physical intuition must be used to propose a constitutive law whose validity can only be tested by comparing the results of experiments with the predictions of the corresponding model. Problems of this type lie at the heart of applied mathematics and physics.

For equation (4.22), the classic constitutive relation—called Darcy’s, Fick’s, or Fourier’s law depending on the physical context—is

$$X = -K \text{grad} \, u + \mu V \quad (4.23)$$

where $K \geq 0$ and $\mu$ are functions of density, position, and time; and $V$ denotes the flow field for the medium in which our substance is moving. The minus sign on the gradient term represents the assumption that the substance diffuses from higher to lower concentrations.

By inserting the relation (4.23) into the balance law (4.22), we obtain the dynamical equation

$$u_t = \text{div}(K \text{grad} \, u) - \text{div}(\mu V) + f. \quad (4.24)$$

Also, if we assume that the diffusion coefficient $K$ is equal to $k^2$ for some constant $k$, the function $\mu$ is given by $\mu(u, x, t) = \gamma u$ where $\gamma$ is a constant,
and $V$ is an incompressible flow field ($\text{div} V = 0$); then we obtain the most often used reaction-diffusion-convection model equation

$$u_t + \gamma \text{grad} u \cdot V = k^2 \Delta u + f. \quad (4.25)$$

In this equation, the gradient term is called the convection term, the Laplacian term is called the diffusion term, and $f$ is the source term.

Let us also note that if the diffusion coefficient is zero, the convection coefficient is given by $\gamma = 1$, the source function vanishes, and $V$ is not necessarily incompressible, then the dynamical equation (4.24) reduces to the law of conservation of mass, also called the continuity equation, given by

$$u_t + \text{div}(uV) = 0. \quad (4.26)$$

Because equation (4.25) is derived from general physical principles, this PDE is used to model all physical processes where reaction, diffusion, or convection is involved.

In case the substrate medium is stationary (that is, $V = 0$), the model equation (4.25) is the diffusion (or heat) equation with a source

$$u_t = k^2 \Delta u + f. \quad (4.27)$$

Using Fourier’s law, this equation is the basic model equation for heat flow.

The derivation of the model equation (4.25) should help to explain the widespread appearance of the Laplacian in applied mathematics: It is the divergence of the gradient vector field that we suppose carries some substance from higher to lower concentrations.

**Exercise 4.19.** Show that the gradient of a function evaluated at a point $p$ points in the direction of maximum increase of the function at $p$. Also, the gradient is orthogonal to each level set of the function at each point on a level set.

**Exercise 4.20.** Discuss the meaning of the divergence of a vector field. In particular, discuss positive divergence, negative divergence, and zero divergence. Give examples. Hint: You may wish to consider the equation

$$\text{div} X(p) = \lim_{\Omega \to \{p\}} \frac{1}{\text{vol} \Omega} \int_{\partial \Omega} X \cdot \eta \, dS.$$
4.4.2 Reaction-Diffusion in One Spatial Dimension: Heat, Mutations, and Traveling Waves

One-dimensional Diffusion

Let us consider the diffusion of heat in a bar with insulated ends (that is, zero heat flux through each end of the bar) under the assumptions that the temperature in every cross-section perpendicular to the bar’s axis has uniform temperature and there are no heat sources or sinks along the bar. We are given the initial temperature distribution along the bar. The problem is to determine the temperature as time increases.

Under our assumptions, we need only consider the spatial distribution of heat along the axis of the bar that we idealize as an interval of real numbers. For a bar of length $L$, we thus let $x$ in the open interval $(0, L)$ denote the spatial coordinate. The basic model equation (4.27) with no heat sources or sinks in this case is

$$u_t = \kappa u_{xx}, \quad (4.28)$$

where $u(x, t)$ is the temperature at position $x$ along the bar at time $t \geq 0$ and $\kappa > 0$ is a constant (called the diffusivity) that depends on the material used to construct the bar. The value of $\kappa$ must be determined by experiment.

A function $f : (0, L) \to \mathbb{R}$ representing the initial temperature along the bar gives the initial condition

$$u(x, 0) = f(x). \quad (4.29)$$

Zero flux conditions at each end of the bar provide the boundary conditions

$$u_x(0, t) = 0, \quad u_x(L, t) = 0. \quad (4.30)$$

These boundary conditions are also called the zero Neumann boundary conditions.

The problem is to find a function $u$ that satisfies the partial differential equation (4.28), the initial condition (4.29), and the boundary conditions (4.30). This is a classic problem first solved by Joseph Fourier in 1822. His basic ideas, which have far reaching consequences, are milestones in the history of science. Fourier’s law of heat conduction—heat flows from regions of high temperature to regions of lower temperature—in the derivation of the PDE (4.28). His mathematical solution of our heat problem uses another important innovation called Fourier series, which will be introduced here.
A fundamental idea that often works for linear PDEs whose spatial domains are rectangular is separation of variables. We simply look for solutions in the form $u(x, t) = X(x)T(t)$, where $X$ and $T$ are unknown functions. Inserting this guess into our PDE yields the formula

$$X(x)T'(t) = \kappa X''(x)T(t),$$

which must hold if $X(x)T(t)$ is a solution. Let us assume for the moment that $X(x)$ and $T(t)$ do not vanish and rearrange the last formula to the form

$$\frac{T'(t)}{\kappa T(t)} = \frac{X''(x)}{X(x)}.$$

If $X(x)T(t)$ is a solution that does not vanish, the left-hand side of the equation a function of $t$ alone and the right-hand side a function of $x$ alone. This implies that the left-hand side and the right-hand side of the equation are equal to the same constant $c$. In other works, there must be a constant $c$ such that

$$T'(t) = c\kappa T(t), \quad X''(x) = cX(x);$$

that is, $X$ and $T$ must be solutions of the given ordinary differential equations.

The ODE $X''(x) = cX(x)$ is easily solved for each of the usual cases: $c > 0$, $c = 0$, and $c < 0$. Indeed, for $c = \lambda^2 > 0$, the general solution is

$$X(x) = ae^{\lambda x} + be^{-\lambda x},$$

for $c = 0$,

$$X(x) = ax + b;$$

and, for $c = -\lambda^2 < 0$,

$$X(x) = a \cos \lambda x + b \sin \lambda x.$$

Our solution must satisfy the boundary conditions. This is possible only for $c = 0$ and $X(x) = b$ or $c < 0$ and

$$X(x) = a \cos \frac{n\pi}{L} x,$$

where $n$ is an integer. Since cosine is an even function, all of the solutions for $c < 0$ are obtained with $n$ ranging over the non negative integers.
Conservation of Mass

For $c = -\lambda^2 = -\left(\frac{n\pi}{L}\right)^2$, the corresponding solution of $T'(t) = c\kappa T(t)$ is

$$T(t) = e^{-\left(\kappa n^2\pi^2/L^2\right)t}.$$  

Thus we have constructed an infinite number of solutions:

$$u_n(x, t) = X_n(x)T_n(t) = e^{-\left(\kappa n^2\pi^2/L^2\right)t} \cos \frac{n\pi}{L}x, \quad n = 0, 1, 2, \ldots, \infty.$$  

By a simple calculation, if $u$ and $v$ are solutions of our PDE and boundary conditions, then so is every linear combination (also called superposition) $au + bv$, where $a$ and $b$ are scalars. If follows that every finite sum

$$u(x, t) = b_0 + \sum_{n=1}^{N} b_n e^{-\left(\kappa n^2\pi^2/L^2\right)t} \cos \frac{n\pi}{L}x$$  

is a solution of the PDE and the boundary conditions.

What about the initial condition? One fact is clear: If

$$f(x) = b_0 + \sum_{n=1}^{N} b_n \cos \frac{n\pi}{L}x$$

for some choice of $b_0, b_1, b_2, \ldots, b_N$, then the function $u$ in display (4.31) is a solution of our PDE, boundary conditions, and initial conditions. This result suggests the question: Which functions $f$ can be written as a sum of cosines? The surprising answer, first given by Fourier, is that most functions defined on the interval $[0, L]$ can be written as an infinite sum of cosines (or sines). We will discuss this result in more detail in subsequent sections. A more precise statement is that every function $f$ defined on $[0, L]$ that is continuously differentiable, except possibly at a finite number of jump discontinuities, can be represented by a convergent Fourier cosine series; that is,

$$f(x) = b_0 + \sum_{n=1}^{\infty} b_n \cos \frac{n\pi}{L}x.$$  

This is a very powerful result. With it, we know that our model equation has a solution for all initial conditions we are likely to encounter. In fact, once we have expressed the initial condition $f$ as a Fourier series,

$$u(x, t) := b_0 + \sum_{n=1}^{\infty} b_n e^{-\left(\kappa n^2\pi^2/L^2\right)t} \cos \frac{n\pi}{L}x$$  

(4.33)
Conservation of Mass

is a solution of our PDE with its boundary and initial conditions.

It turns out that there is a simple method to determine the Fourier coefficients $b_n$ of a function $f$ defined on $[0, L]$. Fourier’s theorem tells us that our function $f$ may be expressed as in display (4.32) and that the sum on the right-hand side of the equation may be integrated term-by-term. Thus,

$$\int_0^L f(x) \, dx = \int_0^L b_0 \, dx + \sum_{n=1}^{\infty} \int_0^L b_n \cos \frac{n\pi}{L} x \, dx.$$

Every integral in the infinite summation vanishes. Hence,

$$b_0 = \frac{1}{L} \int_0^L f(x) \, dx.$$

For each positive integer $m$, we have that

$$\int_0^L f(x) \cos \frac{m\pi}{L} x \, dx = \int_0^L b_0 \cos \frac{m\pi}{L} x \, dx + \sum_{n=1}^{\infty} \int_0^L b_n \cos \frac{n\pi}{L} x \cos \frac{m\pi}{L} x \, dx.$$

As before, the first integral on the right-hand side of the equation vanishes. The integrals of products of cosines behave in the best possible way: they all vanish except for the product where $n = m$. (Check this statement carefully; it is a basic result that makes Fourier series useful.) Using this fact,

$$\int_0^L f(x) \cos \frac{m\pi}{L} x \, dx = \int_0^L b_m \cos^2 \frac{m\pi}{L} x \, dx = b_m \frac{L}{2},$$

and we have the general formula

$$b_m = \frac{2}{L} \int_0^L f(x) \cos \frac{m\pi}{L} x \, dx.$$

All the Fourier coefficients may be computed simply by integrating the product of the given function and an appropriate cosine. A similar result holds for Fourier sine series. Indeed, a function $f$ in the same class of functions may be represented as the Fourier sine series

$$f(x) = \sum_{n=1}^{\infty} a_n \sin \frac{n\pi}{L} x.$$
We have constructed a solution of our initial boundary value problem. But, it would be useless for applied mathematics if there were more than one solution. Which solution would we choose? Fortunately, it is easy to show that our solution is unique. Suppose on the contrary that there were two solutions $u$ and $v$. By superposition, $w := u - v$ is also a solution of the same boundary value problem but with zero initial condition; that is,

$$w_t = \kappa w_{xx}, \quad w(x,0) = 0, \quad w_x(0,t) = 0, \quad w_x(L,t) = 0.$$  

Note that

$$\int_0^L w_t w \, dx = \int_0^L \kappa w_{xx} w \, dx.$$  

By an application of integration by parts on the right-hand integral and taking the time-derivative outside the left-hand integral, we have the equality

$$\frac{d}{dt} \int_0^L \frac{1}{2} w^2 \, dx = \kappa w_x \bigg|_0^L - \int_0^L w_x^2 \, dx.$$  

Using the boundary conditions,

$$\frac{d}{dt} \int_0^L \frac{1}{2} w^2 \, dx = -\kappa \int_0^L w_x^2 \, dx.$$  

It follows that

$$\int_0^L \frac{1}{2} w^2 \, dx$$

is a non negative function that does not increase as $t$ increases. But, at $t = 0$, the initial condition is $w = 0$. Hence, $w$ must be the zero function for all $t > 0$ for which the solution exists; and, $u = v$, as desired.

The analytic solution of our diffusion model equation solves can be used to answer questions and make predictions from the model. When we add heat sources and sinks to our model or make other modifications, especially those that produce nonlinear terms, our analytic solution method will likely fail. In these cases we may use approximate solutions obtained with numerical computations. Our diffusion model serves as a simple example to illustrate a method for approximating the solutions of partial differential equations.

To make a numerical computation, we must be given the diffusivity $k$, the length of the spatial domain $L$ and the initial data $f$. The first step is to discretize space and time. This may be done by choosing a positive integers
Conservation of Mass

\[ N \text{ with corresponding spatial increment } \Delta x := L/N, \text{ a temporal increment } \Delta t, \text{ and agreeing to consider the value of } u \text{ only at the interior spatial domain points with coordinates } x_i := i\Delta x = iL/N \text{ for } i = 1, 2 \ldots N - 1 \text{ and temporal coordinates } j\Delta t \text{ for } j = 0, 1, 2, \ldots, \text{ where we leave open the number of time-steps we plan to compute. In other words, we will approximate the values of } u \text{ at the interior grid-points } U_i^j = u(i\Delta x, j\Delta t).

Having discretized space and time, we will also discretize the PDE by approximating the time and space derivatives. For this, we use Taylor theorem. Recall that for a function \( f \) that is \( (n + 1) \)-times continuously differentiable at a point \( a \), the value of \( f \) at a nearby point \( x \) is given by

\[
f(x) = f(a) + f'(a)(x - a) + \frac{1}{2!}f''(a)(x - a)^2 + \frac{1}{3!}f'''(a)(x - a)^3 + \cdots + \frac{1}{n!}f^n(a)(x - a)^n + \frac{1}{(n + 1)!}f^{n+1}(c)(x - a)^{n+1}
\]

for some number \( c \) that lies between \( a \) and \( x \). We often say that the right-hand side of the last formula is the taylor expansion of the function \( f \) at \( a \). Let us consider a point \((x, t)\) and the nearby point \((x, t + \Delta t)\). By an application of Taylor's formula to the function \( \Delta t \mapsto u(x, t + \Delta t) \) at \( \Delta t = 0 \), we have the expansion

\[
u(x, t + \Delta t) = u(x, t) + u_t(x, t)\Delta t + \frac{1}{2!}u_{tt}(x, c)\Delta t^2.
\]

A rearrangement yields the equation

\[
u_t(x, t) = \frac{u(x, t + \Delta t) - u(x, t)}{\Delta t} - \frac{1}{2!}u_{tt}(x, c)\Delta t.
\]

and the approximation

\[
u_t(x, t) \approx \frac{u(x, t + \Delta t) - u(x, t)}{\Delta t}
\]

with an error of order \( \Delta t \). Using our discretization, we may approximate the left-hand side of our PDE by

\[
u_t(i\Delta x, j\Delta t) \approx \frac{U_{i}^{j+1} - U_{i}^{j}}{\Delta t}.
\]

\[ (4.35) \]
By a similar procedure, note that up to an error of fourth-order in $\Delta x$,

$$u(x - \Delta x, t) = u(x, t) - u_x(x, t)\Delta x + \frac{1}{2}u_{xx}(x, t)\Delta x^2 - \frac{1}{3!}u_{xxx}(x, t)\Delta x^3,$$

$$u(x + \Delta x, t) = u(x, t) + u_x(x, t)\Delta x + \frac{1}{2}u_{xx}(x, t)\Delta x^2 + \frac{1}{3!}u_{xxx}(x, t)\Delta x^3.$$

These equations may be added and rearranged to obtain the approximation

$$u_{xx}(x, t) \approx \frac{u(x - \Delta x, t) - 2u(x, t) + u(x + \Delta x, t)}{\Delta x^2}$$

with an error of order $\Delta x^2$, which is better than our approximation of the time derivative. Using our discretization, we may approximate the right-hand side of our PDE by

$$u_{xx}(i\Delta x, j\Delta t) \approx \frac{U_j^i - 2U_i^j + U_i^{j+1}}{\Delta x^2}. \quad (4.36)$$

By equating the approximations (4.35) and (4.36) and rearranging the result, we have the discrete approximation of our PDE

$$U_i^{j+1} = U_i^j + \frac{\kappa\Delta t}{\Delta x^2}(U_i^{j-1} - 2U_i^j + U_i^{j+1}). \quad (4.37)$$

The left-hand side of equation (4.37) is the value of $u$ at the $(j + 1)$st time step; all values of $u$ on the right-hand side are evaluated at the $j$th time step. Since we are given the values of $u$ at the zeroth time step via the initial condition, we may simply bootstrap up to any desired time by iteration. Well, almost.... We have not yet taken into account the boundary conditions. For example, we are given the values of $U_0^j$ for $i = 1, 2, 3 \ldots \bar{N} - 1$. To obtain the values at $U_0^1$ we need $U_0^0$ at the left end and $U_{\bar{N}}^0$ at the right end of the bar. As a simple approximation of the Neumann boundary conditions (which require $u_x$ to vanish at each end), we may insist that for all time steps

$$U_0^j = U_1^j, \quad U_\bar{N}^j = U_{\bar{N}-1}^j. \quad (4.38)$$

If $u$ is constant over the intervals $[x_0, x_1]$ and $[x_{\bar{N}-1}, x_{\bar{N}}]$, then $u_x$ is zero at the ends of the bar, which have spatial coordinates $x_0$ and $x_{\bar{N}}$. This is of course not the only possible method to insure the boundary conditions are met; but, at least, it is simple to implement in a computer code.
All the ingredients are now in place to approximate the PDE (4.28) with Neumann boundary conditions (4.30) and initial condition (4.29) on a computer. Simply impose the end conditions (4.38) at each step, set \( j = 1 \) and compute \( U_1^j \) for \( i = 1, 2, 3, \ldots N - 1 \) using the equation (4.37), repeat the process to compute \( U_2^j \) using the previously computed values \( U_1^j \), and continue the same way in general to compute \( U_{j+1}^i \) over \( i = 1, 2, 3, \ldots N - 1 \) using the previously computed values \( U_j^i \). The process is stopped when the product of the number of steps and the increment \( \Delta t \) reaches the desired final time.

Our numerical scheme may also be viewed in vector form. Define the \( N - 1 \)-vector \( W^j \) to be the transpose of the row vector \((U_1^j, U_2^j, U_3^j, \ldots, U_{N-1}^j)\), let \( \alpha = \kappa \Delta t / \Delta x^2 \), and define the \((N - 1) \times (N - 1)\)-matrix \( A \) whose main diagonal is \((1 - \alpha, 1 - 2\alpha, 1 - 2\alpha, \ldots, 1 - 2\alpha, 1 - \alpha)\) (that is, the first and last components are \(1 - \alpha\) and the other components are all \(1 - 2\alpha\)), and whose first super-diagonal and first sub-diagonal elements are all \(\alpha\). In case \(N = 5\), the matrix is

\[
A = \begin{pmatrix}
1 - \alpha & \alpha & 0 & 0 \\
\alpha & 1 - 2\alpha & \alpha & 0 \\
0 & \alpha & 1 - 2\alpha & \alpha \\
0 & 0 & \alpha & 1 - \alpha \\
\end{pmatrix}.
\]

The iteration scheme (4.37) may be written in the compact form

\[
W^{j+1} = AW^j.
\]

In other words, the iteration scheme is simply matrix multiplication by \( A \). The special first and last rows of \( A \) are due to the Neumann boundary conditions.

The initial vector \( W^0 \) is given by the initial condition for the PDE. The numerical scheme is implemented by iteration: \( W^1 = AW^0 \), \( W^2 = AW^1 \), and so on; or, in a more compact form

\[
W^{j+1} = A^j W^j
\]

where \( A^j \) means the \( j \)th power of the matrix \( A \) and \( W^j \) means the \( j \)th element in the sequence whose first three elements are \( W^1, W^2, \) and \( W^3 \).

Our numerical scheme has at least two potential difficulties: The approximation of the time derivative \( u_t \) by our discretization is less accurate than our approximation of the spatial derivative \( u_{xx} \), and numerical instabilities
Conservation of Mass

will occur unless the temporal and spatial discretizations are chosen appropriately. Alternative numerical methods that overcome the first difficulty will be discussed in Section 4.4.4.

The condition for numerical stability is

$$\kappa \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}.$$ (4.39)

This requirement is explained in Section 4.4.4. Numerical instabilities that lead to meaningless numerical results are likely to occur if this condition, called the Courant-Friedricks-Lewy condition, is not met (see Exercise 4.25). In practice, the Courant-Friedricks-Lewy condition determines the maximum allowable time-step size for our numerical method after a spatial discretization is set.

To give some indication of how the Courant-Friedricks-Lewy condition arises, suppose that a roundoff error is introduced in the computer at the first computational step. Instead of computing $W^1$, the machine computes $W^1 + \epsilon$ (where $\epsilon$ is an $N - 1$-vector representing the error). Of course, further errors might be introduced at subsequent steps. But, for simplicity, consider the propagation of the first error only and assume that after the first error occurs the arithmetic is exact. Under these conditions, we will compute

$$W^2 = AW^1 + A\epsilon, \quad W^3 = A^2W^1 + A^2\epsilon, \quad W^4 = A^3W^1 + A^3\epsilon, \ldots.$$ 

The error propagates as $A^j\epsilon$. Our method will produce useless results if the norm of this propagated error grows as $j$ increases.

What would happen if the matrix $A$ were the diagonal $2 \times 2$-matrix

$$A = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}?$$

Clearly

$$A^j = \begin{pmatrix} a^j & 0 \\ 0 & b^j \end{pmatrix}.$$ 

If either $|a| > 1$ or $|b| > 1$ and there is a corresponding nonzero element of the vector $\epsilon$, then the size of the propagated error will grow. For example, if $|b| > 1$ and $\epsilon$ is the transpose of the vector $(0.01, 0.035)$ the number $b^j0.035$ will grow to infinity as $j$ goes to infinity. In this case, roundoff errors will be amplified under iteration and our numerical approximation of the PDE will
Conservation of Mass

become useless. On the other hand, if both $|a|$ or $|b|$ are less than or equal to one, then the propagated error will remain bounded as $j$ goes to infinity. The same is true for general diagonal matrices.

The matrix $A$ for our numerical scheme is not diagonal. But, it has a special form: $A$ is symmetric; that is, $A$ is equal to its transpose. Every symmetric matrix is diagonalizable; in other words, if $A$ is symmetric, then there is an invertible matrix $B$ such that $B^{-1}AB$ is diagonal. Also, every eigenvalue of a symmetric matrix is real. Using these facts, let us suppose that every eigenvalue of our matrix $A$ lies in the closed interval $[-1, 1]$. The matrix $C := B^{-1}AB$ has the same eigenvalues as $A$. (Why?). Iterations of a vector $v$ by $C$ remain bounded because $C$ is diagonal and all its eigenvalues have absolute value less than or equal to one. In fact,

$$|C^j v| \leq |v|$$

for all vectors $v$ and all positive nonzero integers $j$. Iteration of $v$ by the matrix $A$ also remains bounded because

$$|A^j v| = |(BCB^{-1})^j v| \leq |BCB^{-1}v| \leq \|B\|\|B^{-1}v\| \leq \|B\|\|B^{-1}\||v|.$$ 

In fact, if the Courant-Friedricks-Lewy condition (4.39) is satisfied, all eigenvalues of $A$ are in the closed interval $[-1, 1]$ (see Section 4.4.4).

**Exercise 4.21.** (a) Solve the diffusion equation on the spatial domain $[0, L]$ with initial condition and zero Dirichlet boundary conditions:

$$u_t = \kappa u_{xx}, \quad u(x, 0) = f(x), \quad u(0, t) = 0, \quad u(L, t) = 0.$$

(b) Show that your solution is unique.

**Exercise 4.22.** Solve the diffusion equation on the spatial domain $[0, L]$ with initial condition and non zero Neumann boundary conditions:

$$u_x(0, t) = a, \quad u_x(L, t) = b,$$

where $a$ and $b$ are real numbers. Hint: Look for a solution $u = v + w$, where $v$ is a function that satisfies the boundary conditions and $w$ satisfies the PDE with zero boundary conditions.

**Exercise 4.23.** Solve the diffusion equation on the spatial domain $[0, L]$ with initial condition and non zero Dirichlet boundary conditions:

$$u(0, t) = a, \quad u(L, t) = b,$$

where $a$ and $b$ are real numbers.
Conservation of Mass

Exercise 4.24. The zero flux boundary condition for the diffusion equation has the physical interpretation that no substance is lost as time increases. Prove this fact by showing that the time-derivative of the total amount of substance (its integral over the spatial domain) vanishes.

Exercise 4.25. (a) Write a computer code to implement the method described in this section to approximate the solution of the diffusion equation in one space-dimension on a finite interval with Neumann boundary conditions and given initial conditions. As a test case, consider the spatial domain to be one unit in length, the diffusivity $\kappa = 1$, and the initial data given by $f(x) = 1 + \cos\pi x$. Compare your numerical results with the analytic solution. (b) Test the Courant-Friedricks-Lewy number by choosing at least the discretizations $\Delta t/\Delta x^2 = 0.4$ and $\Delta t/\Delta x^2 = 0.6$. Discuss your results.

Exercise 4.26. The following is a constitutive model for the ink diffusion experiment in Section 3.3: Let $u(x,t)$ denote the concentration of ink at position $x$ at time $t$ and $L$ the trough length. Also let $a$ and $b$ be positive constants. The model equation of motion for $u$ is the initial boundary value problem

$$u_t = ku_{xx}, \quad u_x(0, t) = u_x(L, t) = 0, \quad u(x, 0) = \begin{cases} a, & 0 \leq x \leq b, \\ 0, & x > b. \end{cases}$$

(a) What constitutive law is used to construct this model? (b) Show that the model predicts the presence of ink at every position along the trough for every positive time. Does this fact invalidate the model? (c) Show that the model predicts that the total amount of ink remains constant in time. (d) Define the diffusion front to be the largest distance from the origin where the ink concentration is 1% of $a$. Use the model to determine the diffusion front. If necessary, choose values for the parameters in the problem. (e) Is the distance of the diffusion front from the origin a linear function of time? If not, what type is this function? (f) Calibrate the model to the data given in Section 3.3 and discuss the model prediction in view of the experimental data. (g) Construct a model that takes into account two (or three) space dimensions. Compare the new front speed with the front speed obtained for the one-dimensional model. (h) Can you refine the model to give a more accurate representation of the diffusion experiment? You may wish to perform a similar diffusion experiment.

Propagation of a Mutant Gene

Consider a population with a mutant gene whose concentration is $u$. Its allele (the parent gene) has concentration $1 - u$. The individuals in the population...
Conservation of Mass

diffuse along a one-dimensional spatial domain (for example a shoreline) and they interact with each other to produce offspring. A simple model (introduced in 1937 independently by R. A. Fisher [26] and A. N. Kolmogorov, I. Petrovskii and N. Piscounov [39]) is

$$u_t = \kappa u_{xx} + au(1 - u).$$ \hspace{1cm} (4.40)

The choice of the model interaction term, given by \(f(u) = au(1-u)\), may be more fully justified using probability theory (see [39]). This term is perhaps the simplest model of interaction (a product) that agrees (qualitatively) with experiments. The constant \(a\) is meant to model the utility of the gene for the organism to survive: \(a > 0\) for an advantageous mutation; \(a < 0\) for a disadvantageous mutation.

There is a natural scientific question: How will an advantageous mutation spread if it occurs in some individual or group of individuals at a specified spatial location?

The quantity \(u(x,t)\) is the number of individuals at time \(t\) at shoreline position \(x\) with the mutant gene divided by the total number of individuals in the population at \(x\). Of course this is an idealization. In reality, we would measure this ratio over some area (a fixed width times a length of shoreline). The concentration is more precisely defined to be the limit of this ratio as the area of the region shrinks to zero at \(x\) for the fixed time \(t\). Thus, \(u\)—the ratio of two tallies—is a dimensionless quantity. The time-derivative \(u_t\) has the dimensions of inverse time. In symbols, we write \([u_t] = 1/T\), where in this formula the square brackets denote units of the enclosed expression and \(T\) denotes the unit of time (perhaps \(T\) is years). The diffusion term has units \([\kappa u_{xx}] = [\kappa]/L^2\): These must agree with the units of the time derivative. Thus, \([\kappa]/L^2 = 1/T\) and \([\kappa] = L^2/T\). Likewise, the units of the interaction term are carried by \(a\) and \([a] = 1/T\); thus, \(a\) has the appropriate units for a rate.

We are unlikely to know good values of the diffusivity \(\kappa\) (a measure of how fast the organisms carrying the mutant gene spread along the shoreline) or \(a\) (which is the growth rate of the population with the mutant gene). Thus, we should not expect to make reliable quantitative predictions; rather, we should use the model to predict the qualitative behavior of the spread of a mutant gene.

The qualitative behavior of the solutions of our model are independent of the (positive) parameters \(\kappa\) and \(a\). In fact, we can simply eliminate the
parameters by a change of variables. Let $t = \tau s$, where $[\tau] = T$ (that is, $\tau$ has the dimensions of time); and, $x = \ell \xi$, where $[\ell] = L$. The change of variables is accomplished by applying the chain rule. Note first that

$$\frac{du}{dt} = \frac{du}{ds} \frac{ds}{dt} = \frac{1}{\tau} u_s,$$

$$\frac{du}{dx} = \frac{du}{d\xi} \frac{d\xi}{dx} = \frac{1}{\ell} u_\xi,$$

$$\frac{d^2 u}{dx^2} = \frac{1}{\ell^2} u_{\xi\xi}.$$

By substitution into the model PDE (4.40), we have

$$\frac{1}{\tau} u_s = \kappa \frac{1}{\ell^2} u_{\xi\xi} + au(1 - u)$$

and, by rearranging, the equation

$$u_s = \kappa \frac{\tau}{\ell^2} u_{\xi\xi} + \tau au(1 - u).$$

We may choose

$$\tau = \frac{1}{a}, \quad \ell = \sqrt{\frac{\kappa}{a}}.$$  \hspace{1cm} (4.41)
to obtain the desired dimensionless model equation

$$u_s = u_{\xi\xi} + u(1 - u).$$

Reverting to the usual notation, we will discuss the PDE

$$u_t = u_{xx} + u(1 - u). \quad (4.42)$$

For our biological problem, we may consider the PDE for $x$ on the whole real line; or, we may impose the Dirichlet condition ($u$ vanishes at the ends of the portion of the shoreline under consideration) or the Neumann condition ($u_x$ vanishes at these points). Both make sense in this biological context.

The model (4.42) is a non-linear PDE; the superposition of solutions is generally not a solution. There is no known general explicit solution (cf. Exercise 4.27). But, it is possible to prove that unique solutions exist for appropriate initial conditions $u(x, 0) = f(x)$ (see, for example, [60]). Let us assume these results.

We may gain some insight into the behavior predicted from this model by assuming there is no diffusion. In this case, we simply have the ODE $\dot{u} = u(1 - u)$ (see Exercise 2.4). There are two steady states $u = 0$ and $u = 1$. Also, if $0 < u(0) < 1$, then the corresponding solution $u(t)$ grows monotonically to $u = 1$ as $t$ goes to infinity. Thus, with the passage of time, the mutant gene eventually is established in the entire population at each location along the shoreline. Is this basic conclusion altered when the model takes diffusion (and thus spatial dependence) into account?

The next obvious step in our analysis is numerical experimentation. Let us suppose that the initial population with the mutant gene is localized to one place along the shoreline. This situation may be modeled by an initial $u$ of the form

$$u(x, 0) = f(x) = \begin{cases} 
0; & x < a, \\
\mu; & a \leq x \leq b, \\
0; & x > b,
\end{cases} \quad (4.43)$$

where $a < b$ and $0 < \mu \leq 1$.

Some of the results of a typical numerical experiment are depicted in Figure 4.6. The spread of the mutant gene is a wave spreading in both directions from the spatial location of the initial population that carried the mutation. The wave speed for this simulation is approximately 10.

To help determine the wave speed in general, let us note that in our scaling (4.41), the characteristic velocity is length divided by time. Our
Conservation of Mass

Figure 4.7: Graphs of the spatial distribution of the mutant gene modeled by PDE (4.42) at times $t = 20, 40, 60$ and $80$ are depicted for the condition (4.43) with $a = -100$, $b = -90$, and $\mu = 1$. The plotted curves are thicker as time increases. The numerical method is forward Euler on the spatial domain $[-100, 100]$ with Neumann boundary conditions, time step 0.1, and a spatial grid of 200 interior points.

The length scale is $\ell = \sqrt{\kappa/a}$ and our time scale is $\tau = 1/a$. There is a unique characteristic velocity given by

\[
\text{characteristic velocity} = \sqrt{\kappa a}. \quad (4.44)
\]

The wave speed should be a function of this characteristic velocity. Thus the dependence of the wave speed of a wave solution of the original model (4.42) should be $h(\sqrt{\kappa a})$ for some scalar function $h$. This argument is a simple example of an application of dimensional analysis, which is often useful to help determine the functional dependence of some phenomenon on the parameters in a model.

The basic prediction obtained from our PDE model for the spread of a mutant gene from a location where the mutation arises is the spread of this gene as a wave moving in both directions away from the primary location with the concentration of the mutant gene approaching one as the wave passes each remote location.

A related scenario is the spread of a mutant gene that is already dominant on one side of a location along our beach but not present on the other side. Some of the results of a numerical experiment are reported in Figure (4.7). These results suggest the local concentration of the mutant gene rises to dominate the population where the mutation is already present and it spreads
Figure 4.8: A portion of the phase portrait for the ODE system (4.49) with $c = 3$ is depicted. The flow crosses each thick line segment in the same direction. The thick curve is an approximate trajectory connecting the rest points at $(0, 0)$ and $(1, 0)$.

to the right at a constant speed with each location being fully saturated as the wave front passes. A mathematical idealization of this situation leads to the question: Is there a solution $u$ of the dimensionless model equation (4.42) such that

$$0 \leq u(x, t) \leq 1, \quad \lim_{t \to -\infty} u(x, t) = 0, \quad \lim_{t \to \infty} u(x, t) = 1? \quad (4.45)$$

The existence of a solution of this type would show that some solutions of the PDE have the same qualitative behavior as a solution of $\dot{u} = u(1 - u)$ (that is, the model differential equation without diffusion) with initial value $u(0)$ restricted to $0 < u(0) < 1$.

Inspection of the Figure (4.7) (or better yet an animation) suggests that the spatial concentration quickly approaches a wave that maintains its profile while moving to the right with constant velocity. Ideally, there is some function $\phi : \mathbb{R} \to \mathbb{R}$ such that

$$u(x, t) = \phi(x - ct), \quad (4.46)$$
where \( c > 0 \) is the wave speed. A solution of this form is called a traveling wave with wave form \( \phi \).

By substitution of the formula (4.46) into the PDE (4.42), we obtain the differential equation

\[
-c\phi'(x - ct) = \phi''(x - ct) + \phi(x - ct)(1 - \phi(x - ct)),
\]

which we may view as an ordinary differential equation for the unknown wave profile \( \phi \) with auxiliary conditions

\[
0 \leq \phi \leq 1, \quad \lim_{s \to -\infty} \phi(s) = 1, \quad \lim_{s \to \infty} \phi(s) = 0.
\]

A basic fact is that \( c \geq 2 \), then there is a solution of the ODE

\[
\ddot{\phi} + c\dot{\phi} + \phi(1 - \phi) = 0
\]

(4.48)

that satisfies the auxiliary conditions.

The second-order ODE (4.48) is equivalent to the system of ODEs

\[
\dot{\phi} = v, \\
\dot{v} = -cv - \phi(1 - \phi)
\]

(4.49)

in the phase plane.

Our system (4.49) has two rest points in the phase plane \((0, 0)\) and \((1, 0)\). The rest point at the origin is asymptotically stable and the rest point at \((1, 0)\) is a saddle point. If \( c \geq 2 \), then the triangle depicted by thick line segments is positively invariant. The horizontal segment connects the rest points, the ray with negative slope is in the direction of the eigenspace corresponding to the negative eigenvalue of the linearized system matrix at \((0, 0)\) with the largest absolute value, and the ray with positive slope is in the direction of the eigenspace corresponding to the unstable manifold of the linearized system matrix at \((1, 0)\). For \( c < 2 \), the eigenvalues of the linearized system at the origin are complex. Thus, the incoming unstable trajectory from the saddle point winds around the origin and \( \phi \) has negative values, violating the condition that \( 0 \leq \phi \leq 1 \) (see [15] for a detailed proof and Exercise 4.34).

The connecting orbit in Figure (4.8) corresponds to the desired function \( \phi \) that defines a traveling wave solution. Of course, \( \phi \) is the first coordinate function of the corresponding solution of the ODE system (4.49). Its graph, which has the profile of the expected traveling wave, is depicted in
Figure 4.9: The figure depicts an approximation of the graph of the function $\phi$ that is the first coordinate function of the solution of the ODE system (4.49) with $c = 3$ along the unstable manifold of the saddle point at $(1, 0)$ that connects this point to the sink at the origin.

Figure 4.9. Kolmogorov, Petrovskii and Piscounov [39] proved that solutions of the PDE (4.42) with initial data such that $0 \leq u(x, 0) \leq 1$, $u(x, 0) = 1$ for $x \leq a$ and $u(x, 0) = 0$ for $x \geq b > a$ approach a traveling wave solution with wave speed $c = 2$ (see Exercise 4.33).

**Exercise 4.27.** Show that $u(x, t) = \phi(x - ct)$ is a solution of the PDE (4.42) in case $c = 5/\sqrt{6}$, $K$ is a constant, and

$$\phi(z) = (1 + Ke^{z/\sqrt{6}})^{-2}.$$  

**Exercise 4.28.** (a) Repeat the experiment reported in Figure 4.6. (b) Repeat the experiment with Dirichlet boundary conditions and compare the results.

**Exercise 4.29.** How does the speed of the wave front(s) of solutions of the PDE (4.42) depend on the amplitude of the initial population with the mutant gene?

**Exercise 4.30.** How does the speed of the wave front(s) of solutions of the PDE (4.42) depend on the length of the spatial interval occupied by the initial population with the mutant gene?

**Exercise 4.31.** Use numerical experiments to test the characteristic velocity approximation in display (4.44). Set an initial mutant gene concentration and vary the parameters $\kappa$ and $a$ in the PDE (4.40).
Exercise 4.32. Reproduce Figures 4.7–4.9.

Exercise 4.33. Use numerical experiments to verify the theorem of Kolmogorov, Petrovksii and Piscounov [39] that solutions of the PDE (4.42) with initial data such that $0 \leq u(x,0) \leq 1$, $u(x,0) = 1$ for $x \leq a$ and $u(x,0) = 0$ for $x \geq b > a$ approach a traveling wave solution with wave speed $c = 2$.

Exercise 4.34. (a) Find the system matrix at each rest point of system 4.49, find the eigenvalues and eigenvectors, and determine the stability types of the rest points. Show the triangle as in Figure 4.8 is positively invariant by proving the vector field corresponding to the system of differential equations points into the region bounded by the triangle along the boundary of the region.

4.4.3 Reaction-Diffusion Systems: The Gray-Scott Model and Pattern Formation

Let us consider two concentrations $u$ and $v$ in some process involving diffusion of each concentration and some interaction between the two substances. We can imagine interacting populations (for example a predator and its prey) or a chemical reaction.

Absent diffusion, the interaction of two species is often modeled by a (nonlinear) system of ordinary differential equations

$$\dot{u} = f(u,v), \quad \dot{v} = g(u,v). \quad (4.50)$$

For example, the basic interaction between a predator concentration $u$ and its prey concentration $v$ might be modeled by

$$\dot{u} = -au + buv, \quad \dot{v} = cv(1 - \frac{1}{k}v) - buv. \quad (4.51)$$

Here $a$ is the death rate of the predator, $c$ is the growth rate of the prey, $k$ is the carrying capacity of the prey’s environment, and $b$ is the success rate of the predator. By taking into account the (spatial) diffusion of the two species, we obtain the reaction-diffusion model

$$u_t = \lambda \Delta u - au + buv, \quad v_t = \mu \Delta v + cv(1 - \frac{1}{k}v) - buv, \quad (4.52)$$

where now $u$ and $v$ are functions of space and time.
At a practical level, the derivation of phenomenological reaction-diffusion models of the form

\[ u_t = \lambda \Delta u + f(u, v), \quad v_t = \mu \Delta v + g(u, v) \] \hspace{1cm} (4.53)

can be as simple as our derivation of a predator-prey model. The derivation of more accurate models of course requires a detailed understanding of the underlying reaction [53, 56].

A famous and important paper [72] by Alan Turing suggests reaction-diffusion models to explain pattern formation in biological systems (morphogenesis). One of his models is given by

\[ u_t = \lambda \Delta u + r(\alpha - uv), \quad v_t = \mu \Delta v + r(uv - v - \beta). \] \hspace{1cm} (4.54)

The patterns we see in nature (for example, animal skin spots and stripes, fish skins, snow flakes, etc) can all be generated by reaction diffusion equations. This fact suggests a broad form of Turing’s principle: Reaction and diffusion are the underlying mechanisms for pattern formation in the natural world. While Turing’s Principle is controversial, the application of mathematics to understand reaction and diffusion is fruitful.

A widely studied reaction-diffusion model for pattern formation is the dimensionless Gray-Scott model [31, 55, 78, 52, 23]

\[ u_t = \lambda \Delta u + F(1 - u) - uv^2, \quad v_t = \mu \Delta v + uv^2 - (F + \kappa)v. \] \hspace{1cm} (4.55)

It is derived from a hypothetical chemical reaction of the form

\[ u + 2v \rightarrow 3v, \quad v \rightarrow P, \]

where the second reaction creates an inert product; the coefficient \( F \) is the dimensionless feed/drain rate (which feeds \( u \) and drains \( u, v, \) and \( P \)), and \( \kappa \) is the dimensionless rate of conversion in the second reaction \( v \rightarrow P \). A standard theory in chemical kinetics, called the law of mass action, assumes that the rate of a reaction is proportional to the product of the concentrations of the chemicals involved in the reaction. For a reaction of the form \( nA + mB \rightarrow C \) the reaction rate is \( r = k[A]^n[B]^m \), where the square brackets denote concentration, \( k \) is the constant of proportionality, and \( n \) and \( m \) are the stoichiometric coefficients that specify the number of molecules of the
Conservation of Mass

The corresponding chemical species that combine in the reaction. Thus, according to the first reaction equation \( u + 2v \rightarrow 3v \), the rate at which \( v \) is increased by the reaction is proportional to \( uvv \)—one molecule of \( u \) plus two molecules of \( v \) combine to form three molecules of \( v \).

For a system of reaction-diffusion equations

\[
U_t = \lambda \Delta U + H(U),
\]

Turing’s fundamental idea—which has had a profound influence on developmental biology—is that spatial patterns can form even for the case of small diffusion for a reaction ODE

\[
\dot{U} = H(U)
\]

that has an attracting steady state. In effect, the diffusion can act against the tendency of the process to proceed to the steady state of the reaction. The implication is that a pattern (for example, the spots and strips of animal skins) might arise from a chemical process involving reaction and diffusion.

How might we investigate a mathematical model for pattern formation?

To obtain a unique solution of our partial differential equation (distributed parameter system), we must specify initial and boundary conditions.

Let us consider a two-dimensional spatial domain \( \Omega \). Our concentration vector \( U = (u, v) \) is a function \( U : \Omega \times [0, T] \rightarrow \mathbb{R}^2 \). The initial condition simply gives the initial spatial distribution of the concentration vector; that is,

\[
U(x, 0) = U_0(x) \quad (4.56)
\]

for some function \( U_0 : \Omega \rightarrow \mathbb{R}^2 \).

The boundary conditions depend on the underlying physical problem. A Dirichlet boundary condition is used for the case where the concentration vector is a known (vector) function \( g : \partial \Omega \times [0, T] \rightarrow \mathbb{R}^2 \); that is,

\[
U(x, t) = g(x, t) \quad (4.57)
\]

whenever \( x \in \partial \Omega \) and \( t \in [0, T] \). A Neumann boundary condition is used for the case where the reactants do penetrate the boundary at some specified rate. In this case, we consider the outer unit-length normal \( \eta \) on \( \partial \Omega \). The Neumann boundary condition states that the normal derivative of the concentration vector is a known function on the boundary; that is,

\[
\eta \cdot \nabla U(x, t) = g(x, t) \quad (4.58)
\]
Conservation of Mass

whenever \( x \in \partial \Omega \) and \( t \geq 0 \).

A computationally convenient (but perhaps less physically realistic) boundary condition is the periodic boundary condition

\[
U(x + a, y) = U(x, y), \quad U(x, y + b) = U(x, y),
\]

where \( a \) and \( b \) are fixed positive constants. In other words, the concentration vector is supposed to be defined on a torus given by a rectangle with opposite sides identified with the same orientations.

A reaction-diffusion system defined on a bounded domain \( \Omega \) with a piecewise smooth boundary, either Dirichlet, Neumann, or periodic boundary conditions, and an initial condition, has a unique solution that exists for some finite time-interval \( 0 \leq t < \tau \) (see [56, 63]). This is our hunting license. In the next section we will consider some qualitative and numerical methods that can be used to obtain some understanding of the evolution predicted by the Gray-Scott model.

4.4.4 Analysis of Reaction-Diffusion Models: Qualitative and Numerical Methods

The mechanism for pattern formation is not well understood. At least it is very difficult to define the properties of the patterns that arise and prove their existence. To show that patterns arise begins with numerical approximations. But, as we will see, qualitative methods for understanding the solutions of differential equations arise naturally in this investigation.

Numerical methods and numerical analysis for differential equations are vast subjects, which are essential tools for applied mathematics. This section is meant as a glimpse into a few numerical methods together with some discussion of the issues that are encountered in practice. Our context will be an analysis of the the Gray-Scott model (4.55). What is the fate of the concentrations of the reactants of this hypothetical chemical reaction?

Euler’s Method

To begin our adventure, let us recall Euler’s method. It is the prototypical method for solving ODEs, for example the system of ODEs

\[
\dot{u} = f(u, v), \quad \dot{v} = g(u, v)
\]

(4.59)
obtained by ignoring the spatial dependence in our reaction-diffusion system (4.53). The idea is to discretize the time derivatives and thus approximate the continuous flow of time by a series of discrete steps that can be computed. To do this and for future reference, note that for a sufficiently smooth function of one variable $F : \mathbb{R} \to \mathbb{R}$ we have the Taylor series expansions

\[
F(x + h) = F(x) + F'(x)h + \frac{1}{2!}F''(x)h^2 + \frac{1}{3!}F'''(x)h^3 + O(h^4),
\]

\[
F(x - h) = F(x) - F'(x)h + \frac{1}{2!}F''(x)h^2 - \frac{1}{3!}F'''(x)h^3 + O(h^4),
\]

where the big $O$ notation is explained in Appendix A.8 and Taylor’s formula in Appendix A.9. Using the first equation of the display, it follows that

\[
F'(x) = \frac{F(x + h) - F(x)}{h} + O(h).
\]

Thus, by replacing $F'(x)$ with the (forward) difference quotient $F(x + h) - F(x)/h$, we make an error of order $h$. The numerical method is to make this replacement; the numerical analysis is to recognize the error has order $h$.

This replacement leads to Euler’s method for approximating solutions of the system of ODEs (4.59): For an initial time $t_0$ and initial state $(u,v) = (u_0, v_0)$, future times at the regular time-increment $\Delta t$ and the corresponding future states are given recursively by

\[
\begin{align*}
  u_n &= u_{n-1} + \Delta tf(u_{n-1}, v_{n-1}), \\
  v_n &= v_{n-1} + \Delta tg(u_{n-1}, v_{n-1}), \\
  t_n &= t_{n-1} + \Delta t.
\end{align*}
\]

(4.61) (4.62) (4.63)

The discretization error is $O(h)$; it is proportional to the step size $h$. The local truncation error, or the error per step, is the norm of the difference between the solution and the approximation after one step. In vector form ($\dot{U} = H(U)$) using Taylor’s formula, this error is

\[
|U(t_0 + \Delta t) - U_1| = |U(t_0) + \Delta tU'(t_0) + \frac{\Delta t}{2!}U''(\tau) - (U_0 + \Delta tH(U_0))|
\]

\[
= \frac{\Delta t}{2!}U''(\tau)
\]

\[
= O(\Delta t^2),
\]

where $\tau$ is a number between $t_0$ and $t_0 + \Delta t$. 
The Reaction Equations

We can apply Euler’s method to the Gray-Scott reaction model (4.55); that is, the system of ODEs

\[ \dot{u} = F(1 - u) - uv^2, \]
\[ \dot{v} = uv^2 - (F + \kappa)v. \]

(4.64)

To compute an approximate solution, we must set values for the parameters \( \kappa > 0 \) and \( F > 0 \), an initial condition \((u(0), v(0))\), and a time step \( \Delta t \). How should we make these choices? This question does not have a simple answer. The choices depend on the application. For the Gray-Scott model, we are interested in the fate of a typical solution. Thus, the choice of initial condition is not essential as long as it is generic. On the other hand, the reaction model has a two-dimensional parameter space. We will be computing for a very long time if we wish to exhaust all the possible parameter values. A wiser course of action is to rely on a fundamental principle of applied mathematics: Think before you compute.

Our current goal is to understand the general behavior of the system (4.64). Does this behavior depend on the parameter values?

The state variables \( u \) and \( v \) are supposed to represent chemical concentrations; thus, their values should be nonnegative. Moreover, the evolution of the system from a physically realistic state \((u(0) \geq 0 \text{ and } v(0) \geq 0)\) should produce only physically realistic states. In other words, the closed first quadrant in the state space should remain invariant under the flow. This fact is easily checked. Simply note that the \( u \)-axis is invariant (because \( \dot{v} = 0 \) whenever \( v = 0 \)), and the vector field points into the first quadrant along the positive \( v \)-axis (because \( \dot{u} > 0 \) whenever \( u = 0 \)).

The simplest dynamics is given by rest points; that is, constant solutions of the system of ODEs given by solutions of the algebraic equations

\[ F(1 - u) - uv^2 = 0, \quad uv^2 - (F + \kappa)v = 0. \]

Note that \((u, v) = (1, 0)\) is a solution for all values of \( \kappa \) and \( F \). Also, \((1/2, F/(2(F + \kappa)))\) is a double root whenever \( F = 4(F + \kappa)^2 \) and \( F > 0 \). Figure 4.10 depicts this curve (a parabola) in the parameter space. It meets the vertical axis at \((\kappa, F) = (0, 0)\) and \((\kappa, F) = (0, 1/4)\); the point \((\kappa, F) = (1/16, 1/16)\) is the point on the curve with the largest first coordinate. Our ODE has three rest points for the parameter vector \((\kappa, F)\) in
Figure 4.10: Plot of the parabola $F = 4(F + \kappa)^2$ with horizontal axis $\kappa$ and vertical axis $F$. Outside the curve, the system (4.64) has one rest point at $(u, v) = (1, 0)$. Inside the curve the system has three rest points.

the region bounded by the curve and the vertical axis and one rest point outside of this region. For parameter values inside the region, the two new rest points have coordinates

$$u = \frac{1}{2} \left(1 \pm \sqrt{1 - 4(F + \kappa)^2/F}\right), \quad v = \frac{F}{F + \kappa}(1 - u). \quad (4.65)$$

The system matrix for the linearization of the system of ODEs at $(1, 0)$ is the diagonal matrix whose main diagonal has the components $-F$ and $-(F + \kappa)$ (see Appendix A.17 for a discussion of linearization and stability). These negative numbers are the eigenvalues of this matrix; hence, $(1, 0)$ is asymptotically stable independent of the parameter values.

The new rest points appear as the parameters cross the curve from outside to inside; on the curve a double root appears “out of the blue.” This is called a saddle-node bifurcation or a blue sky catastrophe. The prototype for this bifurcation (called a normal form) is given by

$$\dot{x} = b - x^2; \quad \dot{y} = \pm y,$$
Figure 4.11: The figure (with horizontal axis $\kappa$ and vertical axis $F$) depicts with a solid curve the parabola $F = 4(F + \kappa)^2$ (the saddle-node bifurcation curve) and the dashed curve $F = 1/2(\sqrt{\kappa} - 2\kappa - \sqrt{\kappa - 4\kappa^3/2})$ (the Hopf bifurcation curve). Outside the saddle-node curve, the system (4.64) has one stable rest point at $(u, v) = (1, 0)$. Inside this curve the system has three rest points. Above the Hopf bifurcation curve and inside the saddle-node curve there are three rest points, two sinks and one saddle. Below the Hopf curve and above the saddle-node curve there is one sink, one source and one saddle. For $F$ decreasing and $\kappa < 0.0325$ (approximately), the Hopf bifurcation is supercritical, for $\kappa > 0.0325$, the Hopf bifurcation is subcritical.

where $b$ is the bifurcation parameter. Note that for $b < 0$ (which corresponds in our case to being outside the parabola $F = 4(F + \kappa)^2$), there are no rest points. The system with $b = 0$ has a semi-stable rest point called a saddle-node. For $b > 0$, there are two rest points: one saddle and one sink in the $+$-sign case, and one saddle and one source in the $-$-sign case. This scenario is exactly what happens for system (4.64). (The rest point at $(1, 0)$ plays no role in this bifurcation.)

The system matrix at the rest points in display (4.65) is

$$A = \begin{pmatrix} -F - v^2 & -2uv \\ v^2 & 2uv - (F + \kappa) \end{pmatrix}. \quad (4.66)$$
At these rest points $uv = F + \kappa$; therefore, we have the formulas

\[
\text{Trace} (A) = \kappa - v^2, \quad \text{Determinant} (A) = (F + \kappa)(v^2 - F). \tag{4.67}
\]

The eigenvalues of $A$ are the roots of the characteristic equation

\[
\text{Determinant} (A - \lambda I) = 0,
\]

where $I$ denotes the $2 \times 2$ identity matrix. By an easy computation, the characteristic equation is seen to have the general form

\[
\lambda^2 - \text{Trace} (A)\lambda + \text{Determinant} (A) = 0
\]

and the roots

\[
\lambda = \frac{\text{Trace}(A) \pm \sqrt{\text{Trace}(A)^2 - 4\text{Determinant}(A)}}{2}.
\]

The stability types of rest points are determined from the signs of $\text{Trace}(A)$ and $\text{Determinant}(A)$. For example, if $\text{Trace}(A) < 0$ and $\text{Determinant}(A) > 0$, then the radicand is either complex or, in case it is real, less than $\text{Trace}(A) < 0$. Hence, both roots have positive real parts and the corresponding rest point is a source.

Consider $\text{Trace}(A)$ on the parabola $F = 4(F + \kappa)^2$. Using the formulas in display (4.65) for the corresponding rest points, it follows that

\[
v = \frac{F}{2(F + \kappa)};
\]

hence,

\[
\text{Trace}(A) = \kappa - v^2
= \kappa - F \left( \frac{F}{4(F + \kappa)^2} \right)
= \kappa - F
= \frac{\sqrt{F}}{2} - 2F.
\]

By simple analysis of the function $F \mapsto \frac{\sqrt{F}}{2} - 2F$, we see that $\text{Trace}(A)$ is positive for $0 < F < 1/16$, zero at $F = 1/16$ and positive for $1/16 < F < 1/4$. Also, by the continuity of $\text{Trace}(A)$ as a function of $\kappa$ and $F$, it maintains
its sign along curves in the parameter space that cross the parabola, except those that cross at the point \((\kappa, F) = (1/16, 1/16)\).

The quantity \(\text{Determinant}(A)\) as a function of \(\kappa\) and \(F\) vanishes on the parabola. To determine its sign along a curve in the parameter space that crosses into the region bounded by the parabola and the coordinate axes, note that \(F > 4(F + \kappa)^2\) in this region; thus, in the bounded region and near the boundary parabola, there is some (small) \(\epsilon > 0\) such that

\[
\frac{4(F + \kappa)^2}{F} = 1 - \epsilon.
\]

The value of \(v\) at the corresponding rest points is

\[
v^2 = \frac{F}{F + \kappa} \left( \frac{1}{2} + \frac{1}{2} \sqrt{1 - (1 - \epsilon)} \right)
= \frac{F}{2(F + \kappa)} (1 \mp \eta),
\]

where \(\eta := \sqrt{\epsilon}\). The sign of \(\text{Determinant}(A)\) is determined by the sign of \(v^2 - F\), which is given by

\[
v_{\pm} - F = \frac{F^2}{4(F + \kappa)^2} \left( 1 \mp 2\eta + \eta^2 \right) - F
= F \left( \frac{F}{4(F + \kappa)^2} (1 \mp 2\eta + \eta^2) - 1 \right)
= F \left( \frac{1}{1 - \eta^2} (1 \mp 2\eta + \eta^2) - 1 \right)
= \frac{2F\eta}{1 - \eta^2} (1 \mp 1 + \eta).
\]

From this computation, it is clear that \(\text{Determinant}(A)\) is positive (respectively, negative) at the rest point whose second component is \(v_-\) (respectively, \(v_+\)). Also, the value of this determinant goes to zero as \(\epsilon > 0\) approaches zero.

In summary, a source and a saddle appear in a saddle node bifurcation upon crossing the parabola into the bounded region at points on the parabola corresponding to \(0 < F < 1/16\); a sink and a saddle appear in a saddle node bifurcation upon crossing the parabola into the bounded region at points on the parabola corresponding to \(1/16 < F < 1/4\).
Another type of bifurcation, called Hopf bifurcation, occurs in the regime where \( A \) has complex eigenvalues; it occurs whenever, along a curve in the parameter space, a pair of complex conjugate eigenvalues crosses the imaginary axis in the complex plane with nonzero speed. A normal form for this important bifurcation is most easily understood in polar coordinates \((r, \theta)\);

\[
\dot{r} = br \pm r^3, \quad \dot{\theta} = \omega + ar^2.
\]

the corresponding system in Cartesian coordinates has the linear part

\[
\dot{x} = bx - \omega y, \quad \dot{y} = \omega x + by.
\] (4.68)

With \( \omega \neq 0 \) and \( a \neq 0 \) fixed, a pair of complex conjugate eigenvalues \((b \pm i\omega)\) crosses the imaginary axis as the bifurcation parameter \( b \) passes through zero in the positive direction. The rest point at the origin thus changes from a sink to a source. For the plus sign and \( b < 0 \), there is an unstable limit cycle (that is, an isolated periodic orbit), which is a circle of radius \( \sqrt{-b} \) that disappears into the rest point at \( b = 0 \). The bifurcation in this case is called a subcritical Hopf bifurcation. For the minus sign, a stable limit cycle (which in this special case is the circle with radius \( \sqrt{b} \)) emerges out of the rest point as \( b \) increases from zero.

We can detect the Hopf bifurcation by finding the curve(s) in the parameter space where the characteristic equation of the linearization at a rest point has pure imaginary eigenvalues. Generically, there is a Hopf bifurcation as a curve in the parameter spaces crosses this “Hopf curve.”

The eigenvalues of \( A \) are pure imaginary whenever its trace vanishes and its determinant is positive. For parameter values in the region of parameter space bounded by the saddle node bifurcation curve, the system matrix of the linearization at the rest point \( v_+ \) has positive determinant. By the formulas in display (4.67), the trace of this matrix at \( v_+ \) is given by \( \kappa - v_+^2 \). Using the formulas in display (4.65) and

\[
G := 1 - \frac{4(F + \kappa)^2}{F},
\]

the trace vanishes if and only if \(( \kappa, F)\) lie on the Hopf curve

\[
\kappa = \frac{F^2}{4(F + \kappa)^2(1 + \sqrt{G})^2}.
\]
After some algebra, the Hopf curve in the first quadrant is also given by

$$(\kappa + F)^2 = F\sqrt{\kappa},$$

a quadratic equation for $F$ that has solutions

$$F = \frac{\sqrt{\kappa} - 2\kappa \pm \sqrt{\kappa - 4\kappa^{3/2}}}{2}.$$ (4.69)

Because the trace of $A$ (given by $\kappa - v^2$) vanishes and the determinant of $A$ (given by $(F + \kappa)(v^2 - F)$) is positive on the Hopf curve, we must have $F < \kappa$ on this curve. By inspection of the graphs of the solutions (4.69), it follows that this condition is satisfied only with the minus sign on the interval $0 < \kappa < 1/16 = 0.0625$. The corresponding graph, depicted in Figure 4.11, is the Hopf curve. It meets the saddle node curve at $(\kappa, F) = (0, 0)$ and $(\kappa, F) = (1/16, 1/16)$.

Outside the saddle-node curve, the system (4.64) has one stable rest point at $(u, v) = (1, 0)$. Inside this curve the system has three rest points. Above the Hopf bifurcation curve and inside the saddle-node curve there are three rest points, two sinks and one saddle. Below the Hopf curve and above the saddle-node curve there is one sink, one source and one saddle. There is a critical value $\kappa_* \approx 0.0325$ of $\kappa$ such that for $F$ decreasing and $\kappa < \kappa_*$, the Hopf bifurcation is supercritical, for $\kappa > \kappa_*$, the Hopf bifurcation is subcritical.

To prove the statements about the subcritical and supercritical Hopf bifurcations is not trivial. An idea for a proof is to translate the corresponding rest point to the origin of a new coordinate system and then transform (by a linear transformation) the resulting system so that its linearization has the normal form for the Hopf bifurcation given in display (4.68). After this process is complete, the resulting second and third-order terms of the right-hand side of the system of differential equations, determine the “stability index,” which gives the direction of the Hopf bifurcation (see [15]). For our differential equation in the form

$$\dot{x} = \alpha x - y + p(x, y), \quad \dot{y} = x + \alpha y + q(x, y)$$

where $p = \sum_{j=1}^{\infty} p_j(x, y)$ and $q = \sum_{j=1}^{\infty} q_j(x, y)$ with

$$p_j(x, y) := \sum_{i=0}^{j} a_{j-i, i} x^{j-i} y^i, \quad q_j(x, y) := \sum_{i=0}^{j} b_{j-i, i} x^{j-i} y^i,$$
Figure 4.12: The left panel depicts the phase portrait for system (4.64) for the parameter values $\kappa = 0.05$ and $F = 0.02725$. An orbit in the unstable manifold of the saddle point is asymptotic to a spiral sink. In the opposite direction on the stable manifold, an orbit is asymptotic to the sink at $(1,0)$. The right panel depicts the phase portrait for the parameter values $\kappa = 0.05$ and $F = 0.0265$. The positions of the stable and unstable manifolds have crossed. An orbit in the stable manifold of the saddle point is asymptotic to an unstable limit cycle, which surrounds a spiral sink (not depicted).

The sign of the stability index is given by the sign of the expression

$$L_4 = \frac{1}{8}(a_{20}a_{11} + b_{21} + 3a_{30} - b_{02}b_{11} + 3b_{03} + 2b_{02}a_{02} - 2a_{20}b_{20} - b_{20}b_{11} + a_{12} + a_{02}a_{11}).$$

(4.70)

If the quantity $L_4 > 0$ (respectively, $L_4 < 0$), then our Hopf bifurcation is subcritical (respectively, supercritical).

The bifurcation diagram in Figure 4.11 is not complete. For example, the existence of the subcritical Hopf bifurcation for $\kappa > \kappa_*$ as $F$ decreases through the Hopf curve means that an unstable limit cycle disappears into the corresponding rest point at this bifurcation. Where did the limit cycle come into existence for some larger value of $F$? The answer is that there is a third bifurcation in our system called a homoclinic loop bifurcation.

Two phase portraits (corresponding to a value of the parameter $F$ slightly larger than its critical value and slightly smaller than its critical value) are
Figure 4.13: The phase portrait for system (4.64) for the parameter values $\kappa = 0.02$ and $F = 0.0039847$. An orbit in the unstable manifold of the saddle point is asymptotic to a stable limit cycle. In the opposite direction on the stable manifold, an orbit is asymptotic to the sink at $(1,0)$.

developed in Figure 4.12. Pay attention to the portion of the unstable manifold of the saddle point (which is at the point where two curves appear to cross) that lies above and to the left of the saddle; the other part of the unstable manifold goes to the sink at $(u,v) = (1,0)$. In the left panel, the unstable manifold lies below the stable manifold; in the right panel, it lies above. Between the parameter values corresponding to these phase portraits, there is a value of $F$ where the stable and unstable manifolds meet to form a (homoclinic) loop. During this process, there is a sink surrounded by the loop. In the second panel the stable manifold winds around this sink. But, it can’t be asymptotic to the sink. So, (by the Poincaré-Bendixson theorem) there must be an unstable periodic orbit surrounding the sink and surrounded by the spiral formed by the stable manifold (see [15]). This is an example of a homoclinic loop bifurcation.

A homoclinic loop bifurcation also occurs for $\kappa < 0.035$ as $F$ decreases to a critical value below the Hopf curve. In this case the homoclinic loop bifurcation absorbs a stable limit cycle. An example of this cycle near the homoclinic loop bifurcation is depicted in Figure 4.13.

The conjectural position of the homoclinic loop curve in the bifurcation diagram is between the saddle-node and Hopf curve for $\kappa < 0.035$ and above
the Hopf curve for $\kappa > 0.035$. Its end points are the same as the end points of the Hopf curve (see Exercise 4.36 and [4] for more details of the bifurcation analysis of this system).

**Exercise 4.35.** Show that there is a critical value $\kappa_* \approx 0.0325$ of $\kappa$ such that, for $\kappa < \kappa_*$, system (4.64) has a supercritical Hopf bifurcation as $F$ decreases through the Hopf curve (4.69).

**Exercise 4.36.** Make a numerical study of the homoclinic loop curve for system (4.64) and plot it together with the saddle-node and Hopf curves. Hint: Public domain software such as AUTO or XPPAUTO can approximate bifurcation curves.

**Exercise 4.37.** Add a periodic forcing term to system (4.64) and compute the stroboscopic Poincaré map (that is, start at some initial condition integrate forward for one period of the forcing, plot the final point, and iterate this process. Vary the amplitude and frequency of the forcing and plot some typical phase portraits of the Poincaré map.

**Exercise 4.38.** Add a periodic parametric forcing term to system (4.64) and compute the stroboscopic Poincaré map (that is, start at some initial condition integrate forward for one period of the forcing, plot the final point, and iterate this process. Here the most interesting scenario is to make $F$ or $\kappa$ periodically time-dependent so that these parameters sweep through some of their bifurcation values (cf. [36]). For example, fix $\kappa = 0.02$ and replace $F$ by

$$F = 0.004 + 0.0005 \sin(\omega t),$$

where $\omega$ is small (perhaps $\omega = 0.0001$) so that $F$ changes relatively slowly. See if you can obtain a long transient that exhibits beats. Our system is not well-suited to dynamic bifurcation because the basin of attraction of the sink at $(1, 0)$ is likely to capture our orbit.

**Exercise 4.39.** A simple model (derived from the Oregonator model) for an oscillating chemical reaction between the concentrations $x$ and $z$ of two chemical species is given by

$$ex' = \frac{fz(\gamma - x)}{\gamma + x} + x(1 - x),$$

$$z' = x - z,$$

where $\epsilon$, $f$, and $\gamma$ are positive parameters. For $\gamma$ fixed at $\gamma = 8 \times 10^{-4}$, determine the parameters $\epsilon$ and $f$ for which oscillations occur. Draw a bifurcation diagram for the parameter space.
Exercise 4.40. The Brusselator model for a hypothetical oscillating chemical reaction is

\[ \begin{align*}
A & \rightarrow B \\
B + X & \rightarrow Y + D \\
2X + Y & \rightarrow 3X \\
X & \rightarrow E
\end{align*} \]

Assume that the concentrations of \( A \) and \( B \) are constant. (a) Write the rate equations for the concentrations of the species \( X \) and \( Y \). (b) Determine a change of variables (including time) for your equation to obtain the dimensionless form

\[ \begin{align*}
\dot{x} &= a - bx + x^2y - x, \\
\dot{y} &= bx - x^2y.
\end{align*} \]

(c) Determine the steady state(s) of the dimensionless system and their stability types. (d) Show that oscillations can occur for some parameter values. (d) Draw a bifurcation diagram for oscillatory solutions in the \((a,b)\) parameter space.

Exercise 4.41. The Oregonator is given by the reactions

\[ \begin{align*}
A + Y & \rightarrow X + P \\
X + Y & \rightarrow 2P \\
A + X & \rightarrow 2X + 2Z \\
2X & \rightarrow A + P \\
B + Z & \rightarrow \frac{1}{2}fY
\end{align*} \]

where \( A, B, \) and \( P \) are assumed constant and \( f \) is a parameter \( \frac{1}{2} < f < 1 \). (a) Show that the rate equations can transformed to the dimensionless system

\[ \begin{align*}
\epsilon x' &= \gamma y - xy + x(1 - x), \\
\delta y' &= \gamma y - xy + fz, \\
z' &= x - z.
\end{align*} \]

Also express the parameters \( \epsilon, \delta, \) and \( \gamma \) as functions of the reaction rates. (b) Assume that the parameters are

\[ \epsilon \approx 4 \times 10^{-2}, \quad \delta \approx 4 \times 10^{-4}, \quad \gamma \approx 8 \times 10^{-4}. \]

Consider \( \delta \) as a small parameter; \( \epsilon \) and \( \gamma \) as parameters. In the limiting case with \( \delta = 0 \), argue that in the region \( x > 0 \) it makes sense to take

\[ y = \frac{fz}{q + x}. \]
What ingredients would be needed to make a rigorous argument? Use the formula for $y$ to obtain the system

$$
\epsilon x' = \frac{f z (\gamma - x)}{\gamma + x} + x(1 - x), \quad z' = x - z
$$

studied in Exercise 4.39.

**Diffusion and Spatial Discretization**

We have succeeded in understanding the bifurcation diagram for the Gray-Scott model without diffusion. What happens when diffusion is introduced?

It seems reasonable that if we add sufficiently small diffusion ($\lambda$ and $\mu$) and choose our parameters $\kappa$ and $F$ in the region where the only steady state is $(u, v) = (1, 0)$, then this spatially constant steady state will also be a stable steady state for the PDE (4.55). On the other hand, it is not at all clear what happens for parameter values close to the bifurcations for the reaction equations (4.64). Let us explore the dynamical behavior of the PDE near these values using numerical approximations.

By adding the Taylor series (4.60) and rearranging terms, note that

$$
F''(x) = \frac{F(x + h) - 2F(x) + F(x - h)}{h^2} + O(h^2).
$$

(4.71)

This approximation is used to discretize the second derivatives in our PDE.

The first step of our numerical procedure is to choose a spatial discretization of a spatial domain. For the Gray-Scott model there is no natural spatial domain; it is just some region in two-dimensional space. Following Pearson [55], let us consider the spatial domain to be the square $[0, L] \times [0, L]$ in the $(x, y)$-plane. A lattice (or grid) in this square is defined by the points $(i\Delta x, j\Delta y)$ (called nodes), for $i = 0, 1, 2, \ldots, m$ and $j = 0, 1, 2, \ldots, n$, where $\Delta x := L/m$ and $\Delta y := L/n$. As in Euler’s method for ODEs, let us also choose a time step $\Delta t$. The concentration $u(i\Delta x, j\Delta y, k\Delta t)$ is denoted by $U_{i,j}^k$ and $v(i\Delta x, j\Delta y, \Delta t)$ by $V_{i,j}^k$.

The next step is to approximate the PDE by difference quotients; this idea leads to many possible alternatives. Perhaps the simplest viable scheme
is the forward Euler method given by

\[
U_{i,j}^{k+1} = U_{i,j}^k + \Delta t \left[ \frac{\lambda}{\Delta x^2} (U_{i+1,j}^k - 2U_{i,j}^k + U_{i-1,j}^k) + \frac{\lambda}{\Delta y^2} (U_{i,j+1}^k - 2U_{i,j}^k + U_{i,j-1}^k) 
+ F(1 - U_{i,j}^k) - U_{i,j}^k (V_{i,j}^k)^2 \right],
\]

\[
V_{i,j}^{k+1} = V_{i,j}^k + \Delta t \left[ \frac{\mu}{\Delta x^2} (V_{i+1,j}^k - 2V_{i,j}^k + V_{i-1,j}^k) + \frac{\mu}{\Delta y^2} (V_{i,j+1}^k - 2V_{i,j}^k + V_{i,j-1}^k) 
+ U_{i,j}^k (V_{i,j}^k)^2 - (F + \kappa)V_{i,j}^k \right];
\]

(4.72)

it is exactly Euler’s method applied to PDEs.

The third step is to impose the boundary conditions. For periodic boundary conditions, we will update the concentrations at (what we will call) the interior nodes corresponding to \(i = 0, 1, 2, \ldots, m - 1\) and \(i = 0, 1, 2, \ldots, m - 1\). To do this, we will require values at the “fictitious nodes” with \((i, j)\)-coordinates \((-1, j)\), for \(j = 0, 1, 2, \ldots, n - 1\) and \((i, -1)\), for \(i = 0, 1, 2, \ldots, m - 1\). These are defined by

\[
U_{-1,j}^k = U_{m-1,j}^k, \quad V_{-1,j}^k = V_{m-1,j}^k;
\]

\[
U_{i,-1}^k = U_{i,n-1}^k, \quad V_{i,-1}^k = V_{i,n-1}^k.
\]

Likewise we define values at the unassigned portion of the boundary of the square:

\[
U_{m,j}^k = U_{0,j}^k, \quad V_{m,j}^k = V_{0,j}^k;
\]

\[
U_{i,n}^k = U_{i,0}^k, \quad V_{i,n}^k = V_{i,0}^k.
\]

(4.73)

To complete the grid (for graphics), we can also define \(U_{m,n}^k = U_{0,0}^k\).

To implement this procedure, we start at \(k = 0\) and assign (initial) values of the concentrations at all the interior nodes. The periodic boundary conditions are used to assign values to the unassigned portion of the boundary of the square and at the fictitious nodes. We then update the values of the concentrations (to \(U_{ij}^1\) and \(V_{ij}^1\)) at all the interior nodes using the equations (4.72), and update the values of the concentrations at the reminder of the boundary and the fictitious nodes using the formulas (4.73) and (4.74). This process is continued until some preassigned final time is reached, or until some other test applied to the array of concentrations is met.

To reproduce the numerical experiments in [55] (which are reported with color graphics), use the assignments

\[
L = 2.5, \quad m = n = 256, \quad \lambda = 2 \times 10^{-5}, \quad \mu = 10^{-5}, \quad \Delta t = 1.0
\]
Figure 4.14: The figure depicts the $u$-concentration for a computer generated approximate state of the Gray-Scott model for the parameter values $\kappa = 0.06$, $F = 0.038$, $\lambda = 2 \times 10^{-5}$ and $\mu = 10^{-5}$. The system evolves from the depicted state to states with similar configurations for at least 200,000 time steps of unit length.

with $(\kappa, F)$ chosen near the bifurcation curves in Figure 4.11 and impose the initial data as follows: Set the initial concentrations to the spatial steady state value $(u, v) = (1, 0)$, reset the values in the $20 \times 20$ central square of nodes to $(u, v) = (0.5, 0.25)$, and then perturb every grid point value by a random number that changes the originally assigned value by no more than 1%. Integrate forward for some number of time steps, usually chosen in the range 100,000–200,000, and plot the final approximate concentrations.

The result of a typical numerical experiment, for the parameter values $\kappa = 0.06$ and $F = 0.038$, 100,000 time steps of unit length, and a rendering of the value of the function $(x, y) \mapsto u(100000, x, y)$ into a gray scale on the interval $[0, 1]$, is depicted in Figure 4.14. This pattern seems to evolve over a long time-interval and eventually it reaches a (time-independent) steady state that retains the basic qualitative features in the figure (see Exercise 4.43).

**Numerical Stability**

A virtue of the forward Euler method is that it is easy to program. Its error per step is $O(\Delta x^2 + \Delta y^2)$ in space and $O(\Delta t)$ in time. Unfortunately, there
is a hidden danger: the discretization can introduce instabilities that are not present in the PDE.

To understand how a numerical instability might occur, imagine that the values of the concentrations at the interior grid points are written as a column vector \( W \) (of length \( 2mn \)). The update equations (4.72) can be recast in the matrix form

\[
W^{k+1} = AW^k + H(W^k),
\]

where \( A \) is the matrix corresponding to the discretization of the second derivatives and \( H \) is the nonlinear function whose components are the reaction terms. Is this process stable?

The linear part of our update equation is \( W^{k+1} = AW^k \). At some step of our numerical computation a vector error \( \epsilon \) might be introduced into the computed value of \( W^k \). Let us suppose the error occurs in the computation of \( W^1 \) so that the computed value is \( W^1 + \epsilon \) instead of \( W^1 \). This error will propagate in future steps as follows:

\[
W^2 = AW^1 + A\epsilon, \quad W^3 = A^2W^1 + A^2\epsilon, \ldots, W^{k+2} = A^{k+1}W^1 + A^{k+1}\epsilon.
\]

The error will not cause too much trouble if its propagation by the iteration process remains bounded; that is, if there is some constant \( M \) such that \( |A^k \epsilon| \leq M|\epsilon| \) for all \( k > 1 \). A sufficient condition for the desired stability is that all eigenvalues of \( A \) are in the open unit disk in the complex plane; or, equivalently, the spectral radius of \( A \) is strictly less than one. If at least one eigenvalue of \( A \) lies outside the closed unit disk, then a generic error will grow without bound. (The error will not grow in the unlikely situation that it remain in the eigenspace of an eigenvalue which is in the open unit disk, but all other errors will grow.) In case there are eigenvalues on the unit disk, a stability analysis requires additional information. Consider two simple examples:

\[
A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.
\]

In this example, \( A \) is the \( 2 \times 2 \) identity matrix; its eigenvalues are both one. Since \( A^kv = v \) for every vector \( v \) and every integer \( k \), we have that \( |A^kv| = |v| \). The propagation of \( v \) remains bounded. The eigenvalues of \( B \) are also both equal to one; but, for this matrix

\[
\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^k \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_1 + kv_2 \\ v_2 \end{pmatrix},
\]

we see that a generic error will grow without bound.
Thus, if \( v_2 \neq 0 \), the propagation of \( v \) by iteration of \( B \) grows without bound. These examples are indicative of the general situation. To fully analyze the general case requires some concepts from linear algebra, especially the Jordan canonical form. A simple result, adequate for our purposes, states that if all the eigenvalues of a matrix \( A \) lie in the closed unit disk and the matrix is diagonalizable (that is, there is an invertible matrix \( B \) such that \( B^{-1}AB \) is a diagonal matrix), then errors remain bounded under iteration by \( A \).

To apply these observations to the stability of our numerical methods, let us consider the \( A \)-matrix for the forward Euler method in case there is only one spatial dimension. The forward Euler update equation is given by

\[
U_i^{k+1} = U_i^k + \lambda(U_{i+1}^k - 2U_i^k + U_{i-1}^k),
\]

for \( i = 1, 2, \ldots, m \), where \( \lambda := \mu \Delta t/\Delta x^2 \) and \( \mu \geq 0 \). Using the rearrangement

\[
U_i^{k+1} = (1 - 2\lambda)U_i^k + \lambda U_{i+1}^k + \lambda U_{i-1}^k
\]

and assuming that \( U_{-1}^k = 0 \) and \( U_{m+1}^k = 0 \), it follows that

\[
A = \\
\begin{pmatrix}
1 - 2\lambda & \lambda & 0 & 0 & 0 & \cdots & \lambda \\
\lambda & 1 - 2\lambda & \lambda & 0 & 0 & \cdots & 0 \\
0 & \lambda & 1 - 2\lambda & \lambda & 0 & \cdots & 0 \\
\vdots & & & & & & \vdots \\
0 & 0 & \cdots & 0 & \lambda & 1 - 2\lambda & \lambda \\
\lambda & 0 & 0 & \cdots & 0 & \lambda & 1 - 2\lambda
\end{pmatrix}.
\]

The matrix \( A \) is symmetric (\( A \) is equal to its transpose). By a result from operator theory, a symmetric matrix has real eigenvalues. Thus, the propagation of errors will remain bounded if all eigenvalues lie in the closed interval \([-1, 1]\]. By an application of Gerschgorin’s theorem (A.5), the eigenvalues of \( A \) all lie in the interval \([1 - 4\lambda, 1]\). Thus, there is the possibility of an eigenvalue outside the unit disk if \( 1 - 4\lambda < -1 \); that is, if \( \lambda \geq 1/2 \). To eliminate this possibility our stability criterion is

\[
\mu \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}.
\]

Instabilities that lead to meaningless numerical results might occur if this condition is not met (see Exercise 4.45). The forward Euler method is called...
conditionally stable because the condition (4.76) must be met to avoid instabilities. Unfortunately, the number one is an eigenvalue of $A$ (see Exercise 4.42). Thus, while the propagation of errors will remain bounded, the errors will not be reduced under iteration by $A$.

**Exercise 4.42.** Show that one is an eigenvalue of the matrix (4.75).

**Exercise 4.43.** (a) Perform numerical experiments (using the forward Euler method with periodic boundary conditions) that reproduces a pattern for the Gray-Scott model similar to the pattern in Figure 4.14. (b) Demonstrate that the qualitative results for part (a) do not change with the choice of the choice of step sizes in space and time. For example, repeat the experiment(s) in part (a) with a spatial grid containing twice the number of nodes. (b) Use the code written for part (a) to find a pattern generated by the Gray-Scott model that is clearly different from the pattern in Figure 4.14 and does not correspond to a constant value of $u$.

**Exercise 4.44.** Explore the parameter space for the Gray-Scott model (using the forward Euler method with no flux boundary conditions).

**Exercise 4.45.** Write a forward Euler code for the one dimensional heat equation $u_t = \lambda u_{xx}$ with Neumann boundary conditions on the interval $[0,1]$. Taking into account the stability condition (4.39), demonstrate with carefully designed numerical experiments that numerical instability occurs when the stability condition is not met.

**Exercise 4.46.** Determine the stability criterion analogous to inequality (4.39) for the forward Euler scheme with periodic boundary conditions for the PDE (4.55).

We have explored some of the qualitative long-term behavior of the Gray-Scott model. The forward Euler method converges to the solution of the PDE as the discretization step sizes in space and time approach zero. But, how do we know that the qualitative behavior—for example, the qualitative behavior depicted in Figure 4.14—is present in the solution of the PDE? This is a difficult question. The correct answer: we don’t know! The only way to be sure is to prove a mathematical theorem. On the other hand, we can gain confidence in the correctness of the results of a numerical experiment in several ways that will be discussed. For example, we might compute the answer with different step sizes in (space and time) and confirm that the qualitative features of the computed values do not change (see Exercise 4.43); or, we might recompute with a different numerical scheme.
We should start with a moderately small step size (perhaps the step size equal to one) and decrease the step size until the result of the experiment does not appear to change with the choice of the step size. A decrease in step size theoretically decreases the truncation error. On the other hand, as we decrease the step size, we increase the number of numerical operations (additions, subtractions, multiplications, and divisions) thus increasing the effects of roundoff error due to the finite number of decimal digits stored in the computer. Another source of error is called condition error. One manifestation of this type of error is due to the big O estimates that determine the order of the numerical method. Recall that these estimates state that the error in some approximation is proportional to a power of the step size. But, the size of the constant of proportionality is ignored. It might happen that these constants are very large for our application, in which case it is called ill-conditioned. The errors due to roundoff and condition tend to accumulate with subsequent iterations of our numerical method. Hence, the global error will generally decrease with step size to some minimum value as the order estimates dominate; but, it will increase (perhaps not monotonically) as the step size is decreased further due to the accumulation of roundoff and condition errors.

Quantitative PDE: A computational challenge

The interesting qualitative behavior of the Gray-Scott model is the pattern formation. Because it is difficult to quantify a particular pattern, we will consider a simpler quantitative problem as a vehicle for discussing the fundamental problem of numerical computation: Do we get the correct answer?

Let us suppose we are interested in an accurate and efficient computation of the concentrations $u$ and $v$ over some finite time-interval on the spatial domain $[0, L] \times [0, L]$ with (zero) Neumann boundary conditions and given initial data. We must define (or have our scientific collaborator with expertise in chemistry define) what is meant by “accurate.” For an applied problem, the desired accuracy will often be determined by the accuracy of the measurements in some experiment or the accuracy of the measurements of some physical parameters. The relative error (percent error) for approximation of numbers larger than one is usually more meaningful than the absolute error. Recall that for numbers $a$ and $b$, where $b$ is viewed as an approximation of
Conservation of Mass

For the concentrations in the Gray-Scott model, which will all be in the interval \([0, 1]\), the absolute error is the better choice. In practice, we do not know the number \(a\) that we are trying to approximate. But, it is sometimes possible to obtain a theoretical error estimate.

Let us suppose that an approximation of the solution of the Gray-Scott PDE is desired with an absolute error of less than \(10^{-2}\).

To meet the expectations and desires of our imaginary collaborator, let us set the following mathematical problems:

**Problem 4.47.** For the Gray-Scott model (4.55) on the spatial domain \([0, L] \times [0, L]\) with (zero) Neumann boundary conditions, with \(L = 2.5\), the diffusions

\[\lambda = 2 \times 10^{-5}, \quad \mu = 10^{-5},\]

the parameters

\[F = 0.0225, \quad \kappa = 0.05,\]

and initial data (at time \(t = 0\))

\[u(x, y) = 1 - 0.5 \frac{2^8}{L^8} (x^2 y^2 (x - L)^2 (y - L)^2), \quad (4.77)\]

\[v(x, y) = 0.25 \frac{2^8}{L^8} (x^2 y^2 (x - L)^2 (y - L)^2), \quad (4.78)\]

approximate the average values of the concentrations \(u\) and \(v\) over the spatial domain at time \(T = 1024\) with an absolute error of less than \(10^{-2}\).

**Problem 4.48.** Using the data of Problem 4.47, approximate the average values of the concentrations \(u\) and \(v\) at time \(T = 10\) with an error of less than \(10^{-2}\).

The initial state in Problem 4.47 satisfies the boundary conditions and is chosen to lie between the steady states of the system. (No technical meaning is intended for the word “between.”)

In practice, an “efficient” method is one that can be used to obtain the desired accuracy in a short time. This definition is not precise; the idea is that efficient methods are fast.

Our theoretical estimate of the global error for the forward Euler method is \(O(\Delta t)\) and \(O(\Delta x^2 + \Delta y^2)\). This means that we can expect the absolute
error, which is defined to be the norm of the difference between the solution and its approximation, to be proportional to a quantity controlled by the sizes of $|\Delta t|$ and $|(\Delta x, \Delta y)|^2$. There are many possible choices for norms. In a finite-dimensional space all norms are equivalent, so the choices are not of fundamental importance. On the other hand, the choices of norms might be dictated by the physical problem.

A practical choice for the norm might lead to the representation of the theoretical error estimate in the form

$$|u - u_{\text{appx}}| + |v - v_{\text{appx}}| \leq M(|\Delta t| + \Delta x^2 + \Delta y^2),$$

where $M$ is some unknown constant and the norms on the left-hand side of the inequality are Euclidean norms, maximum norms, or $\ell_1$-norms. The maximum norm would be appropriate if we desire the computed concentrations at all points in the spatial grid to differ from the exact values by no more than some specified amount; that is, we would hope to achieve the result

$$\max_{i,j,k}\{|u(k\Delta t, i\Delta x, j\Delta y) - u_{ij}^k|\} < 10^{-2},$$

$$\max_{i,j,k}\{|v(k\Delta t, i\Delta x, j\Delta y) - v_{ij}^k|\} < 10^{-2},$$

where $i = 0, 1, 2, \ldots, m - 1, j = 0, 1, 2, \ldots, n - 1, \text{and } k = T/\Delta t$.

For Problems 4.47 and 4.48, we are challenged to approximate the average values of the states $u$ and $v$ over the spatial domain. Our numerical method produces approximations of the state variables on a grid of points covering the square $[0, L] \times [0, L]$. The desired averages are the integrals of the states over this domain divided by its area. Hence, we must approximate the integrals. For definiteness and simplicity, the trapezoidal rule is a viable choice. With $k\Delta t = T$ (that is, $k$ corresponds to the final time $T$), we wish to achieve the error estimates

$$\frac{1}{L^2} \int_0^L \int_0^L u(T, x, y) \, dx \, dy - \frac{1}{L^2} \text{Trap}(u) < 10^{-2},$$

$$\frac{1}{L^2} \int_0^L \int_0^L v(T, x, y) \, dx \, dy - \frac{1}{L^2} \text{Trap}(v) < 10^{-2},$$

(4.79)

where

$$\text{Trap}(u) := \frac{\Delta x \Delta y}{4} \sum_{i,j} (u_{i,j}^{k\Delta t} + u_{i+1,j}^{k\Delta t} + u_{i+1,j+1}^{k\Delta t} + u_{i,j+1}^{k\Delta t}).$$
There is no simple method known to prove that the desired error is achieved. Of course, if we know the error exactly, we know the solution. So we cannot expect to prove that a numerical approximation achieves a desired error bound except in some special cases. On the other hand, we can test our numerical results in various ways to achieve a high confidence level that they are correct (cf. [10]).

Use A Convergent Algorithm

The numerical algorithms (for instance, the forward Euler method) used in this book are known to converge to the corresponding exact solutions. The proofs for their convergence are given in books on theoretical numerical analysis. The issue of convergence is fundamental. But, the mathematics required to present the proofs is beyond the scope of this book.

Use A Numerically Stable Algorithm

Our numerical code must be either unconditionally stable; or, we must be careful to satisfy appropriate stability criteria. For example, to use the forward Euler method in one space dimension we must respect its stability criterion (4.39). The appropriate stability criterion for two space dimensions will be discussed in the next subsection (see display (4.81)).

Test Code Against Known Solutions

The first confidence building step is to test the numerical code that implements an algorithm against an example where the exact solution is known. The preferred choice for an example is a special case of the problem under consideration. For the Gray-Scott model with Neumann boundary data, it seems that no explicit solution is known. The next best choice is to use an exact solution for a model with similar features.

Let us test the forward Euler method with Neumann boundary data against the exact solution

\[
\begin{align*}
v(t, x, y) &= e^{-F - 2\lambda x^2/L^2} e^{\pi x \cos \pi L} \cos \pi y, \\
\end{align*}
\]

\[
\begin{align*}
u(t, x, y) &= 1 + v(t, x, y)
\end{align*}
\]
of the linear system
\[
\begin{align*}
u_t &= \lambda \Delta u + F(1 - u), \\
v_t &= \lambda \Delta v - F v, \\
_{ux}(t, 0, y) &= {ux}(t, L, y) = {uy}(t, x, 0) = {uy}(t, x, L) = 0, \\
v_x(t, 0, y) &= v_x(t, L, y) = v_y(t, x, 0) = v_y(t, x, L) = 0
\end{align*}
\]  

(4.80)

defined on the square $[0, L] \times [0, L]$.

The Neumann boundary condition, which in our case states that the normal derivative vanishes on the boundary, is implemented using the central difference approximation of the first derivative. For example, the vanishing of the partial derivative $u_x(t, 0, y)$ (along the left-hand boundary of the spatial domain) leads to the numerical approximation
\[
u(k\Delta t, -\Delta x, j\Delta y) = u(k\Delta t, \Delta x, j\Delta y).
\]

After the boundary data is imposed, the forward Euler difference equations are applied to all grid points including the grid points on the boundary of the rectangular domain. The required values of the concentrations outside the spatial domain (which are required in the difference equations applied to grid points on the boundary of the spatial domain) are assigned by the Neumann boundary data to values at grid points in the spatial domain.

How should we make a numerical experiment to test against the known solution of our linear problem? The simplest idea would be to simply choose a grid size, compute over some time interval and compare the answer with the exact answer. How do we judge the success of our experiment? Which time-step size and grid size should we choose? How small an error is acceptable?

To answer the questions posed in the last paragraph, let us design an experiment that uses some numerical analysis.

Recall the stability criterion for the forward Euler method: For the diffusion equation in the case of one space-dimension, the method is stable if
\[
\frac{\lambda \Delta t}{\Delta x^2} \leq \frac{1}{2}.
\]

For two space-dimensions and $\Delta x = \Delta y$, the stability criterion is
\[
CFL := \frac{\Delta t}{\Delta x^2} \leq \frac{1}{4}.
\]  

(4.81)
Conservation of Mass

The denominator on the right-hand side of this inequality comes from the factor 4 in the forward Euler difference equations written in the form

\[ U_{i,j}^{k+1} = (1 - 4 \text{ CFL})U_{i,j}^k + \text{CFL}[(U_{i+1,j}^k + U_{i-1,j}^k) + U_{i,j+1}^k + U_{i,j-1}^k], \]

\[ V_{i,j}^{k+1} = (1 - 4 \text{ CFL})U_{i,j}^k + \text{CFL}[(U_{i+1,j}^k + U_{i-1,j}^k) + U_{i,j+1}^k + U_{i,j-1}^k]. \]

The quantity \( 1 - 4 \text{ CFL} \) (which replaces \( 1 - 2\lambda \) in our stability analysis of the one-dimensional forward Euler scheme) appears along the diagonal of the matrix representation of these difference equations. The abbreviation CFL stands for Courant-Friedricks-Lewy. Indeed, this quantity is often called the Courant-Friedricks-Lewy number (see [18]).

According to theory, the global error for the forward Euler method is \( O(\Delta t) \) and \( O(\Delta x^2) \). If we fix CFL \( \leq 1/4 \) and choose the time-step

\[ \Delta t = \frac{\text{CFL} \Delta x^2}{\lambda}, \]

then the global error should be \( O(\Delta x^2) \). This observation is the basis for a useful test: Put \( \lambda = 2 \times 10^{-5} \), \( \mu = \lambda \), \( F = 10^{-3} \) and \( L = 2.5 \); for \( \ell = 3, 4, \ldots, 8 \), set \( m = 2^\ell \), \( \Delta x = \Delta y = L/m \), and choose the grid \((i\Delta x, j\Delta y)\) with \( i = 0, 2, \ldots, m \) and \( j = 0, 2, \ldots, m \); define

\[ \text{CFL} = \lambda \frac{256^2}{L^2} \approx 0.2097, \]

\[ \Delta t = \text{CFL} \frac{\Delta x^2}{\lambda} = 2^{16-2\ell}, \]

and apply the forward Euler method recursively for \( 2^{2\ell-6} \) time steps so that the final time is \( T = 1024 \). For each choice of \( \ell \) compute the error \( e_\ell \) to be the maximum norm of the difference of the exact and computed values at the final time over all interior grid points; and, for \( \ell > 3 \), also compute the quantity \( e_{\ell-1}/e_\ell \). Note that if

\[ e_{\ell-1} = K\Delta x^2 \quad \text{and} \quad e_\ell = K \left( \frac{\Delta x}{2} \right)^2, \]

then

\[ \frac{e_{\ell-1}}{e_\ell} = 4; \]

in other words, our code is performing properly if the computed quotients of the errors approaches 4 as we increase the grid size. On the other hand, let
<table>
<thead>
<tr>
<th>Interior Grid</th>
<th>∆t</th>
<th>error</th>
<th>pe/ce</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 × 8</td>
<td>1024</td>
<td>0.362346</td>
<td></td>
</tr>
<tr>
<td>16 × 16</td>
<td>256</td>
<td>0.05383</td>
<td>6.7313</td>
</tr>
<tr>
<td>32 × 32</td>
<td>64</td>
<td>0.0126742</td>
<td>4.24721</td>
</tr>
<tr>
<td>64 × 64</td>
<td>16</td>
<td>0.00312647</td>
<td>4.05384</td>
</tr>
<tr>
<td>128 × 128</td>
<td>4</td>
<td>0.000779079</td>
<td>4.01304</td>
</tr>
<tr>
<td>256 × 256</td>
<td>1</td>
<td>0.000194612</td>
<td>4.00323</td>
</tr>
</tbody>
</table>

Figure 4.15: Numerical experiment for system (4.80) with \( \lambda = 10^{-3} \), \( F = 10^{-1} \) and total integration time \( T = 1024 \). The last column is the previous maximum absolute error (over all the nodes in the spatial domain) divided by the current absolute error.

us note that the time-step size decreases as the grid size increases. For very large grids, roundoff errors are likely to accumulate and destroy the expected second-order global error.

The results of a numerical experiment are given in Figure 4.15. It seems that the code is performing as expected with a second-order global error relative to the spatial discretization provided that the time step is chosen so that the Courant-Friedricks-Lewy condition is met. Moreover, with a grid size of 256 × 256 and \( \Delta t = 1 \), the computed value agrees with the exact value to three decimal places; the error is less than \( 10^{-3} \).

Warnings for the novice programmer: Implementing algorithms into computer code for solving partial differential equations is complicated. Standard mistakes occur when indexing arrays. Be careful when coding arrays with indices starting at zero (or some other number not equal to one). Remember that in most computer systems, the equality sign is used to mean replacement; for example, the statement \( a = a + 1 \) replaces the current value of \( a \) with the value \( a + 1 \). When updating arrays while solving difference equations, it might be tempting to replace an array \( U_{ij} \) with its updated value as defined perhaps by the forward Euler formulas. But, using the same symbol \( U_{ij} \) could lead to trouble: the original value \( U_{ij} \) might be needed later. Be sure to incorporate comments on every block within your code. Without thoughtful comments, your code will be incomprehensible when you return to check, debug, or modify its content. Write code using subroutines that can be checked independently. This method helps to preserve modularity. Also, well-tested subroutines can be used without modification in other projects.
Declare the types of all variables used in your code. The only way to debug a code is by printing out the values of variables at strategic locations within your code.

**Test Code Using a Posteriori Estimates**

How can we test a code to see if it performs as expected with respect to the theoretical global error in case the exact solution is not available? Our answer to this question uses a gem of numerical analysis: Richardson extrapolation. We will discuss this method and apply it to the Gray-Scott model.

By adding the formulas (4.60) and rearranging the resulting expression, we have the centered difference approximation of the second derivative given by

\[ U_0^0(h) := \frac{F(x + h) - 2F(x) + F(x - h)}{h^2} = F''(x) + 2 \sum_{j=1}^{\infty} \frac{F^{(2j)}(x)h^{2j}}{(2j)!}. \] (4.82)

Take note of the form of the error, which is a Taylor series in \( h^2 \). Richardson extrapolation requires knowledge of the form of the error written as a series in powers of some parameter, which is usually the size of the discretization as in this example. We will proceed under the assumption that the form of the error is a power series in even powers of the discretization size \( h \); but, this requirement is not essential (see Exercise 4.50 and compare [10, Appendix A]). Let us assume \( U_0^0(h) \) is used to approximate some quantity \( U \) such that

\[ U_0^0(h) = U + \sum_{j=1}^{\infty} K_{2j}^0 h^{2j}, \] (4.83)

where the \( K_{2j} \) are constants. This is exactly the case for the centered difference formula used to approximate the second derivative as soon as we fix \( x \).

Since the formula 4.83 is supposed to hold for all \( h \), it holds for \( h/2 \), for which we have

\[ U_0^0\left(\frac{h}{2}\right) = U + \sum_{j=1}^{\infty} K_{2j}^0 \frac{h^{2j}}{2^{2j}}. \] (4.84)

The fundamental observation of Richardson extrapolation is simple: The formulas \( U^1(h) \) and \( U^1(h/2) \) both approximate \( U \) with a \( O(h^2) \) error. By
Richardson extrapolation is used in several different ways. A direct application yields high-order approximation formulas from low-order formulas.

For instance, by an application of formula (4.86) with $\ell = 1$ to the centered difference (4.82), we produce the approximation

$$F''(x) = -\frac{1}{3h^2}(F(x+h) - 16F(x + \frac{h}{2}) + 30F(x) - 16F(x - \frac{h}{2}) + F(x-h)) + O(h^4).$$

(4.87)
Conservation of Mass

Discretization Approximation Error
\begin{align*}
h = 1/2^2 & \quad -0.646260455 \quad 0.00337648457 \\
h = 1/2^4 & \quad -0.649425496 \quad 0.000211442825 \\
h = 1/2^6 & \quad -0.649636113 \quad 8.26053417E-07 \\
h = 1/2^8 & \quad -0.649636935 \quad 3.6373629E-09 \\
h = 1/2^{10} & \quad -0.649636924 \quad 1.48132472E-08 \\
h = 1/2^{12} & \quad -0.649636745 \quad 1.93627182E-07 \\
h = 1/2^{14} & \quad -0.649635315 \quad 1.62413866E-06 \\
h = 1/2^{16} & \quad -0.649536133 \quad 0.000100806268 \\
\end{align*}

Figure 4.17: The centered difference approximation of the second derivative of \( \sin x \) at \( x = 1/\sqrt{2} \) and the approximation error are listed.

Note that the higher-order accuracy requires two additional function evaluations. To avoid the extra function evaluations, we can compute a sequence of approximations using the original centered difference formula for step sizes

\[ h, \quad \frac{h}{2}, \quad \frac{h}{2^2}, \quad \ldots, \quad \frac{h}{2^j}, \quad \ldots \]

and apply Richardson extrapolation to this sequence of numbers. An example of this procedure for the centered difference formula is tabulated in Figure 4.16. Note that accuracy to over 12 digits is obtained using \( h = \Delta x = 1/128 \) at the bottom of the third column of the Richardson table.

The centered difference approximation of the second derivative (of a sufficiently smooth) function certainly converges to the second derivative of the function. This fact is consistent with the data in Figure 4.16. On the other hand, in a numerical computation with finite decimal approximations of numbers, the round off error and the condition error (which arises here from the error made in the function evaluations) will destroy the expected convergence as the discretization size decreases. The typical situation for a numerical approximation is for the error of the computation to decrease with the discretization size until it reaches some optimal value and then the error begins to increase as the roundoff and condition errors dominate. This phenomenon is illustrated by the table in Figure 4.17. Using ten-digit arithmetic, the optimal error for approximating the second derivative of \( \sin \) at \( 1/\sqrt{2} \) is on the order of \( 10^9 \) (which is perhaps not too surprising) with a discretization size of \( 1/2^{12} \approx 0.000244141 \) (which might be surprising).
Let us return to our original question on assessing the order of the error of a numerical procedure with respect to a discretization parameter applied to approximate a quantity that is not known. Good estimates are possible using a reinterpretation of Richardson’s extrapolation.

Suppose the theoretical error for an order-two procedure \( U^0(h) \) is known to be as in display (4.83), and we wish to verify that our implementation is performing with the expected \( O(h^2) \) error. We again compute \( U^0(h/2) \) and subtract the error formulas to obtain

\[
U^0(h) - U^0(h/2) = \frac{3}{4} K_2 h^2 + O(h^4).
\]

Next, we substitute

\[
K_2 h^2 = U^0(h) - U + O(h^4)
\]

to obtain the formula

\[
U^0(h) - U^0(h/2) = \frac{3}{4} (U^0(h) - U) + O(h^4);
\]

or, its rearrangement

\[
U^0(h) - U = \frac{4}{3} (U^0(h) - U^0(h/2)) + O(h^4).
\] (4.88)

The quantity

\[
\text{RE}(h) := \frac{4}{3} (U^0(h) - U^0(h/2)),
\] (4.89)

which can be computed without knowing \( U \), is a \( O(h^4) \) approximation of the error \( U^0(h) - U \). Under our assumption that the error is \( O(h^2) \), we can substitute \( U^0(h) - U = O(h^2) \) in formula (4.88) and conclude that

\[
\text{RE}(h) = O(h^2).
\]

The desired test of our procedure is obtained by computing \( \text{RE}(h) \). In fact, we can be confident that our procedure is performing at the theoretical second-order if

\[
\frac{\text{RE}(h)}{\text{RE}(h/2)} \approx 2^2.
\]

The factor \( 4/3 \) cancels in the quotient; it does not play a role in this test.
Figure 4.18: The order estimates in this figure are for an implementation of the forward Euler method using the data of Problem 4.48. The floating point numbers are computed for the average of the concentrations \((u + v)/2\) using formula (4.91).

<table>
<thead>
<tr>
<th>(h)</th>
<th>32 × 32</th>
<th>64 × 64</th>
<th>128 × 128</th>
<th>256 × 256</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\Delta t = 1/4)</td>
<td>0.955307</td>
<td>0.955303</td>
<td>0.955301</td>
<td>0.955300</td>
</tr>
<tr>
<td>(\Delta t = 1/8)</td>
<td>0.977365</td>
<td>0.977362</td>
<td>0.977361</td>
<td>0.977361</td>
</tr>
<tr>
<td>(\Delta t = 1/16)</td>
<td>0.988587</td>
<td>0.988585</td>
<td>0.988585</td>
<td>0.988585</td>
</tr>
<tr>
<td>(\Delta t = 1/32)</td>
<td>0.994266</td>
<td>0.994266</td>
<td>0.994266</td>
<td>0.994265</td>
</tr>
<tr>
<td>(\Delta t = 1/64)</td>
<td>0.997126</td>
<td>0.997126</td>
<td>0.997126</td>
<td>0.997126</td>
</tr>
<tr>
<td>(\Delta t = 1/128)</td>
<td>0.998561</td>
<td>0.998561</td>
<td>0.998561</td>
<td>0.998561</td>
</tr>
</tbody>
</table>

Figure 4.19: The error estimates in this figure are for an implementation of the forward Euler method using the data of Problem 4.48. The floating point numbers are computed by averaging the absolute error estimates (4.93) corresponding to the two concentrations.

<table>
<thead>
<tr>
<th>(h)</th>
<th>32 × 32</th>
<th>64 × 64</th>
<th>128 × 128</th>
<th>256 × 256</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\Delta t = 1/4)</td>
<td>0.0068953</td>
<td>0.006898</td>
<td>0.00690026</td>
<td>0.00690022</td>
</tr>
<tr>
<td>(\Delta t = 1/8)</td>
<td>0.00353214</td>
<td>0.00353455</td>
<td>0.00353486</td>
<td>0.00353496</td>
</tr>
<tr>
<td>(\Delta t = 1/16)</td>
<td>0.00178909</td>
<td>0.00178955</td>
<td>0.00178951</td>
<td>0.00178949</td>
</tr>
<tr>
<td>(\Delta t = 1/32)</td>
<td>0.000900304</td>
<td>0.000900353</td>
<td>0.000900337</td>
<td>0.000900331</td>
</tr>
<tr>
<td>(\Delta t = 1/64)</td>
<td>0.000451592</td>
<td>0.00045157</td>
<td>0.000451573</td>
<td>0.00045157</td>
</tr>
<tr>
<td>(\Delta t = 1/128)</td>
<td>0.000226156</td>
<td>0.000226133</td>
<td>0.000226138</td>
<td>0.000226139</td>
</tr>
</tbody>
</table>
Using the same idea in the general case, where the error is given by an order estimate whose first non-zero term is $K_j h^j$, the corresponding ratio is

$$\frac{U^0(2h) - U^0(h)}{U^0(h) - U^0(\frac{h}{2})} = 2^j,$$

(4.90)

where $j$ is the order of the method (see Exercise 4.50). In other words, the order of the method is given (approximately) by

$$j = \log_2 \frac{U^0(2h) - U^0(h)}{U^0(h) - U^0(\frac{h}{2})}.$$  

(4.91)

**Exercise 4.49.** Apply the test (4.90) to confirm that Euler’s method is order one for the Gray-Scott system with no diffusion.

**Exercise 4.50.** Show that the Richardson extrapolation formula for an approximation scheme $U^0(h)$, where

$$U^0(h) = U + \sum_{j=1}^{\infty} K_j h^j,$$

is given inductively by

$$U^j(h) = \frac{2^j U^{j-1}(\frac{h}{2}) - U^{j-1}(h)}{2^j - 1},$$

(4.92)

and the error estimate (corresponding to (4.89)) is

$$RE^j(h) := 2(U^0(h) - U^0(h/2)).$$

(4.93)

In Problem 4.47, the partial differential equation includes diffusion and nonlinear reaction. The forward Euler method is a second-order scheme for approximating the diffusion process provided that we choose the time step so that the CFL number is fixed and less than $\frac{1}{4}$. On the other hand, the forward Euler method applied to the reaction ODE is first-order by Exercise 4.49. Thus, we can expect the forward Euler method to be first-order when applied to the Gray-Scott PDE, which combines reaction and diffusion. To confirm this estimate by a numerical experiment, we can compute (with the data in
Figure 4.20: Four Richardson extrapolation tables for the average concentration \((u + v)/2\) using the formula (4.92) applied to an implementation of the forward Euler method for the data of Problem 4.48 are listed. The spatial grid sizes for the tables (top to bottom) are \(32 \times 32, 64 \times 64, 128 \times 128,\) and \(256 \times 256.\)

<table>
<thead>
<tr>
<th>(h)</th>
<th>(\Delta t = 1/4)</th>
<th>(\Delta t = 1/8)</th>
<th>(\Delta t = 1/16)</th>
<th>(\Delta t = 1/32)</th>
<th>(\Delta t = 1/64)</th>
<th>(\Delta t = 1/128)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\Delta t)</td>
<td>(U^1)</td>
<td>(U^2)</td>
<td>(U^3)</td>
<td>(U^4)</td>
<td>(U^5)</td>
<td>(U^6)</td>
</tr>
<tr>
<td>(\Delta t = 1/4)</td>
<td>0.453113</td>
<td>0.453069</td>
<td>0.453046</td>
<td>0.453035</td>
<td>0.453029</td>
<td>0.453026</td>
</tr>
<tr>
<td>(\Delta t = 1/8)</td>
<td>0.453069</td>
<td>0.453027</td>
<td>0.453024</td>
<td>0.453024</td>
<td>0.453024</td>
<td>0.453023</td>
</tr>
<tr>
<td>(\Delta t = 1/16)</td>
<td>0.453035</td>
<td>0.453035</td>
<td>0.453039</td>
<td>0.453043</td>
<td>0.453045</td>
<td>0.453031</td>
</tr>
<tr>
<td>(\Delta t = 1/32)</td>
<td>0.453029</td>
<td>0.453039</td>
<td>0.453034</td>
<td>0.453034</td>
<td>0.453034</td>
<td>0.453033</td>
</tr>
<tr>
<td>(\Delta t = 1/64)</td>
<td>0.453024</td>
<td>0.453034</td>
<td>0.453033</td>
<td>0.453034</td>
<td>0.453034</td>
<td>0.453033</td>
</tr>
<tr>
<td>(\Delta t = 1/128)</td>
<td>0.453024</td>
<td>0.453034</td>
<td>0.453033</td>
<td>0.453034</td>
<td>0.453034</td>
<td>0.453033</td>
</tr>
</tbody>
</table>

Figure 4.21: The order estimates in this figure are for an implementation of the forward Euler method using the data of Problem 4.47. The floating point numbers are computed for the average of the concentrations \((u + v)/2\) using formula (4.91).

<table>
<thead>
<tr>
<th>(h)</th>
<th>(\Delta t = 1/4)</th>
<th>(\Delta t = 1/8)</th>
<th>(\Delta t = 1/16)</th>
<th>(\Delta t = 1/32)</th>
<th>(\Delta t = 1/64)</th>
<th>(\Delta t = 1/128)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\Delta t)</td>
<td>(32 \times 32)</td>
<td>(64 \times 64)</td>
<td>(128 \times 128)</td>
<td>(256 \times 256)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\Delta t = 1/4)</td>
<td>1.83847</td>
<td>1.07313</td>
<td>0.983787</td>
<td>1.01402</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\Delta t = 1/8)</td>
<td>1.49828</td>
<td>1.51129</td>
<td>0.982416</td>
<td>1.00673</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\Delta t = 1/16)</td>
<td>1.26883</td>
<td>1.50625</td>
<td>0.989738</td>
<td>1.00323</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\Delta t = 1/32)</td>
<td>1.14148</td>
<td>1.33578</td>
<td>0.994583</td>
<td>1.00157</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\Delta t = 1/64)</td>
<td>1.0734</td>
<td>0.997854</td>
<td>0.997225</td>
<td>1.00077</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\Delta t = 1/128)</td>
<td>0.904951</td>
<td>0.867657</td>
<td>0.998597</td>
<td>1.00038</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conservation of Mass

<table>
<thead>
<tr>
<th>$h$</th>
<th>$32 \times 32$</th>
<th>$64 \times 64$</th>
<th>$128 \times 128$</th>
<th>$256 \times 256$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta t = 1/4$</td>
<td>$2.96827 \times 10^{-8}$</td>
<td>$0.0482585$</td>
<td>$0.0856009$</td>
<td>$0.180438$</td>
</tr>
<tr>
<td>$\Delta t = 1/8$</td>
<td>$1.08615 \times 10^{-8}$</td>
<td>$0.0308683$</td>
<td>$0.0435401$</td>
<td>$0.0936266$</td>
</tr>
<tr>
<td>$\Delta t = 1/16$</td>
<td>$4.62043 \times 10^{-9}$</td>
<td>$0.0168982$</td>
<td>$0.0221594$</td>
<td>$0.0459164$</td>
</tr>
<tr>
<td>$\Delta t = 1/32$</td>
<td>$2.12758 \times 10^{-9}$</td>
<td>$0.00875027$</td>
<td>$0.0111671$</td>
<td>$0.0232914$</td>
</tr>
<tr>
<td>$\Delta t = 1/64$</td>
<td>$1.02037 \times 10^{-9}$</td>
<td>$0.00444146$</td>
<td>$0.00560387$</td>
<td>$0.0117545$</td>
</tr>
<tr>
<td>$\Delta t = 1/128$</td>
<td>$4.99665 \times 10^{-10}$</td>
<td>$0.00223628$</td>
<td>$0.0028068$</td>
<td>$0.00590156$</td>
</tr>
</tbody>
</table>

Figure 4.22: The error estimates in this figure are for an implementation of the forward Euler method using the data of Problem 4.47. The floating point numbers are computed by averaging the absolute error estimates (4.93) corresponding to the two concentrations.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$U^0$</th>
<th>$U^1$</th>
<th>$U^2$</th>
<th>$U^3$</th>
<th>$U^4$</th>
<th>$U^5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta t = 1/4$</td>
<td>$0.485018$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta t = 1/8$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta t = 1/16$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta t = 1/32$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta t = 1/64$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td></td>
</tr>
<tr>
<td>$\Delta t = 1/128$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
<td>$0.485018$</td>
</tr>
</tbody>
</table>

Figure 4.23: Four Richardson extrapolation tables for the average concentration $(u + v)/2$ using the formula (4.92) applied to an implementation of the forward Euler method for the data of Problem 4.47 are listed. The spatial grid sizes for the tables (top to bottom) are $32 \times 32$, $64 \times 64$, $128 \times 128$, and $256 \times 256$. 

Problem 4.48) using a fixed spatial grid and the formula (4.91). The results of an experiment are given in Figure 4.18 and indicate that the forward Euler approximations are indeed order-one. Richardson absolute error estimates, computed using formula (4.93) and given in Figure 4.22 suggest that the absolute error for step sizes smaller than 1/64 and for grid sizes at least $64 \times 64$ is less than $10^{-3}$. The Richardson tables strongly suggest that the average over both concentrations is approximately 0.4530.

A conservative analysis suggests that we can be confident that the computed values are correct with an error of less than $10^{-2}$ for a spatial-grid size of $128 \times 128$ and a time step of about $1/16$. A similar accuracy seems to be achieved with larger time-steps (perhaps $1/2$, $1/4$, and $1/8$) together with one Richardson extrapolation. To reach time $T = 10$, a time-step of $1/128$ requires 1280 steps. On the other hand, the proposed Richardson extrapolation with time steps $1/2$ and $1/4$ requires 60 time steps. Clearly, the Richardson extrapolation offers a more efficient procedure.

For a less conservative and more efficient procedure, we can perform our computation with a $64 \times 64$ spatial grid and a time step of $1/4$ and remain confident that our computed results meet the $10^{-2}$ error tolerance.

Note that Richardson extrapolation is used only on the computed values at the final time in the experiments reported in this section. Clearly, the accuracy of the code can be improved by incorporating Richardson extrapolation at each time step (see Exercise 4.52).

While some appreciation of the meaning of the word “efficiency” for a computational algorithm can be gained by reading books, its true meaning is best understood by direct experience of the time required to perform computations. The choice between codes that produces a result in a few seconds or a few hours may not be of fundamental importance if the code is only run once to solve a textbook problem. On the other hand, an improvement of just a few percent might be essential for the success of a production code.

Since we are dealing with a system of equations in Problem 4.48, it is natural to consider both concentrations $u$ and $v$ in our error analysis. Thus, the corresponding averaged values are reported in this section. Taking into account our analysis and the computed values of $u$ and $v$ not listed but used to obtain the averaged concentrations, the result of our forward Euler approximations strongly suggest that (at time $T = 10$ and with an error of less than $10^{-2}$) the average $u$ concentration is 0.8084 and the average $v$ concentration is 0.09768.

Computational results for Problem 4.47 are listed in Figures 4.21, 4.22,
and 4.23. They suggest that the forward Euler scheme is operating as an order-one method for the grid sizes $128 \times 128$ and $256 \times 256$. The computed error is on the order of $10^{-2}$, but this is not assured. In this case, the average of the concentrations in Figure 4.23 do not seem to converge beyond the first decimal point. Taking the corresponding extrapolated values for the averaged concentrations $u$ and $v$, reasonable approximations for these averages at $T = 1024$ are 0.5 for $u$ and 0.1 for $v$. The accuracy of the computation seems to degrade over time. This is probably due to the accumulation of truncation error. An error of $10^{-3}$ per step might accumulate to an error of 1.0 after 1000 steps.

Problem 4.47 is not a typical textbook exercise; it is a genuine computational challenge.

Our reaction-diffusion model has an important complicating feature that has not yet been mentioned: The system evolves on two time-scales. The diffusion evolves on a slow time-scale of order $10^{-5}$ due to the diffusion coefficients; the reaction evolves on a fast time-scale of order one. How can we take this into account to obtain a more accurate numerical method?

At this point, we should have a high level of confidence in our computed averages at $T = 10$ and a low level of confidence in our computation at $T = 1024$. Certainly, more testing is warranted.

**Exercise 4.51.** Discuss which is better (taking into account efficiency and accuracy): Richardson extrapolation applied to three runs of forward Euler (with the data of Problem 4.47) for the step sizes $1/4$, $1/8$, and $1/16$ and a $256 \times 256$ spatial grid or one run with step size $1/128$ and a $128 \times 128$ spatial grid.

**Exercise 4.52.** Repeat the numerical work for Problems 4.47 and 4.48 with a forward Euler integration that incorporates Richardson extrapolation after the second time step: compute two time steps, extrapolate, use the extrapolated value to be the computed value, compute another step, etc. Compare the results with the results reported in this section.

**Exercise 4.53.** Repeat the numerical work for Problems 4.47 and 4.48 with one change: use Dirichlet boundary conditions such that the concentrations $u$ and $v$ on the boundary are held constant at $u = v = 0.5$. 
**The Crank-Nicolson Method**

The forward Euler method is easy to program, it runs fast, and with acceleration (extrapolation applied to increase accuracy) it can be used to produce reasonably efficient and accurate results. On the other hand, it has at least two weaknesses: the ratio of the time-step size and the spatial discretization size must be small to avoid numerical instability (see Exercise 4.45) and the method is first-order in time. Both of these deficiencies have been overcome with the development of difference methods that are unconditionally stable and second-order in time. One such method that is often used is due to John Crank and Phyllis Nicolson [20]; its derivation, order estimates, and implementation are discussed in this section.

Consider a solution $t \mapsto u(t)$ of the ODE $\dot{u} = f(u)$; it satisfies the equation

$$u'(t) = f(u(t)).$$

By integrating both sides of this identity with respect to the independent variable, we obtain the integral equation

$$u(t + \Delta t) - u(t) = \int_t^{t+\Delta t} f(u(s)) \, ds.$$

The left-hand rectangle rule approximation of the integral (that is, $((t + \Delta t) - t) \times f(u(t))$) yields the forward Euler method; the trapezoidal rule approximation

$$u(t + \Delta t) \approx u(t) + \frac{\Delta t}{2} (f(u(t + \Delta t)) + f(u(t)))$$

yields the trapezoidal method also called the *implicit improved Euler method*

$$U^{k+1} = U^k + \frac{\Delta t}{2} (f(U^{k+1}) + f(U^k)). \quad (4.94)$$

As might be expected, the trapezoidal method is more accurate than Euler’s method; in fact, it is a second-order method. To prove this fact, we simply assume that the solution and its approximation agree at time $t$ and estimate the norm of the difference between the true and approximate
solutions at \( t + \Delta t \) as follows:

\[
\begin{align*}
|u(t + \Delta t) - U^1| &= |u(t + \Delta t) - u(t) - \frac{\Delta t}{2}(f(u(t + \Delta t)) + f(u(t)))| \\
&= |u'(t)\Delta t + \frac{1}{2}u''(t)\Delta t^2 - \frac{\Delta t}{2}(2f(u(t)) + f'(u(t))u'(t)\Delta t)| \\
&\quad + O(\Delta t^3) \\
&= |f(u(t))\Delta t + \frac{1}{2}f'(u(t))u'(t)\Delta t^2 \\
&\quad - \frac{\Delta t}{2}(2f(u(t)) + f'(u(t))u'(t)\Delta t)| + O(\Delta t^3) \\
&= O(\Delta t^3).
\end{align*}
\]

Since the local truncation error is \( O(\Delta t^3) \), the method is second-order; that is

\[
\frac{u(t + \Delta t) - u(t)}{\Delta t} = -\frac{1}{2}(f(u(t + \Delta t)) + f(u(t))) + O(\Delta t^2)
\]

(cf. Exercise 4.55).

The trapezoidal method (4.94) is implicit. Its implementation requires, at each time step, the solution of the equation

\[
U^{k+1} - \frac{\Delta t}{2}f(U^{k+1}) = U^k + \frac{\Delta t}{2}f(U^k)
\]

for \( U^{k+1} \) to update the state variable. This equation is nonlinear whenever the ODE is nonlinear.

The premier method for approximating the solutions of nonlinear equations is Newton’s method, which is certainly one of the most important algorithms in analysis. Recall that if we wish to find a root \( a \) of a function \( g : \mathbb{R}^n \to \mathbb{R}^n \), we may suppose that if \( x \) is close to \( a \), then (by Taylor’s Theorem)

\[
g(a) \approx g(x) + Dg(x)(a - x) \approx 0.
\]

By solving for \( a \) in the formula \( g(x) + Dg(x)(a - x) = 0 \), we expect to find an approximation of this vector. The solution is

\[
a = x - Dg(x)^{-1}g(x),
\]

where \( a \) is an approximation of the root of \( g \). This observation is turned into Newton’s method: Choose an approximation \( x^0 \) of the expected root \( a \) and compute the sequence \( \{x^k\}_{k=0}^\infty \) according to the iteration process

\[
x^{k+1} = x^k - Dg(x^k)^{-1}g(x^k).
\]
Under the three assumptions (1) the starting value \( x^0 \) is sufficiently close to \( a \), (2) the function \( g \) is continuously differentiable, and (3) \( Dg(a) \) is invertible, the sequence of iterates defined by Newton’s method converges to the desired root \( a \). Moreover, the rate of convergence is quadratic; that is, for \( r = 2 \) and some positive number \( C \),

\[
|x^{k+1} - a| \leq C|x^k - a|^r. \tag{4.96}
\]

To approximate a solution \( x = a \) of the equation \( g(x) = 0 \) for \( g : \mathbb{R} \to \mathbb{R} \) by Newton’s method (the scalar case), choose some initial approximation \( x^0 \) of \( a \) and iterate using the formula

\[
x^{k+1} = x^k - \frac{g(x^k)}{g'(x^k)}. \tag{4.97}
\]

In the vector case, it is best to avoid computing the inverse of the matrix \( Dg(x^k) \). Instead, Newton’s method is implemented in an alternate form: Solve the linear system

\[
Dg(x^k)y = -g(x^k) \tag{4.98}
\]

for \( y \) and define \( x^{k+1} = y + x^k \).

For an arbitrary convergent sequence \( \{x^k\}_{k=1}^{\infty} \) with limit \( a \), we say that the order of convergence is \( r \) and the asymptotic error is \( C \) whenever we have the inequality (4.96). In other words, the error at index \( k + 1 \) (which we will denote by \( \text{err}(k + 1) \)) is approximately proportional (with a fixed constant \( C \) of proportionality) to the \( \text{err}(k)^r \); in particular, the error at index \( k + 1 \) is \( O(\text{err}(C)^r) \). The quadratic convergence of Newton’s method makes it a pillar of numerical (and theoretical) analysis (cf. [35]).

As a simple example of the convergence properties of Newton’s method, let us find the real root, \( x = 1 \), of the cubic polynomial \( g(x) = x^3 + x^2 - x - 1 \). The results of an implementation of Newton’s method with initial guess \( x = 2 \) are given in Figure 4.24. The asymptotic error computed with \( r = 2 \) as in display (4.96) seems to converge to \( C = 1 \). Thus, the experiment confirms that Newton’s method is quadratically convergent in this case.

For comparison, note that the root \( x = 1 \) can also be found as a fixed point of the function

\[
G(x) = x - \frac{1}{8}(x^3 + x^2 - x - 1). \tag{4.99}
\]
Conservation of Mass

<table>
<thead>
<tr>
<th>iterate</th>
<th>Newton</th>
<th>asymptotic error</th>
<th>composition</th>
<th>asymptotic error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.0</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.4</td>
<td></td>
<td>0.875</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.1</td>
<td>0.4</td>
<td>0.929932</td>
<td>0.125</td>
</tr>
<tr>
<td>3</td>
<td>1.0087</td>
<td>0.625</td>
<td>0.962554</td>
<td>0.560547</td>
</tr>
<tr>
<td>4</td>
<td>0.00007</td>
<td>0.869565</td>
<td>0.980582</td>
<td>0.534420</td>
</tr>
<tr>
<td>5</td>
<td>1.0</td>
<td>0.987124</td>
<td>0.990104</td>
<td>0.518548</td>
</tr>
<tr>
<td>6</td>
<td>1.0</td>
<td>0.999888</td>
<td>0.995003</td>
<td>0.509662</td>
</tr>
</tbody>
</table>

Figure 4.24: The iterates and asymptotic errors, using $|x^{k+1} - a|/|x^k - a|^r$ with $r = 2$ for Newton’s method and $r = 1$ for composition, to approximate the real root $x = 1$ of the polynomial $x^3 + x^2 - x - 1$.

<table>
<thead>
<tr>
<th>iterate</th>
<th>Newton</th>
<th>Aitken</th>
<th>Steffensen</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.1</td>
<td>0.92737</td>
<td>0.92737</td>
</tr>
<tr>
<td>3</td>
<td>1.0087</td>
<td>1.01026</td>
<td>1.00308</td>
</tr>
<tr>
<td>4</td>
<td>1.00007</td>
<td>1.00285</td>
<td>1.00000</td>
</tr>
<tr>
<td>5</td>
<td>1.00000</td>
<td>1.00076</td>
<td>1.00000</td>
</tr>
<tr>
<td>6</td>
<td>1.00000</td>
<td>1.00020</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

Figure 4.25: The iterates for Newton’s, Aitken’s and Steffensen’s methods applied to find the root $(x = 1)$ of the polynomial $x^3 + x^2 - x - 1$.

Starting from an initial guess $x = x_0$, the composition method (also called fixed point iteration) is to compute the compositional iterates of the function $G$; that is, to proceed inductively using the formula $x^{k+1} = G(x^k)$. In other words, we view $G$ as a dynamical system on the real line and seek the real root of the polynomial as a stable fixed point of this dynamical system. The factor $1/8$ is used to ensure that $G$ has a stable fixed point at $x = 1$ (see Exercise 4.64). The data in Figure 4.24 suggests that the iterates of $G$ are linearly convergent to the fixed point.

Newton’s method gives the correct root with an error of less than $10^{-2}$ after three iterates; the composition method requires five. To achieve an error less than $10^{-3}$, Newton’s method requires four iterates; the composition method requires nine. Let us also note that the computational cost of the two methods is comparable. It should be clear that Newton’s method is superior.

We have discussed acceleration by Richardson extrapolation of certain low-order methods with discretization errors. It is also possible to accelerate some linearly convergent sequences by a staple of numerical analysis called
Aitken’s $\Delta^2$-method. The idea is simple. For a linearly convergent sequence $\{x_k\}_{k=1}^{\infty}$ with limit $x^\infty$, we can expect that (for large $k$)

$$|x^{k+1} - x^\infty| \approx C|x^k - x^\infty|,$$

for some constant $C$. In addition, let us assume that the errors all have the same signs. By taking one more iterate, we have two approximations

$$x^{k+1} - x^\infty \approx C(x^k - x^\infty), \quad x^{k+2} - x^\infty \approx C(x^{k+1} - x^\infty).$$

By eliminating $C$, we find the relation

$$(x^{k+1} - x^\infty)^2 \approx (x^{k+2} - x^\infty)(x^k - x^\infty)$$

and solve for $x^\infty$ to obtain

$$x^\infty \approx \frac{x^k x^{k+2} - x^{k+1}}{x^{k+2} - 2x^{k+1} + x^k},$$

or equivalently,

$$x^\infty \approx x^k - \frac{(x^k - x^{k+1})^2}{x^{k+2} - 2x^{k+1} + x^k}.$$ 

Of course, this last relation is turned into a numerical method by replacing the unknown $x^\infty$ by the $k$th Aitken $\Delta^2$ approximation; that is,

$$\text{Aitken}(k) := \frac{x^k x^{k+2} - x^{k+1}}{x^{k+2} - 2x^{k+1} + x^k}. \quad (4.100)$$

Figure 4.25 lists the Aitken approximations obtained by applying formula (4.100) to the sequence of iterates of the function $G$ defined in display (4.99). The convergence is faster than the linear convergence of the simple iterates of $G$, but the new sequence does not converge quadratically (see Exercise 4.65). This deficiency can be remedied: Instead of applying Aitken’s method directly to the sequence of iterates, compute two iterations, apply formula (4.100) to $x_0$, $x_1$ and $x_2$ to obtain Aikan(0), use this value—which should be a better approximation of the fixed point than $x_3$—as the new $x_0$, and repeat this process. This algorithm is called Steffensen’s method. The Steffensen sequence for our test example is listed in Figure 4.25. For this example, it performs as well as Newton’s method (see Exercise 4.65).

The assumption that the errors $x^k - x^\infty$ in our sequence all have the same signs might not be satisfied; for example, the sequence given by $x^0 = 1$
and \( x^{k+1} = g(x^k) \) (where \( g(x) := -x/2 \)) converges to zero but alternates in sign. For a situation like this where the iterates alternate in sign, Steffensen’s method can be applied to every second iterate; or, in other words, it can be applied to the sequence generated by \( g^2(x) := g(g(x)) \).

Since we have Newton’s method, why are Aitken’s and Steffensen’s methods used? Answer: Aitken’s method applies to general linearly convergent sequences; hence, it can be applied in situations where no quadratically convergent method is known. Steffensen’s method is valuable in cases where the derivative of the function being iterated is difficult to obtain or expensive to evaluate (cf. Exercise 4.66). As might be expected, Richardson extrapolation, Aitken’s \( \Delta^2 \), and Steffensen’s are only a few of the important methods that have been developed to accelerate convergence of low-order methods. Of course, extrapolation methods, like all numerical methods, do not always work. The complete story is beyond the scope of this book (cf. [10]).

The Crank-Nicolson method is simply the trapezoidal method adapted to partial differential equations, where we view our parabolic PDE as an abstract evolution equation \( \dot{u} = f(u) \), where \( f \) is a differential operator. Thus, the Crank-Nicolson method for our basic reaction-diffusion PDE (4.53) is the scheme

\[
U_{i,j}^{k+1} = U_{i,j}^k + \frac{\Delta t}{2} \left[ \frac{\lambda}{\Delta x^2} (U_{i+1,j}^{k+1} - 2U_{i,j}^{k+1} + U_{i-1,j}^{k+1} + U_{i+1,j}^k - 2U_{i,j}^k + U_{i-1,j}^k) \\
+ \frac{\lambda}{\Delta y^2} (U_{i,j+1}^{k+1} - 2U_{i,j}^{k+1} + U_{i,j-1}^{k+1} + U_{i,j+1}^k - 2U_{i,j}^k + U_{i,j-1}^k) \\
+ (f(U_{i,j}^{k+1}, V_{i,j}^{k+1}) + f(U_{i,j}^k, V_{i,j}^k)) \right],
\]

\[
V_{i,j}^{k+1} = V_{i,j}^k + \frac{\Delta t}{2} \left[ \frac{\lambda}{\Delta y^2} (V_{i,j+1}^{k+1} - 2V_{i,j}^{k+1} + V_{i,j-1}^{k+1} + V_{i,j+1}^k - 2V_{i,j}^k + V_{i,j-1}^k) \\
+ \frac{\lambda}{\Delta x^2} (V_{i+1,j}^{k+1} - 2V_{i,j}^{k+1} + V_{i-1,j}^{k+1} + V_{i+1,j}^k - 2V_{i,j}^k + V_{i-1,j}^k) \\
+ (g(U_{i,j}^{k+1}, V_{i,j}^{k+1}) + g(U_{i,j}^k, V_{i,j}^k)) \right],
\]

(4.101)

where the symbols \( f \) and \( g \) are now redefined to denote the reaction terms in the PDE.

To implement the Crank-Nicolson scheme directly, we must solve a nonlinear system of equations to compute each update. This can be done using Newton’s method or Steffensen’s method (see Exercise 4.63).

Another possibility is to make the trapezoidal method into an explicit second-order method. Indeed, the idea is simple: Compute the implicit up-
date using Euler’s method. This results in the explicit second-order scheme

$$U^{k+1} = U^k + \frac{\Delta t}{2} (f(U^k + \Delta t f(U^k)) + f(U^k)),\quad (4.102)$$

which is often called the improved Euler method (see Exercise 4.67).

A third possibility is to make the method explicit (using the Euler approximation as in equation (4.102)) for the reaction terms and leave the update equations implicit for the diffusion terms. To complete each time step for this algorithm, we must solve a system of linear equations. This method is theoretically second-order with respect to both the space and time discretization sizes; in other words, the method is $O(\Delta t^2)$ and $O(\Delta x^2 + \Delta y^2)$.

To write a numerical code, we must incorporate a method for approximating the solution of a large system of linear equations. This subject is a branch of numerical analysis; a glimpse is provided here.

Let us begin by choosing a discretization of the spatial domain: the nodes are \((i-2)\Delta x, (j-2)\Delta y\) for \(\Delta x = L/m\) and \(\Delta y = L/n\) with \(i = 2, 3, \ldots, m+2\) and \(j = 2, 3, \ldots, n+2\). The corners of our spatial rectangle are labeled by the indices \((2, 2)\) (lower left), \((2, n + 2)\) (upper left), \((m + 2, n + 2)\) (upper right) and \((m + 2, 2)\) (lower right). Each of the \((m + 1)(n + 1)\)-nodes in the closed rectangle will correspond to two linear equations, one for \(U^k_{ij}\) and one for \(V^k_{ij}\). The resulting systems of linear equations can and will be solved separately.

The equations for the approximate state variables at the nodes must be ordered in some convenient manner; that is, we need a bijective function defined on the set of nodes whose range is the set of integers \(\{1, 2, 3, \ldots, (m + 1)\times(n + 1)\}\). We will use the bijection that corresponds to the ordering of the nodes on the spatial grid from top left to bottom right along rows from left to right. The required function, called “nodes”, is given by

$$\text{nodes}(i, j) = i - (m + 1)j + (n + 1) + mn + 2\quad (4.103)$$

(see Exercise 4.68).

Using our bijection, we can write our system of linear equations (for the unknown updates of the state variables at the nodes) in matrix form

$$AU^{k+1} = b,\quad (4.104)$$

where the \((m + 1)(n + 1)\times(m + 1)(n + 1)\)-matrix \(A\) and the \((m + 1)(n + 1)\)-vector \(b\) are defined using the Crank-Nicolson scheme. This procedure is
accomplished in three steps:

1. Use the Neumann boundary condition to set the values of $U$ and $V$ at the fictitious nodes that lie outside the spatial domain, which are needed to compute the discretized spatial second derivatives; for example, $U_{1,j}^k := U_{3,j}^k$ along the left-hand edge of the spatial domain.

2. Define the matrix $A$ using the form of the equations at each node; for example, the generic nodes($i,j$)-row of the matrix $A$ for the $U$ variables is given by the assignments

$$A(\text{nodes}(i,j), \text{nodes}(i,j)) := 1 + \lambda \left( \frac{\Delta t}{\Delta x^2} + \frac{\Delta t}{\Delta y^2} \right),$$

$$A(\text{nodes}(i,j), \text{nodes}(i+1,j)) := -\lambda \frac{\Delta t}{2\Delta x^2},$$

$$A(\text{nodes}(i,j), \text{nodes}(i-1,j)) := -\lambda \frac{\Delta t}{2\Delta x^2},$$

$$A(\text{nodes}(i,j), \text{nodes}(i,j-1)) := -\lambda \frac{\Delta t}{2\Delta y^2},$$

$$A(\text{nodes}(i,j), \text{nodes}(i,j+1)) := -\lambda \frac{\Delta t}{2\Delta y^2},$$

and all other components are set to zero in this row. Unfortunately, the matrix contains many nongeneric rows. Perhaps, an example will help to illustrate the extra complications due to the Neumann boundary conditions.

For the case $m = n = 2$ with $eta := \lambda \Delta t/(2\Delta x^2)$, the matrix $A$ is the $9 \times 9$-matrix

$$
\begin{pmatrix}
1 + 4\beta & -2\beta & 0 & -2\beta & 0 & 0 & 0 & 0 & 0 \\
-\beta & 1 + 4\beta & -\beta & 0 & -2\beta & 0 & 0 & 0 & 0 \\
0 & -\beta & 1 + 4\beta & -\beta & 0 & -2\beta & 0 & 0 & 0 \\
-\beta & 0 & -\beta & 1 + 4\beta & -\beta & 0 & -\beta & 0 & 0 \\
0 & -\beta & 0 & -\beta & 1 + 4\beta & -\beta & 0 & -\beta & 0 \\
0 & 0 & -\beta & 0 & -\beta & 1 + 4\beta & -\beta & 0 & -\beta \\
0 & 0 & 0 & -2\beta & 0 & -\beta & 1 + 4\beta & -\beta & 0 \\
0 & 0 & 0 & 0 & -2\beta & 0 & -\beta & 1 + 4\beta & -\beta \\
0 & 0 & 0 & 0 & 0 & -2\beta & 0 & -\beta & 1 + 4\beta
\end{pmatrix}.
$$

(4.106)

The factor two appears due to the boundary conditions. Note that the matrix structure is naturally separated into five strips from top to bottom: The top strip is the first row, strip two is rows 2 through $m+1$, strip three (the generic strip) is rows $m+2$ through $(m+1)(n+1) - (m+1)$, strip four is rows $(m+1)(n+1) - m$ to $(m+1)(n+1) - 1$, and strip five is the last row. The upper nonzero diagonals are the first and the $(m+1)$ upper diagonals;
the lower nonzero diagonals are the first and the \((m + 1)\) lower diagonals.

(3) The vector \(b\) is defined accordingly:

\[
\begin{align*}
b_{ij} &:= b(\text{nodes}(i, j)) := (1 - \lambda \left( \frac{\Delta t}{\Delta x^2} + \frac{\Delta t}{\Delta y^2} \right) )U^k_{i,j} \\
&\quad + \lambda \frac{\Delta t}{2\Delta x^2} (U^k_{i+1,j} + U^k_{i-1,j}) \\
&\quad + \lambda \frac{\Delta t}{2\Delta y^2} (U^k_{i,j+1} + U^k_{i,j-1}) \\
&\quad + \frac{\Delta t}{2} (f(\text{FEU}^k_{i,j}, \text{FEV}^k_{i,j}) + f(U^k_{i,j}, V^k_{i,j}))
\end{align*}
\]

where \((\text{FEU}^k_{i,j}, \text{FEV}^k_{i,j})\) is the forward Euler approximation of the full reaction-diffusion PDE with boundary conditions imposed of the approximate states \((U^{k+1}_{i,j}, V^{k+1}_{i,j})\) computed from the known approximation \((U^k_{i,j}, V^k_{i,j})\) and the same \(\Delta t\), \(\Delta x\), and \(\Delta y\) used for the Crank-Nicolson scheme.

Our choice of ordering determines the structure of the matrix \(A\) and vector \(b\); but, in a computer code to implement the Crank-Nicolson method, we should avoid defining the matrix \(A\)—there is no need to reserve storage for its zero elements. Since the matrices we are contemplating are large (for instance, a 129 \(\times\) 129 grid requires solving a system of 16641 equations), it is natural to take advantage of the structure of our (sparse) matrices, which have only five diagonals with nonzero components. We do not want the computer to waste our time computing values that are known to be zero.

At this juncture, we could enter the world of numerical linear algebra. This is a vast subject (see [12, 71]), which has certainly been influenced by the necessity of dealing with the matrix systems that arise in solving partial differential equations. This book is mainly about differential equations in applied mathematics, so we will not develop the theory in detail here. On the other hand, we need a viable method to approximate the solutions of the (large) linear systems that arise in implementations of implicit schemes such as the Crank-Nicolson algorithm. Hence, we must at least have a working knowledge of some fundamental results of this subject.

There are two basic methods for the numerical solution of linear systems: Gaussian elimination (which should be a familiar method from linear algebra) and fixed point iteration.

A viable implementation of Gaussian elimination for our (large and sparse) matrices demands adaptations that take into account the matrix structures.
In the special case of tridiagonal matrices (which arise for PDEs with one spatial dimension), there is an adaptation of Gaussian elimination that is simple to program, fast, and practical (see [12]). For banded matrices (such as the matrix in display (4.106)), there are similar but more complicated methods. We should also note that in applications to PDE we will often have to solve $AW = b$ repeatedly for the same system matrix $A$ but with different $b$. Efficient numerical schemes compute and store a factorization of $A$, for example a factorization $A = LU$ into a lower and an upper triangular matrix, so that $AW = b$ is solved by back substitution.

In keeping with our study of evolution equations (dynamical systems), we will employ an iterative method to solve linear systems, which is often called successive relaxation. In its simplest form, it is a generalization of the idea we used (recall display (4.99)) to approximate the solution of the nonlinear equation $g(x) = 0$ by iteration of the function $G(x) = x - \omega g(x)$ for a choice of $\omega$ that makes the desired solution a stable fixed point of the dynamical system defined by $G$. Indeed, we can hope to approximate the solution of our linear system $(Az = b)$ by iterating the linear transformation $T$ given by

$$Tz = z - \omega(Az - b), \quad (4.107)$$

where $\omega$ is a nonzero real number. There are at least three important questions: (1) What is a sufficient condition to guarantee that the process converges? (2) What is the best value to choose for the parameter $\omega$? (3) What is a good choice for $z$ to start the process?

Question (3) does not have a definite answer. Surely, the optimal choice of the initial point is the solution vector for the equation $Az = b$! Since this is the value we wish to find, we will have to settle for a less than optimal choice. One useful idea for our problem is to note that we plan to solve our system many times as we step along in time; hence, we can use current values of the concentrations at the nodes as the starting value to obtain the updated concentrations.

Questions (1) and (2) are answered by some numerical analysis. Recall that a fixed point of a map is asymptotically stable whenever all eigenvalues of the derivative of the map at the fixed point lie inside the open unit disk in the complex plane (see Exercise 4.64). Fortunately, the derivative of a linear transformation is itself and does not depend on the point at which it is evaluated. The derivative of our proposed function (4.107), which defines
our linear dynamical system, is
\[ DT = I - \omega A, \]
where \( I \) is the identity matrix. The eigenvalues of \( DT \) are in correspondence with the eigenvalues of \( A \); in fact, \( \mu \) is an eigenvalue of \( DT \) if and only if \( (1 - \mu)/\omega \) is an eigenvalue of \( A \). By the assignments given in display (4.105), all the elements on the main diagonal of \( A \) are equal and (in case \( \Delta x = \Delta y \)) of the form \( 1 + 2\alpha \) where \( \alpha = \lambda \Delta t/\Delta x^2 \). The sum of the absolute values of the off diagonal elements in each row of \( A \) (where each element is equal to \(-\alpha/2\)) is at most \( 2\alpha \). Also, note that \( A \) is a symmetric matrix; therefore, it has real eigenvalues. By Gerschgorin's theorem (see Appendix A.5), the eigenvalues of \( A \) lie in the interval \([1, 1 + 4\alpha]\), which consists of only positive real numbers.

Let \( \sigma(A) \) denote the set of eigenvalues (the spectrum) of \( A \) and note that \( \xi \in \sigma(A) \) corresponds to the eigenvalue \( \mu = 1 - \xi \omega \) in \( \sigma(DT) \). The maximum absolute value of the eigenvalues of \( T \) (its spectral radius \( \rho(T) \)) is given by
\[ \rho(T) = \min_{\omega \in \mathbb{R}} \max_{\xi \in \sigma(A)} |1 - \xi \omega|. \quad (4.108) \]
Under the assumption that \( \rho(T) < 1 \), the optimal value of \( \omega \) is the value at which the minimum occurs. By Exercise 4.72,
\[ \rho(T) = \frac{2\alpha}{1 + 2\alpha}, \quad \omega = \frac{1}{1 + 2\alpha}. \]
Thus, in this case, we have that \( \rho(T) < 1 \) as desired and the optimal choice \( \omega = 1/(1 + 2\alpha) \). The sequence of iterates of \( T \) will be linearly convergent to the solution of \( Az = b \) for every starting vector.

The implementation of our iterative method \( z^{\ell+1} = Tz^{\ell} \) should take advantage of the structure of the matrix \( A \). In particular, there is no reason to store this matrix. The update from \( z^{\ell} \) to \( z^{\ell+1} = Tz^{\ell} \) is given, for \( \gamma = 1, 2, 3, \ldots (m + 1)(n + 1) \), by
\[ z^{\ell+1}_\gamma = (1 - \omega(1 + \lambda(\Delta t/\Delta x^2 + \Delta t/\Delta y^2)))z^{\ell}_\gamma 
- \omega \lambda \frac{\Delta t}{2\Delta x^2} z^{\ell+1}_{\gamma+1} - \omega \lambda \frac{\Delta t}{2\Delta x^2} z^{\ell-1}_{\gamma-1} 
- \omega \lambda \frac{\Delta t}{2\Delta y^2} z^{\ell}_{\gamma+m+1} - \omega \lambda \frac{\Delta t}{2\Delta x^2} z^{\ell}_{\gamma-(m+1)}. \quad (4.109) \]
where \( z^\ell_p \) is set to zero if the index
\[
p \in \{ \gamma + 1, \gamma - 1, \gamma + m + 1, \gamma - (m + 1) \}
\]
is not in the range of \( \gamma \). We will have to encode a stopping procedure for our iteration process; for example, the procedure can be stopped as soon as an iterative update does not change by more than some prespecified tolerance. After gaining some experience by monitoring the performance of our code, we can perhaps simply iterate a fixed number of times to decrease the execution time. Our final update \( Z_\gamma \) is used to update the concentration \( U^{k+1}_{i,j} \) by
\[
U^{k+1}_{i,j} = Z_{\text{nodes}(i,j)}.
\]

The iteration process can be accelerated by an application of Steffensen’s method. For applications of this method to systems, there are at least two natural questions: (1) Can Steffensen’s method be applied componentwise? (2) How can we determine the signs of the errors to make sure the hypothesis for the convergence of Steffensen’s method is satisfied?

The idea underlying the answers to our questions can be illustrated by examining the iteration of linear transformation of the plane. Let us suppose that \( T \) is a linear transformation of the plane whose spectrum is inside the open unit disk. In this case, the sequence of vectors \( \{x^k\}_{k=0}^{\infty} \) defined by the iteration process \( x^{k+1} = Tx^k \) (for an arbitrary choice of \( x^0 \)) converges to zero. Under the assumption that \( T \) is symmetric and positive definite, the eigenvalues of \( T \) are positive real numbers \( \lambda_1 \leq \lambda_2 \) and there is a basis of \( \mathbb{R}^2 \) consisting of eigenvectors of \( T \). In this basis, \( T \) has the matrix representation
\[
\begin{pmatrix}
\lambda_1 & 0 \\
0 & \lambda_2
\end{pmatrix}
\]
The eigenspaces (that is, the subspaces spanned by the eigenvectors) in this basis correspond to the coordinate axes. Pick a vector in the plane, say the vector with components \( x_0 \) and \( y_0 \) and iterate the process. The \( k \)th iterate has components \( \lambda_1^k x_0 \) and \( \lambda_2^k y_0 \). Note that if \( \lambda_1 < \lambda_2 \), then the \( k \)th iterate becomes nearly parallel to the vertical coordinate axis. In other words, the dot product of the \( k \)th iterate and the usual basis vector \( e_1 \) is small compared with the dot product of the \( k \)th iterate and \( e_2 \). The first two vectors are nearly orthogonal. After a few iterations, the iterates align with the eigenspace corresponding to the largest eigenvalue. Thus, up to a small
error, the iterates converge monotonically to zero along this one-dimensional subspace. Also, all the components of the iterates converge linearly to zero and the errors with respect to each component have the same signs.

Exactly the same behavior occurs in general as long as the matrix that determines the iteration process is positive definite and symmetric. More generally, it suffices to have a positive real eigenvalue in the open unit disk that is larger than the absolute values of the real parts of all the other eigenvalues.

In practice, we should compute several iterates so that the expected alignment takes place, apply Steffensen’s method, compute several more iterations so that the expected alignment takes place, and continue in this manner until the iterations are no longer changing up to some preassigned tolerance or a maximum preset number of iterations is exceeded. It is possible to check internally that the alignment has occurred; for example, the computed differences $x_i^{k+1} - x_i^k$ for the components can be tested to see if they are all of the same sign and approximately the same magnitude (see Exercise 4.76).

While the basic iterative method we have just described is viable (especially when it is accelerated via Aitken extrapolation), it is seldom used because there are (usually) better methods due to Jacobi, Gauss, Seidel and others. The idea is to write the matrix $A$, which appears in the definition of $T$ in display (4.107), as a sum and move one of the summands to the left-hand side of the corresponding iteration scheme. With an appropriate choice of the splitting of $A$ into a sum of matrices, it is possible to obtain new iteration schemes that converge rapidly for the problem at hand.

Suppose $A$ is an $n \times n$-matrix and $\omega$ is a nonzero scalar. To solve for $z$ in the linear system

$$0 = -\omega(Az - b),$$

the usual splitting of $A$ is the sum of three $n \times n$-matrices: $D$, whose main diagonal is the main diagonal of $A$ and all its other components are zero; $L$, the lower triangular part of $A$ (that is, $L$ consists of the components of $A$ below its main diagonal and all of its other components are zero); and $U$, similarly, the upper triangular part of $A$. Using this notation, the linear system is recast in the form

$$0 = Dz - Dz - \omega(Dz + Lz + Uz - b),$$

which may be rearranged to obtain the matrix equation

$$(D + \omega L)z = ((1 - \omega)D - \omega U)z + \omega b.$$
If $D + \omega L$ is invertible, the unknown vector $z$ is a fixed point of the linear transformation
\[ \zeta \mapsto (D + \omega L)^{-1}(((1 - \omega)D - \omega U)\zeta + \omega b). \tag{4.110} \]

The idea is to obtain approximations of the unknown $z$ by iteration of this transformation starting from an initial guess. This method is usually called successive over relaxation (SOR). It is called an over relaxation method if $\omega > 1$, an under relaxation method if $\omega < 1$, and the Gauss-Seidel method if $\omega = 1$.

The formula (4.110) requires the matrix $S := D + \omega L$ to be invertible, which is the case if and only if every element on its main diagonal is not zero. In general, the inversion of matrices should be avoided in numerical computation because the number of operations required to invert a matrix increases rapidly with its size. In fact, the number of operations is on the order of $n^3$ for an $n \times n$-matrix. In contrast, the SOR method is practical because the required inversion of a lower triangular matrix is accomplished simply and efficiently by back substitution. By inspection of the matrix $S$, it follows immediately that the components of $z$ in the linear system
\[ Sz = v, \]
where $v$ is a given $n$ vector, are
\[ z_1 = \frac{1}{S_{ii}}v_1, \]
\[ z_i = \frac{1}{S_{ii}}(v_i - \omega \sum_{j=1}^{i-1} S_{ij}z_j). \tag{4.111} \]

In this scheme, $z_1$ is used to solve for $z_2$, $z_1$ and $z_2$ are used to solve for $z_3$, and so on; that is, the components of the solution are used to solve for subsequent values as soon as they are obtained. This is the underlying reason why the SOR methods are expected to converge more rapidly than the naive iteration method (4.107).

To implement SOR, choose an initial approximation $z^0$ and compute successive approximations $z^k$ by the scheme
\[ v^k := [(1 - \omega)D - \omega U]z^k + \omega b, \]
\[ (D + \omega L)z^{k+1} = v^k. \tag{4.112} \]
The components \( v_i^k \), \( i = 1, 2, 3, \ldots, n \), of \( v^k \) are
\[
v_i^k = (1 - \omega)a_{ii}z_i^k - \omega \sum_{j=i+1}^{n} a_{ij}z_j^k + \omega b_i, \tag{4.113}
\]
and the second step is completed using the back substitution formulas in display (4.111).

The SOR algorithm will converge if the eigenvalues of the matrix
\[
(D + \omega L)^{-1}((1 - \omega)D - \omega U)
\]
all lie inside the unit circle in the complex plane. The parameter \( \omega \) may be adjusted to make the eigenvalues as close to zero as possible. Of course, an analysis of the structure of the decomposition of matrix \( A = D + U + L \) or numerical computation of the required eigenvalues for a particular problem might lead to a theorem that would ensure convergence. The method does not converge for every \( n \times n \)-matrix \( A \).

The SOR method for the five-diagonal matrix \( A \) arising in the Crank-Nicolson scheme (4.104) can be programmed so that only the nonzero elements of the matrix are used (cf. equation (4.109)). For a positive-definite symmetric matrix, the spectrum of the matrix
\[
(D + \omega L)^{-1}([(1 - \omega)D - \omega U]
\]
lies in the open unit disk in the complex plane whenever \( 0 < \omega < 2 \); hence SOR converges whenever \( A \) is a positive-definite symmetric matrix and \( \omega \) is in this range (see [73]). In general, it is a difficult problem to determine the optimal value of \( \omega \).

Computational results for Problem 4.48 using the Crank-Nicolson algorithm, where the solutions of the corresponding matrix systems are approximated using Gauss-Seidel iteration (SOR with \( \omega = 1 \) and with the stopping criterion: successive iterations that differ in Euclidean norm by less than \( 10^{-6} \), are reported in Figures 4.26 and 4.27. The order estimates in Figure 4.26 suggest that this implementation of the algorithm is performing as expected at order-two in time. The computed value of the average over both concentrations is consistently (over the viable grid and step sizes) 0.453021. The averaged \( u \) concentration is consistently 0.808316 and the averaged \( v \) concentration is 0.0977254. These values agree with the values obtained using the forward Euler method reported on page 102. We can have a high
Conservation of Mass

<table>
<thead>
<tr>
<th>$h$</th>
<th>$32 \times 32$</th>
<th>$64 \times 64$</th>
<th>$128 \times 128$</th>
<th>$256 \times 256$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta t = 1/4$</td>
<td>1.93461</td>
<td>1.93451</td>
<td>1.93497</td>
<td>1.92266</td>
</tr>
<tr>
<td>$\Delta t = 1/8$</td>
<td>1.96882</td>
<td>1.96848</td>
<td>1.96688</td>
<td>2.00855</td>
</tr>
<tr>
<td>$\Delta t = 1/16$</td>
<td>1.98589</td>
<td>1.98510</td>
<td>1.98407</td>
<td>1.83655</td>
</tr>
<tr>
<td>$\Delta t = 1/32$</td>
<td>1.99551</td>
<td>1.99388</td>
<td>1.99383</td>
<td>2.07241</td>
</tr>
<tr>
<td>$\Delta t = 1/64$</td>
<td>2.0028</td>
<td>1.99951</td>
<td>1.99816</td>
<td>2.01279</td>
</tr>
</tbody>
</table>

Figure 4.26: The order estimates in this figure are for an implementation of the Crank-Nicolson method using the data of Problem 4.48 and the Gauss-Seidel iteration with the stopping criterion: successive iterations that differ in Euclidean norm by less than $10^{-6}$. The floating point numbers are computed for the average of the concentrations $(u + v)/2$ using formula (4.91).

<table>
<thead>
<tr>
<th>$h$</th>
<th>$32 \times 32$</th>
<th>$64 \times 64$</th>
<th>$128 \times 128$</th>
<th>$256 \times 256$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta t = 1/4$</td>
<td>0.00022356</td>
<td>0.00022363</td>
<td>0.00022355</td>
<td>0.00022420</td>
</tr>
<tr>
<td>$\Delta t = 1/8$</td>
<td>0.00005730</td>
<td>0.00005731</td>
<td>0.000057347</td>
<td>0.00005776</td>
</tr>
<tr>
<td>$\Delta t = 1/16$</td>
<td>0.00001450</td>
<td>0.00001451</td>
<td>0.000014569</td>
<td>0.00001497</td>
</tr>
<tr>
<td>$\Delta t = 1/32$</td>
<td>3.65 $\times 10^{-6}$</td>
<td>3.65 $\times 10^{-6}$</td>
<td>3.664 $\times 10^{-6}$</td>
<td>4.14 $\times 10^{-6}$</td>
</tr>
<tr>
<td>$\Delta t = 1/64$</td>
<td>9.1 $\times 10^{-7}$</td>
<td>9.1 $\times 10^{-7}$</td>
<td>9.2 $\times 10^{-7}$</td>
<td>9.8 $\times 10^{-7}$</td>
</tr>
<tr>
<td>$\Delta t = 1/128$</td>
<td>2.3 $\times 10^{-7}$</td>
<td>2.3 $\times 10^{-7}$</td>
<td>2.3 $\times 10^{-7}$</td>
<td>2.4 $\times 10^{-7}$</td>
</tr>
</tbody>
</table>

Figure 4.27: The error estimates in this figure are for an implementation of the Crank-Nicolson method using the data of Problem 4.48 and Gauss-Seidel iteration with the stopping criterion: successive iterations that differ in Euclidean norm by less than $10^{-6}$. The floating point numbers are computed by averaging the absolute error estimates (4.93) corresponding to the two concentrations.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$32 \times 32$</th>
<th>$64 \times 64$</th>
<th>$128 \times 128$</th>
<th>$256 \times 256$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta t = 1/4$</td>
<td>1.20052</td>
<td>2.97511</td>
<td>1.97269</td>
<td>2.06121</td>
</tr>
<tr>
<td>$\Delta t = 1/8$</td>
<td>1.42807</td>
<td>4.57200</td>
<td>2.17178</td>
<td>-0.457879</td>
</tr>
<tr>
<td>$\Delta t = 1/16$</td>
<td>1.65060</td>
<td>0.41385</td>
<td>1.94150</td>
<td>0.03870</td>
</tr>
<tr>
<td>$\Delta t = 1/32$</td>
<td>0.99641</td>
<td>1.79209</td>
<td>1.96255</td>
<td>3.38090</td>
</tr>
<tr>
<td>$\Delta t = 1/64$</td>
<td>4.77933</td>
<td>1.94427</td>
<td>1.98211</td>
<td>3.95762</td>
</tr>
<tr>
<td>$\Delta t = 1/128$</td>
<td>-4.55519</td>
<td>1.97100</td>
<td>1.99247</td>
<td>1.69587</td>
</tr>
</tbody>
</table>

Figure 4.28: The order estimates in this figure are for an implementation of the Crank-Nicolson method using the data of Problem 4.47 and Gauss-Seidel iteration with the stopping criterion: successive iterations that differ in Euclidean norm by less than $10^{-6}$. The floating point numbers are computed for the average of the concentrations $(u + v)/2$ using formula (4.91).
Figure 4.29: The error estimates in this figure are for an implementation of the Crank-Nicolson method using the data of Problem 4.47 and Gauss-Seidel iteration with the stopping criterion: successive iterations that differ in Euclidean norm by less than $10^{-6}$. The floating point numbers are computed by averaging the absolute error estimates (4.93) corresponding to the two concentrations.

<table>
<thead>
<tr>
<th>$h$</th>
<th>32 × 32</th>
<th>64 × 64</th>
<th>128 × 128</th>
<th>256 × 256</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta t = 1/4$</td>
<td>3.42×10^-9</td>
<td>1.91×10^-3</td>
<td>2.42×10^-3</td>
<td>2.08×10^-3</td>
</tr>
<tr>
<td>$\Delta t = 1/8$</td>
<td>1.70×10^-9</td>
<td>4.80×10^-4</td>
<td>5.63×10^-4</td>
<td>1.00×10^-3</td>
</tr>
<tr>
<td>$\Delta t = 1/16$</td>
<td>8.49×10^-10</td>
<td>1.20×10^-4</td>
<td>1.43×10^-4</td>
<td>8.21×10^-4</td>
</tr>
<tr>
<td>$\Delta t = 1/32$</td>
<td>4.24×10^-10</td>
<td>3.01×10^-5</td>
<td>3.60×10^-5</td>
<td>1.06×10^-4</td>
</tr>
<tr>
<td>$\Delta t = 1/64$</td>
<td>2.11×10^-10</td>
<td>7.54×10^-6</td>
<td>9.05×10^-6</td>
<td>3.00×10^-5</td>
</tr>
<tr>
<td>$\Delta t = 1/128$</td>
<td>1.07×10^-10</td>
<td>1.89×10^-6</td>
<td>2.97×10^-6</td>
<td>1.40×10^-5</td>
</tr>
</tbody>
</table>

Figure 4.30: The average concentrations in this figure are for an implementation of the Crank-Nicolson method using the data of Problem 4.47 and Gauss-Seidel iteration with the stopping criterion: successive iterations that differ in Euclidean norm by less than $10^{-6}$. The floating point numbers are the computed averages of the concentrations $(u + v)/2$.

<table>
<thead>
<tr>
<th>$h$</th>
<th>32 × 32</th>
<th>64 × 64</th>
<th>128 × 128</th>
<th>256 × 256</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta t = 1/4$</td>
<td>0.485018</td>
<td>0.463835</td>
<td>0.352109</td>
<td>0.335842</td>
</tr>
<tr>
<td>$\Delta t = 1/8$</td>
<td>0.485018</td>
<td>0.463835</td>
<td>0.352000</td>
<td>0.335848</td>
</tr>
<tr>
<td>$\Delta t = 1/16$</td>
<td>0.485018</td>
<td>0.463835</td>
<td>0.352098</td>
<td>0.335842</td>
</tr>
<tr>
<td>$\Delta t = 1/32$</td>
<td>0.485018</td>
<td>0.463835</td>
<td>0.352097</td>
<td>0.335842</td>
</tr>
<tr>
<td>$\Delta t = 1/64$</td>
<td>0.485018</td>
<td>0.463835</td>
<td>0.352097</td>
<td>0.335842</td>
</tr>
<tr>
<td>$\Delta t = 1/128$</td>
<td>0.485018</td>
<td>0.463835</td>
<td>0.352097</td>
<td>0.335842</td>
</tr>
</tbody>
</table>
Conservation of Mass

level of confidence that these values are within 1% of the exact corresponding values.

Computational results for Problem 4.47, using the Crank-Nicolson algorithm and Gauss-Seidel iteration, are reported in Figures 4.28 and 4.29. The order estimates in Figure 4.28 suggest that this implementation of the algorithm is performing at the expected order-two in time. The computed values of the averages over both concentrations converge, and these values are in good agreement with the values obtained by the forward Euler method. Perhaps the most trustworthy results are for the 128 × 128-grid with $\Delta t = 1/4$. The corresponding computed value for the averaged $u$ concentration is 0.531848 and the averaged $v$ concentration is 0.139837. These values also agree with the values obtained using the forward Euler method reported on page 102. The confidence level for these results extends to two or three decimal places at best.

How can we obtain more accurate results for Problem 4.47?

As already mentioned, we could use the trapezoidal together with Newton’s method to approximate the reaction terms.

We could use a direct method (some adapted form of Gaussian elimination) to solve the large linear systems that arise in the Crank-Nicolson method. This would avoid errors due to iteration. The difficulty is to incorporate the matrix structure into the method. Also, we could incorporate a more accurate approximation for the boundary conditions; for example, we could use an approximation based on the formula

$$ F'(a) = \frac{1}{12\Delta x} (F(a-2\Delta x) - 8F(a-\Delta x) + 8F(a+\Delta x) - F(a+2\Delta x)) + O(\Delta x^2). $$

(4.114)

Another idea is to use a more accurate spatial discretization. Instead of centered differences for the second derivatives, we could use higher-order approximations, for instance, the approximation

$$ F''(a) = -\frac{1}{12\Delta x^2} (F(a - 2\Delta x) - 16F(a - \Delta x) + 30F(a) - 16F(a + \Delta x) + F(a + 2\Delta x)) + O(\Delta x^4), $$

(4.115)

which is an alternate form of equation (4.87) (see Exercise 4.73).

Of course, there are many other methods to try. Can you devise a method that produces more accurate results for Problem 4.47?
Exercise 4.54. Suppose that \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is a smooth function. Prove that the initial value problem \( \dot{u} = f(u) \), \( u(0) = u_0 \) is equivalent to the integral equation

\[
  u(t) = u_0 + \int_0^t f(u(s)) \, ds.
\]

Exercise 4.55. Suppose that the local truncation error with discretization step size \( h \) for some method is \( O(h^{n+1}) \). Assume that the global error for the computation is the number of steps times the error per step. If the computation is carried out over an interval of length \( L \) show that the global error is proportional to \( Lh^n \); that is, the global error is \( O(h^n) \).

Exercise 4.56. Prove that Newton's method is quadratically convergent.

Exercise 4.57. (a) Derive the vector form of Newton's method and use it to write a numerical code formulated to approximate the rest points of the system of ODEs given by

\[
\begin{align*}
  \dot{U}_{ij} &= \epsilon(U_{i+1,j} - 2U_{i,j} + U_{i-1,j} + U_{i,j+1} - 2U_{i,j} + U_{i,j-1}) - U_{i,j}V_{i,j}^2 + F(1 - U_{ij}) , \\
  \dot{V}_{ij} &= \mu(V_{i+1,j} - 2V_{i,j} + V_{i-1,j} + V_{i,j+1} - 2V_{i,j} + V_{i,j-1}) + U_{i,j}V_{i,j}^2 - (F + \kappa)V_{ij},
\end{align*}
\]

for \( i = 1,2,3,\ldots m \) and \( j = 1,2,3,\ldots n \). Check that your implementation of Newton's method is quadratically convergent. (b) Find all the rest points in case \( \lambda = 2.0 \times 10^{-5} \), \( \mu = 10^{-5} \), \( \kappa = 0.05 \), \( F = 0.02725 \), \( m = 32 \), and \( n = 32 \). (c) Repeat part (b) for \( m = n = 64 \), 128, and 256. Check for convergence of the rest points to values independent of the choice of the grid size.

Exercise 4.58. (a) Derive the backward Euler method for \( \dot{x} = f(x) \):

\[
x^{k+1} = x^k + \Delta t f(x^{k+1}).
\]

(b) Show this method is first order. (c) Write a numerical code to implement backward Euler that incorporates Newton’s method to solve for the state variable update at each time step. (d) Repeat the computations for Problems 4.47 and 4.48 using the backward Euler method and compare results obtained with the forward Euler method.

Exercise 4.59. Recall Exercise 2.12 concerning the ODE

\[
\begin{align*}
  \dot{x} &= 1, \\
  \dot{y} &= axy,
\end{align*}
\]

where \( a \) is a parameter. The solution of the initial value problem with data \( x(0) = y(0) = -1 \) has the exact value \( (x(2), y(2)) = (1, -1) \) at \( t = 2 \) independent of \( a \). Apply forward Euler, backward Euler, improved Euler, and the trapezoidal method to this ODE, and determine the largest value of \( a \) for which each of your numerical codes returns the correct answer.
Exercise 4.60. An alternative to Newton’s method for vector functions is “Newton’s method one variable at a time (NOVAT).” Suppose we wish to apply Newton’s method to find a zero of the function \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n \) and this function is given in coordinates by

\[
y_1 = f_1(x_1, x_2, \ldots, x_n), \quad y_2 = f_2(x_1, x_2, \ldots, x_n), \quad \ldots, \quad y_n = f_n(x_1, x_2, \ldots, x_n).
\]

The alternative iteration scheme is

\[
x_{1}^{k+1} = x_{1}^{k} - \frac{f_{1}(x_{1}^{k}, x_{2}^{k}, \ldots, x_{n}^{k})}{\frac{\partial f_{1}}{\partial x_{1}}(x_{1}^{k}, x_{2}^{k}, \ldots, x_{n}^{k})},
\]

\[
x_{2}^{k+1} = x_{2}^{k} - \frac{f_{2}(x_{1}^{k}, x_{2}^{k}, \ldots, x_{n}^{k})}{\frac{\partial f_{2}}{\partial x_{2}}(x_{1}^{k}, x_{2}^{k}, \ldots, x_{n}^{k})},
\]

\[\vdots\]

\[
x_{n}^{k+1} = x_{n}^{k} - \frac{f_{n}(x_{1}^{k}, x_{2}^{k}, \ldots, x_{n}^{k})}{\frac{\partial f_{n}}{\partial x_{n}}(x_{1}^{k}, x_{2}^{k}, \ldots, x_{n}^{k})}.
\]

(a) Write a code to implement NOVAT and test its convergence rate on several examples where the root is known in advance. Does it converge quadratically? Report your results. (b) Code an ODE solver using the trapezoidal method together NOVAT to solve the implicit equation for the updated state variable. Apply your code to several ODE systems and report your result. (c) Show that this version of Newton’s method applied to a linear system \( Ax = b \) produces Gauss-Seidel iteration. (d) Construct an example where Newton’s method converges but NOVAT does not. (e) Is there an example where NOVAT converges but Newton’s method does not? (f) Determine the criterion for convergence of NOVAT.

Exercise 4.61. Test the code for Exercise 4.60 on the ODE of Exercise 2.12. Compare the results obtained with the improved Euler method.

Exercise 4.62. (a) Working definition: An ODE with solutions that converge (or diverge) exponentially fast to (or from) other solutions is called stiff. There is no universally accepted definition of stiffness. An alternative definition might be the following: An ODE is stiff if some numerical methods require small step sizes to make accurate approximations of some of its solutions. The simplest example of a stiff ODE is \( \dot{x} = \lambda x \), where \( \lambda \neq 0 \). If \( \lambda \) is negative, then all solutions approach the zero solution exponentially fast. Write the formulas and prove this fact. We may gain some insight on the performance of ODE solvers on stiff equations by applying them to this equation. For the analysis, it suffices to consider the solution of the ODE with initial condition \( x(0) = 1 \). Show that Euler’s method applied to
the test ODE gives the iteration scheme \( x^{k+1} = (1 + \lambda \Delta t) x^k \). In case \( \lambda < 0 \), we
know the solution decays. Show that the Euler approximation will decay if and only
if \( |1 + \lambda \Delta t| < 1 \) and that this puts a restriction on the step size \( \Delta t \): the step size
must be smaller than \( 2/|\lambda| \). Thus, if \( \lambda \) is large, the step size must be small to have
a chance of making an accurate approximation. Show that the same restriction
applies in case \( \lambda > 0 \). An alternative view is the restriction \( |\lambda \Delta t| < 2 \). The set of
points \( z \) in the complex plane (which must be considered for some more complicated
elements) with \( |z| < 2 \) is called the region of absolute stability. (b) Determine the
region of absolute stability for the improved Euler method. (c) Show that the region
of absolute stability for the trapezoidal method is the entire complex plane; that
is, there is no restriction on the step size. This result tells us that the trapezoidal
method is useful to approximate solutions of stiff ODEs. It should perform well with
larger step sizes than will be required for the one step methods we have discussed.
The trapezoidal method is a viable method for many stiff ODEs if the accuracy
requirements are not too restrictive. Higher order methods have been devised for
stiff equations. To learn more, consult any book on the numerical analysis of ODE
solvers. (d) Is the ODE in Exercise 2.12 stiff? Discuss your answer. (e) Consider
the ODE

\[
\begin{align*}
\dot{x} &= \lambda x - y - \lambda (x^2 + y^2)x, \\
\dot{y} &= x + \lambda y - \lambda (x^2 + y^2)y
\end{align*}
\]

with initial data \((x(0), y(0)) = (\xi, 0)\). Derive the exact solution

\[
\begin{align*}
x(t) &= \frac{\xi e^{\lambda t}}{\sqrt{1 + \xi^2(e^{2\lambda t} - 1)}} \cos t, \\
y(t) &= \frac{\xi e^{\lambda t}}{\sqrt{1 + \xi^2(e^{2\lambda t} - 1)}} \sin t.
\end{align*}
\]

Hint: Change to polar coordinates and recall Bernoulli’s differential equation. (f)
Show that the ODE of part (e) is stiff. (g) Test codes for improved Euler and
trapezoidal integration with \( \lambda \) small (perhaps \( \lambda = 1 \)) and big (perhaps \( \lambda = 10 \))
for \( \rho = 4 \) and integration until \( t = 6 \). Do both codes return the correct answer?
What is the maximum allowable step size for your codes to return an answer with
relative error less than 1%?

**Exercise 4.63.** Write a code for solving the Gray-Scott model using the Crank-Nicolson method and the trapezoidal method for the reaction terms. Use Newton’s method to solve the nonlinear equations for the updates. Apply the code to Problems 4.47 and 4.48.

**Exercise 4.64.** Consider the function \( G : \mathbb{R}^n \to \mathbb{R}^n \) as a dynamical system; that
is, the initial state \( x = x_0 \) in \( \mathbb{R}^n \) evolves according to the rule \( x^{k+1} = G(x^k) \). A
fixed point of \( G \) is a state \( x_0 \) such that \( G(x_0) = x_0 \). A fixed point is called stable
if for each \( \epsilon > 0 \) there is a \( \delta > 0 \) such that all (forward) iterates of every initial
state in the (open) ball of radius \( \delta \) centered at the fixed point are in the ball of radius \( \epsilon \). A fixed point is called asymptotically stable if it is stable and in addition if the sequence of iterates of every state starting in some ball of radius \( \delta \) centered at the fixed point converges to the fixed point. (a) Prove that a fixed point \( x_0 \) is asymptotically stable if \( n = 1 \) and derivative \( |G'(x_0)| < 1 \). (b) The function \( G(x) = x - \alpha(x^3 + x^2 - x - 1) \) has a fixed point at \( x = 1 \) for every \( \alpha \in \mathbb{R} \). For which \( \alpha \) is this fixed point asymptotically stable. (c) Prove that a fixed point \( x_0 \) is asymptotically stable if all the eigenvalues of the derivative \( DG(x_0) : \mathbb{R}^n \to \mathbb{R}^n \) are in the open unit disk in the complex plane.

Exercise 4.65. (a) Show that Aitken’s \( \Delta^2 \) method applied to the sequence of iterates of the function \( G \) defined in display (4.99) with starting value \( x_0 = 2 \) is not quadratically convergent. (b) Show that Steffensen’s method is quadratically convergent. (c) It is possible to prove that Aitken’s \( \Delta^2 \) method applied to a sequence of iterates of a function is superlinearly convergent; that is,

\[
\lim_{k \to \infty} \frac{|Aitken(k+1) - x^{\infty}|}{|Aitken(k) - x^{\infty}|} = 0.
\]

Give some (convincing) numerical evidence for this fact.

Exercise 4.66. Consider the following boundary value problem: Find a solution \( t \mapsto \phi(t) \) of the differential equation \( \ddot{x} + x - x^3 = 0 \) that satisfies the conditions \( \phi(0) = 0, \phi(2) = 0, \) and \( \phi(t) > 0 \) for \( 0 < t < 2 \). (a) Approximate the solution (is there only one?) by the shooting method; that is, let \( t \mapsto x(t, \xi) \) denote the solution of the ODE such that \( x(0, \xi) = 0 \) and \( \dot{x}(0, \xi) = \xi \). Use a numerical method to find a root of the equation \( x(2, \xi) = 0 \). Suggestion: Draw the phase portrait of the ODE and use it to determine a reasonable starting value \( \xi_0 \). Perform some numerical experiments to determine a value for \( \lambda \) so that the iterates of \( G(\xi) = \xi - \lambda x(2, \xi) \) converge. Accelerate the convergence using Steffensen’s method. (b) Shoot with Newton’s method. Hint: Find the required derivative by solving (numerically) a variational equation. (c) The suggested methods in parts (a) and (b) may not be the best. Don’t shoot! Approximate the solution of the BVP by viewing the ODE as a PDE with Dirichlet boundary conditions. (d) Discuss the accuracy and efficiency of various methods for solving the boundary value problem. Perhaps you can find a new method that is better than any known method.

Exercise 4.67. Show that the improved Euler method (4.102) is second order.

Exercise 4.68. (a) Prove that the nodes function defined in display (4.103) is bijective. (b) Show how to determine the formula for the nodes function by supposing the function is affine and solving a system of linear equations.
Exercise 4.69. Consider the \((x, y)\) data

\[(0, 0.09), (1/4, 0.12), (1/2, 0.16), (3/4, 0.20), (1, 0.26), (5/4, 0.32), (3/2, 0.39),
(7/4, 0.46), (2, 0.54), (9/4, 0.61), (5/2, 0.68), (11/4, 0.75), (3, 0.80), (13/4, 0.85),
(7/2, 0.88), (15/4, 0.91), (4, 0.93), (17/4, 0.95), (9/2, 0.96), (19/4, 0.97), (5, 0.98)\].

The theory underlying the experiment that produced the data implies that the data should be on the graph of a function of the form

\[f(x, a, b) = \frac{e^{a+bx}}{10 + e^{a+bx}},\]

where \(a\) and \(b\) are parameters. Use nonlinear regression to find the best constants \(a\) and \(b\). To do this, denote the data points with \((x_i, y_i)\), for \(i = 1, \ldots, 21\), and define a new function by

\[\Gamma(a, b) = \sum_{i=1}^{21} (y_i - f(x_i, a, b))^2.\]

The desired \((a, b)\) is the minimum of \(\Gamma\). To find this point, use calculus. Take the derivative and set it equal to zero. Use Newton’s method to solve the resulting nonlinear equations.

Exercise 4.70. Consider the \((x, y)\) data

\[(250, 0.0967), (500, 0.159), (750, 0.201), (1000, 0.233), (1250, 0.257), (1500, 0.276),
(1750, 0.291), (2000, 0.305), (2250, 0.315), (2500, 0.325), (2750, 0.333), (3000, 0.340),
(3250, 0.346), (3500, 0.351), (3750, 0.356), (4000, 0.360)\].

The theory (Michaelis-Menten kinetics) underlying the experiment that produced the data implies that the data should be on the graph of a function of the form

\[f(x, a, b) = \frac{ax}{b + x},\]

where \(a\) and \(b\) are parameters. Use nonlinear regression, as in Problem 4.69, to find the best constants \(a\) and \(b\).

Exercise 4.71. Is Newton’s method, which is quadratically convergent, a useful choice for solving linear systems of equations?
Exercise 4.72. Show that

\[ \rho(T) = \frac{2\alpha}{1 + 2\alpha}, \quad \omega = \frac{1}{1 + 2\alpha} \]

in the context of equation (4.108). Hint: The maximum occurs at an end point of the interval \([1 - \alpha, 1 + 3\alpha]\). Draw graphs of the functions \(\omega \mapsto |1 - (1 - \alpha)\omega|\) and \(\omega \mapsto |1 - (1 + 3\alpha)\omega|\).

Exercise 4.73. (a) Derive formula (4.114) directly from Taylor series approximations of \(F\). (b) Derive formula (4.115) directly from Taylor series approximations of \(F\).

Exercise 4.74. Write a code to approximate solutions of the Gray-Scott model using the Crank-Nicolson method with periodic boundary conditions, and compare the results with those obtained using the forward Euler method.

Exercise 4.75. [Method of Lines] The method of lines is a useful numerical method for some PDEs. For the case of reaction-diffusion equations, the idea is very simple: Discretize in space but not in time and treat the resulting equations (one for each spatial node) as a (perhaps large) system of ODEs (with time as the independent variable). The forward Euler method is exactly an example of the method of lines where Euler’s method is used to solve the system of ODEs. More sophisticated methods can be used to solve the ODEs. (a) Solve the BVP (4.80) by the method of lines using the improved Euler (explicit) method to solve the resulting systems of ODEs. Note: This is not the Crank-Nicolson method; the method of lines is explicit. (b) Solve Problem 4.48 by the method of lines using the improved Euler (explicit) method to solve the system of ODEs. The next part of this problem requires standard numerical methods for approximating the solutions of ODEs that are not explained in this book, but are found in standard textbooks (see, for example, [12]). (c) Solve the BVP (4.80) by the method of lines using 4th-order Runge-Kutta. (d) Solve the BVP (4.80) by the method of lines using Runge-Kutta-Fehlberg, a method that includes adaptive step size control. (e) Solve the BVP (4.80) by the method of lines using a 4th-order (Adams) predictor-corrector, a multi-step method that uses several previously computed steps. (f) Solve the BVP (4.80) by the method of lines using your favorite ODE integration method.

Exercise 4.76. Implement Steffensen’s method (incorporating the ideas discussed on Page 115) to solve five-diagonal \((m + 1)^2 \times (m + 1)^2\)-matrix systems \(Ax = b\) whose nonzero diagonals are the main diagonal of \(A\), the lower \((m+1)st\) diagonal, the lower first diagonal, the upper first diagonal, and the upper \((m+1)st\) diagonal. (a) Use your code to solve the system \(Ax = b\), where the components of the vector \(b\) all equal to one, all components on the main diagonal of \(A\) are equal to five
and all elements on the lower \((m+1)\)st diagonal, the lower first diagonal, the upper first diagonal, and the upper \((m+1)\)st diagonal are equal to one. (b) Incorporate Steffensen’s method into a Crank-Nicolson code and use it to solve the BVP \((4.80)\).

**Exercise 4.77.** [Research Project] The method of lines (Exercise 4.75) yields a window into pattern formation for the Gray-Scott model: The system of ODEs can be viewed as a perturbation problem where the unperturbed system is the reaction at each node and the discretized diffusion is the couplings between the nodes. The persistence and bifurcation from structures in the reaction equations might be used to account for the pattern formation. (a) Show that spatial pattern formation requires distinct diffusion coefficients. Hint: Use the system of PDEs and subtract the two equations. (b) Show that the coupled systems given by the method of lines have periodic solutions for sufficiently small diffusion coefficients.

**Exercise 4.78.** Recall the simplified Oregonator reaction model in Exercise 4.39. Add diffusion and explore spatial pattern formation for this model using the methods developed in this chapter.

**Exercise 4.79.** Suppose that the current dimensionless concentration of a substance \(u\) in space is \(u(w, t)\) at the dimensionless spatial position \(w = (x, y, z)\) and dimensionless time \(t\). Suppose the measured current concentration is approximately

\[
\frac{z(1 - e^{-(1-x^2-y^2)})}{(1 + z^2)^2}
\]

for \(w\) in the half cylinder \(\{(x,y,z) : x^2 + y^2 \leq 1, z \geq 0\}\) and zero everywhere else. The process that led to the current state is believed to be modeled by the three-dimensional diffusion equation \(u_t = \frac{1}{2} \Delta u\). Determine the dimensionless concentration at the point \(w = (1,0,2)\) at time \(t = -1\). Hint: Solutions of the diffusion equation do not make sense for negative time. To treat the problem, consider the process beginning at time \(t = -1\) and moving forward to time zero. Alternatively, reverse the direction of time in the dynamical equation; that is, work with the PDE \(u_t = -\frac{1}{2} \Delta u\). While it is possible to use the numerical methods developed in this chapter to obtain an approximation of the desired value, alternative methods are more efficient. The key result states that the solution of \(u_t = k \Delta u\) with bounded continuous initial data \(u(w, 0) = f(w)\) is

\[
u(w, t) = \int_{\mathbb{R}^3} K(w - \omega) f(\omega) \, d\omega,
\]

where \(K\), called the heat kernel (or diffusion kernel), is given in \(n\)-dimensional space by

\[
K(p) = \frac{1}{(4\pi kt)^{n/2}} e^{-|p|^2/(4kt)}.
\]
The desired concentration value may now be determined by approximating an integral over three-dimensional space. A numerical method that is widely used due to its simplicity is Monte Carlo integration. Suppose we desire the value of the integral \( \int_a^b g(x) \, dx \). We may use a random number generator to generate a finite sequence \( \{x_j\}_{j=1}^N \) of (uniformly distributed) random numbers in the interval \([a, b]\). An approximate value of the integral is given by

\[
\int_a^b g(x) \, dx \approx \frac{b-a}{N} \sum_{j=1}^N g(x_j).
\]

Try this method on a few test cases to convince yourself that it gives reasonable approximations and discuss why the method should work. The integral of the heat kernel (for all \( k \) and \( t \) such that \( kt > 0 \)) over all of space is exactly one. Check this by Monte Carlo integration. (A better way to apply Monte Carlo integration in this case is to notice the relation between the heat kernel and the normal distribution function from probability theory.) Use Monte Carlo integration to solve the original problem. Check the result by using an alternative method to approximate the solution of the heat equation.

**Exercise 4.80.** [Modeling Project] According to folklore, starting a fire in a fireplace warms the room in which the fireplace resides and cools the outlying rooms. The mechanism for this process is the movement of air required by the fire. Air from the outlying rooms used by the fire is replaced by cold outside air. Develop a mathematical model to describe this physical situation and determine to what extent this bit of folklore is true.

### 4.5 Excitable Media: Transport of Electrical Signals on Neurons

The most important discovery of the 20th Century in biophysics is the understanding of how nerves transmit information. The basic discovery relates the transport of ions of sodium and potassium (also sodium and calcium) across the outer membrane of a nerve cell to electrical signals that may propagate along the membrane after an appropriate stimulation. Alan Hodgkin and Andrew Huxley (working in the early 1950s) described the biological basis of the ion transport, created a mathematical model, and explained experimental data on electrical signals excited in squid giant axons; they were awarded the Nobel Prize in Physiology or Medicine in 1963. The original
Hodgkin-Huxley model is a system of four ODEs. Simplifications of the basic model, modifications for other excitable media (for example muscle cells) and spatial dependence have been extensively investigated. One of the most influential simplifications of the Hodgkin-Huxley model was introduced by Richard Fitzhugh, who also pioneered its mathematical and numerical analysis. An electric circuit analog for a similar model was constructed by Jin-Ichi Nagumo. Their two-state model, which is still widely used, describes the qualitative electrical behavior of stimulated nerve cells. We will investigate this model.

Excitable media in biology are far from being completely understood. Much contemporary work is focused on ion transport. Living membranes contain a variety of ion channels (across the membrane) that are selective to specific ions. The transport mechanisms and the switches that open and close ion channels are fundamental to the function of many biological processes. Also, networks of nerve cells and other excitable media are ubiquitous in biology. The study of such networks may lead to an understanding of how the brain works. Mathematics is playing an increasingly important role in this area of interdisciplinary research.

4.5.1 The Fitzhugh-Nagumo Model

The Fitzhugh-Nagumo model treats the nerve membrane as the electric circuit depicted in Figure 4.31. The differential equations for the important states, the voltage $V_D$ across the diode and the current through the inductor $I_L$ are obtained using standard circuit theory.

An electric circuit is a network of electrical components (capacitors, resistors, inductors, diodes, transistors, batteries, etc) connected by wires. The basic physics of electro-magnetism, which is encoded in Maxwell’s laws and the Lorentz force law, may be simplified to a few basic rules that are most often used to approximate the currents and voltages in an electrical circuit. A current $I$ is the rate of change of charge $Q$ with respect to time at a point in the circuit:

$$\frac{dQ}{dt} = I.$$

The magnitude of a current in a circuit is measured in amperes; the sign of the current determines the direction of the flow of electrons. For historical reasons—Benjamin Franklin being responsible—the direction of a current in a circuit is usually taken to be opposite to the direction of the flow of electrons.
An electron carries one unit of negative electric charge. A proton carries a unit of positive charge. Two charged particles attract or repel according to Coulomb’s law: the magnitude of the force on a charged particle due to a second charged particle is proportional to the product of the charges and inversely proportional to the square of the distance between them. The direction of the force is from the charged particle toward the second charge; that is, the force on charge $q_1$ at position $r_1$ due to the charge $q_2$ at position $r_2$ is

$$\text{Coulomb force on } q_1 = k \frac{q_1 q_2}{|r_1 - r_2|^3} (r_1 - r_2),$$

where $k$ is Coulomb’s constant. The force on $q_2$ is the negative of this given force. Charged particles with the same signs repel; charged particles with opposite signs attract. A static charge (no motion relative to an inertial coordinate system) produces an electric field

$$E = \nabla \phi$$

where

$$\phi = k \frac{q}{r}$$
is called the electric potential and \( r \) is the distance from the charge that produces the field. The field of a collection of charges is the sum of the (vector) fields produced by each charge in the collection. Voltage is a scalar, measured in volts, defined to be the potential difference at two positions in an electric field. A test charge (a small charge placed in an electric field) moves from positions of high potential to positions of lower potential. Note that, because we differentiate the potential to obtain the electric field, the electric potential and the voltage are defined up to the addition of a constant. The electric field produced by moving charges is not so simple. In this case, electricity and magnetism are entwined and an electromagnetic field is produced. A useful approximation of the fundamental field theory of electromagnetism can be used to analyze electric circuits. Much of the electromagnetic theory for currents and voltages in wires connected into circuits may be reduced to two rules due to Gustav Kirchhoff:

**Kirchhoff’s Current Rule** The sum of the currents at a node in an electric circuit (where two or more wires are joined) is zero.

**Kirchhoff’s Voltage Rule** The sum of the potential differences of (voltages across) circuit elements around every loop in a circuit is zero.

The current rule is essentially a statement of conservation of charge; the voltage rule is a statement of conservation of energy.

Figure 4.32: Schematic \( I-V \) relation for the diode in the Fitzhugh-Nagumo circuit. The axes are offset to show zero current through the diode for voltages across the device less than the threshold voltage at the position of the vertical axis.
We are now ready to consider the currents and voltages in the Fitzhugh-Nagumo circuit. While there are systematic ways to solve circuits, the process for simple circuits is an application of Kirchhoff’s laws for enough nodes and loops to obtain a closed system of differential equations. We begin by assigning positive directions to the currents. The assigned directions are not important as long as we remember that currents can be negative (which simply means that the flow in the circuit might be opposite to some of our choices). Circuit elements affect the current flowing through them. The effect of each element in the Fitzhugh-Nagumo circuit, a capacitor, a tunnel diode, a resistor, an inductor and a battery, is given mathematically by a function that relates the current $I$ flowing through the element and the voltage $V$ across it. These $I$-$V$ relations are the building blocks of circuits. The relations for the most basic elements are as follows:

**Resistor** The $I$-$V$ relation is Ohm’s law $V = IR$, where $R$ is a factor of proportionality called the resistance.

**Capacitor** A capacitor is a storage device, which at a basic level stores charge. The total charge $Q$ produces a field such that the voltage across the capacitor is $V = Q/C$, where the factor $C$ is called the capacitance. Using the definition of current as the rate of change of charge, the $I$-$V$ relation for a capacitor is

$$\frac{dV}{dt} = \frac{I}{C}.$$

**Inductor** An inductor is a coil of wire. Its effect is a result of Faraday’s law of induction, which states that a nonzero current produces a magnetic field. A magnetic field is produced in the space bounded by the coil that opposes changes in the current through the coil. The $I$-$V$ relation in a circuit is

$$V = L \frac{dI}{dt},$$

where the factor $L$ is called the inductance. (Note: Faraday’s law is symmetric: the motion of a magnet in the space bounded by the coil produces a current in the coil. This is the basic principle underlying the electric motor.)

**Battery** A battery is a storage device that produces a constant voltage across its terminals.
Diode Diodes are circuit elements that cause currents to flow in a specified direction. In modern circuits, diodes are semiconductor devices that come in several varieties. Generally their $I$-$V$ relations are non-linear. The diode in the Fitzhugh-Nagumo circuit has an $I$-$V$ relation in the form of a cubic, as depicted in Figure 4.32. The current through the diode is zero for voltages across the diode that are less than the threshold voltage indicated by the position of the vertical axis in the figure.

The top node of the Fitzhugh-Nagumo circuit, with the current directions as in Figure 4.31, is a junction of four wires that carry the currents $I_C$ of the capacitor, $I_D$ of the diode, the induced current $\hat{I}$ and the current through the circuit branch containing the resistor, inductor, and battery. By Kirchhoff’s current rule, the current through each of the latter elements is the same. Thus, we may choose the current through one of these elements, for example the current $I_L$ through the inductor to represent the current in the branch. With these choices, an application of the current rule yields the relation

$$\hat{I} + I_C = I_D + I_L.$$

Using the $I$-$V$ relation for a capacitor and the relation $I_D = F(V_D)$ for the diode (where $F$ is the function whose graph is depicted in Figure 4.32), we have the equation

$$\hat{I} + C \frac{dV_C}{dt} = F(V_D) + I_L.$$

Kirchhoff’s voltage rule applied to the leftmost loop in the circuit implies that $V_C = -V_D$; therefore,

$$\hat{I} - C \frac{dV_D}{dt} = F(V_D) + I_L. \tag{4.116}$$

The cubic shape of the $I$-$V$ curve for the diode may be approximated by shifting and lifting the cubic function given by $g(z) = Az^3 - Bz$; that is,

$$F(V_D) = g(V_D - \lambda) + \mu,$$

where $A$, $B$, $\lambda$ and $\mu$ are positive constants and $V_D$ exceeds the threshold voltage for the diode. In the analysis to follow the operating conditions will always be assumed to exceed the threshold value.
We may now replace $F(V_D)$ in equation (4.116) to obtain

$$\dot{I} - C\frac{dV_D}{dt} = g(V_D - \lambda) + \mu + I_L$$

or

$$C\frac{dV_D}{dt} = -g(V_D - \lambda) - (\mu + I_L) + \dot{I}. \quad (4.117)$$

The voltage rule applied to the rightmost loop in the circuit yields the relation

$$V_D = V_R + V_L + V_B.$$

The voltage through the resistor is $RI_R$ and the voltage through the inductor is $LdI_L/dt$. Moreover, the current through the resistor is the same as the current through the inductor. Thus we have the differential equation

$$L\frac{dI_L}{dt} = V_D - RI_L - V_B$$

or

$$L\frac{dI_L}{dt} = V_D - R(\mu + I_L) + R\mu - V_B. \quad (4.118)$$

With the change of variables $\hat{V} = V_D - \lambda$ and $\hat{W} = I_L + \mu$, we have the basic form of the circuit equations:

$$C\frac{d\hat{V}}{dt} = B\hat{V} - A\hat{V}^3 - \hat{W} + \hat{I},$$

$$L\frac{d\hat{W}}{dt} = \hat{V} - R\hat{W} + (\lambda + R\mu - V_B). \quad (4.119)$$

By scaling $\hat{V}$, $\hat{W}$ and time to $V$, $W$ and $s$ respectively, this system may be converted to the dimensionless form

$$\frac{dV}{ds} = V - \frac{1}{3}V^3 - W + I,$$

$$\frac{dW}{ds} = aV - bW + c, \quad (4.120)$$

where $a$, $b$, and $c$ are nonnegative constants (see Exercise 4.81).

The Hodgkin-Huxley circuit is supposed to model the electrical activity at a point on a neuron. The process of opening and closing ion channels is modeled by diffusion of the voltage (corresponding to the dimensionless state
Conservation of Mass

The spatial dependence is considered to be one dimensional with respect to a measure of distance $\hat{x}$ in the axial direction of the neuron. The spatial dependence is modeled as diffusion. As in heat conduction, the mathematical model is $\delta \frac{\partial^2 V}{\partial \hat{x}^2}$, where $\delta$ is the diffusivity. By adding this term to the right-hand side of the circuit model and scaling also the spatial variable, we obtain the dimensionless form of the Fitzhugh-Nagumo equations:

$$
\frac{\partial V}{\partial s} = \frac{\partial^2 V}{\partial x^2} + V - \frac{1}{3}V^3 - W + I,
$$

$$
\frac{\partial W}{\partial s} = aV - bW + c
$$

(4.121)

(see Exercise 4.82). The state variable $V$ is a representation of the voltage; it is also called the action or membrane potential, $W$ is called the recovery variable, and $I$ is the stimulus. In keeping with tradition, we may forget that the model is a dimensionless version of the original model and revert to the usual time variable; that is, we will consider this model equation in the form

$$
\frac{\partial V}{\partial t} = \frac{\partial^2 V}{\partial x^2} + V - \frac{1}{3}V^3 - W + I,
$$

$$
\frac{\partial W}{\partial t} = aV - bW + c
$$

(4.122)

where we may imagine that $t$ is the temporal independent variable and $x$ is the spatial variable.

The Fitzhugh-Nagumo model is not meant to be predictive; rather its purpose is to capture the main qualitative features of the electrical activity along a neuron. The most important prediction of the model (which agrees with experiments) is the existence of a threshold stimulus impulse that produces traveling voltage (and recovery) waves which propagate away from the spatial position of the stimulus. The membrane potential traveling wave is the mechanism responsible for carrying information along the neuron.

For definiteness, let us consider the case first explored by Fitzhugh [27], which for our version of the model has parameter values

$$
a = 0.08, \quad b = (0.08)(0.8), \quad c = (0.7)(0.08).
$$

(4.123)

In this case,

$$
a > b > 0 \text{ and } c > 0
$$

and all these parameters are much smaller than one.
Figure 4.33: The left panel depicts a computer generated phase portrait of the ODE system (4.124) for the parameter values (4.123) and stimulus $I = 0$ with the spiral sink at $(V, W) \approx (-1.199, -0.624)$ marked with a disk. The right panel is a $V$-profile for a trajectory starting half a unit below the rest point.

Some insight may be gained by considering the model without diffusion; that is, the model restricted to a single point in space:

$$\begin{align*}
\dot{V} &= V - \frac{1}{3}V^3 - W + I, \\
\dot{W} &= aV - bW + c,
\end{align*}$$

(4.124)

Figure 4.33 depicts the phase portrait of the ODE model (4.124) with stimulus $I = 0$. Note the presence of the vertical isocline $W = V - V^3/3$ (the curve in the phase plane where $\dot{V} = 0$), is clearly visible. Away from this curve, $\dot{V}$ is at least an order of magnitude larger than $\dot{W}$ because for our example regime the parameters $a$, $b$ and $c$ are at least an order of magnitude smaller than one, which we take as the coefficient of the $V$ component of the vector field. Thus, the flow moves rapidly toward the vertical isocline. Since this curve can be crossed in only one direction in the half planes bounded by the horizontal isocline $W = (aV + c)/b$, there is a layer near the horizontal isocline where the flow moves relatively slowly and stays near the isocline. Some solutions that start near and below the rest point, traverse a nearly closed loop and are eventually confined to a small neighborhood of the rest point, which of course they reach in their forward limit as the temporal parameter goes to infinity. An action potential $V$-profile for such a solution is also shown. Figure 4.34 depicts the phase portrait with stimulus $I = 1$. As
Figure 4.34: The left panel depicts a computer generated phase portrait of
the ODE system (4.124) for the parameter values (4.123) and stimulus \( I = 1 \)
with the source at \((V,W) \approx (0.409, 1.386)\) marked with a disk. The right
panel is a \( V \)-profile for the periodic trajectory surrounding the rest point.

Figure 4.35: The left panel depicts a computer generated phase portrait of
the ODE system (4.124) for the parameter values (4.123) and stimulus \( I = 2 \)
with the source at \((V,W) \approx (1.334, 2.543)\) marked with a disk. The right
panel is a \( V \)-profile for a trajectory starting half a unit above the rest point.
the stimulus is increased through a critical value \((I \approx 0.3313)\) at which the rest point changes stability (with a pair of complex conjugate eigenvalues passing through the imaginary axis in the complex plane) a stable limit cycle is produced via a Hopf bifurcation. The action potential \(V\) corresponding to the limit cycle has multiple spikes. A second critical value is reached as \(I\) is increased further and the limit cycle disappears in a Hopf bifurcation. Figure 4.35 depicts the phase portrait with stimulus \(I = 2\), which is larger than the second critical value (see Exercise 4.84), and a typical action potential profile. The pulses (spikes) in voltages observed in the figures may be interpreted as voltage changes induced by a stimulus at a site along a neuron.

Diffusion in the Fitzhugh-Nagumo equations models the spatial coupling among ion channels along a neuron. A stimulus above a threshold value but not too large produces an action potential that moves as time increases as in Figure 4.36. The voltage pulse caused by the stimulus splits into two parts that move left and right away from the spatial position of the stimulus. This wave of voltage (the action potential) is the information carried along a neuron.

As in the example depicted in Figure 4.37, a sufficiently strong stimulus spatially spread to one end of the nerve axon produces an action potential front that moves away from the stimulated region. The front in the figure seems to converge to a fixed profile; that is, it seems to converge to a traveling wave solution of the PDE.

Recall that a solution \((V,W)\) of the PDE (4.122) is called a traveling wave if there are two functions \(f\) and \(g\) and a number \(\gamma \neq 0\), called the wave speed, such that \(V(x,s) = f(x - \gamma t)\) and \(W(x,s) = g(x - \gamma t)\). The pair \((f,g)\) is called the wave profile. We will assume that \(\gamma > 0\) so that the solution is in the form of a wave traveling to the right along the \(x\)-axis with speed \(\gamma\). The expected physical wave is in the form of a pulse, which we define as a traveling wave solution such that \(V\) converges to the same fixed value as \(|x - \gamma t|\) grows to infinity and \(W\) behaves the same way except that the values at infinity may be different from the values of \(V\). For our biological application these values at infinity must be the rest state values \((V_0,W_0)\) given by the rest point of the ODE (4.124); these values correspond to the rest state of the nerve: the corresponding spatially constant solution of the PDE.

The natural setting for the existence of a traveling wave solution (which must be a solution of the PDE defined on the whole real line) is the PDE (4.122)
Figure 4.36: The upper left panel depicts an approximation of the action potential (lower graph) and the recovery variable (upper graph) at \( t = 0.84 \) for the Fitzhugh-Nagumo model (4.122) with the parameter values (4.123); stimulus \( I \) given by \( I(x, t) = 0.69 \) for \( |x| < 5 \) and \( 0 \leq t \leq 1 \) and by \( I(x, t) = 0 \) otherwise; initial \( V \) and \( W \) set equal to their rest state values (approximately \((-1.199, -0.624))\); and Dirichlet boundary conditions set at \( \pm 50 \) equal to these rest state values. The panels left to right and top to bottom depict the action potential and recovery variable at \( t = 0.84, 10.5, 16.8, 24.5, 35, \) and 56.
Figure 4.37: The upper left panel depicts an approximation of the action potential (lower graph) and the recovery variable (upper graph) at $t = 0.07$ for the Fitzhugh-Nagumo model (4.122) with parameter values (4.123); stimulus $I = 0$; initial $V$ given by $V = 1$ for $x < -40$, $(V_0 - 1.0)(x + 30)/10 + V_0$ for $-40 \leq x \leq -30$ and $V_0 \approx -1.199$ (the corresponding ODE steady state first component), and $V = V_0$ for $x > -30$; initial $W$ set to its ODE (4.124) rest state value (approximately $-0.624$); and Dirichlet boundary conditions set at $\pm 50$ equal to the ODE rest state values. The panels left to right and top to bottom depict the action potential and recovery variable at $t = 0.84, 2.8, 11.2, 14, 56,$ and $70$. 
Conservation of Mass

with zero stimulus and the boundary conditions \((V(x, t), W(x, t)) \to (V_0, W_0)\) as \(|t| \to \infty\).

To seek a traveling wave solution, substitute \(V(x, s) = f(x - \gamma t)\) and \(W(x, s) = g(x - \gamma t)\) into the PDE and try to solve for \(f\), \(g\), and \(\gamma\). This substitution results in a family of differential equations parameterized by \(\gamma\):

\[
\begin{align*}
  f'' + \gamma f' + f - \frac{1}{3} f^3 - g &= 0, \\
  \gamma g' + af - bg + c &= 0
\end{align*}
\]  

(4.125)

with independent variable \(s = x - ct\). To satisfy the boundary conditions, we desire a solution such that

\[
\lim_{|s| \to \infty} (f(s), g(s)) = (V_0, W_0)
\]  

(4.126)

where \((V_0, W_0)\) is a rest point of the ODE (4.124). The new ODE (4.125) is second-order; it is equivalent to the first-order system

\[
\begin{align*}
  \dot{u} &= v, \\
  \dot{v} &= -\gamma v - u + \frac{1}{3} u^3 + w, \\
  \dot{w} &= \gamma \left( -au + bw - c \right).
\end{align*}
\]  

(4.127)

The point \((V_0, W_0)\) corresponds to the steady state \((u, v, w) = (V_0, 0, W_0)\) for the new system (4.125). Thus, we seek a homoclinic orbit (a trajectory that is asymptotic in forward and backward time to the same steady state) of the three-dimensional system (4.127).

Proof of the existence of traveling wave solutions for the three-dimensional system (4.127) is much more complicated than for Fisher’s equation: the new system of ODEs is three dimensional instead of two dimensional and it is no longer true that there is a continuum of parameters \(\gamma\) for which a traveling wave exists—homoclinic orbits are (usually) unstable to perturbations; heteroclinic orbits from a saddle to a sink are stable. Another way to see the difficulty is simply by inspection of the system (4.127). There are three equations, but four unknowns: \(u\), \(v\), \(w\), and \(\gamma\). Thus, the problem is underdetermined. A solution method is discussed in the next section.

Exercise 4.81. Determine a scaling of the state variables and time that converts the system (4.119) to the dimensionless form (4.120)
Exercise 4.82. Show that an appropriate rescaling after the addition of a diffusion term to the first equation in system (4.119) yields the dimensionless equations (4.121).

Exercise 4.83. Reproduce Figures (4.33)–(4.35).

Exercise 4.84. (a) Determine the values of the stimulus $I$, for system (4.124) with parameter values (4.123), for which Hopf bifurcations occur. (b) Using numerical experiments, find the value of $I$ for which the area bounded by the limit cycle is maximal.

Exercise 4.85. Discuss via numerical experiments the dependence of the wave speed on initial data for the Fitzhugh-Nagumo model and compare with Fisher’s model.

### 4.5.2 Numerical Traveling Wave Profiles

A beautiful approach to the existence problem for traveling waves, which will be briefly discussed here, has resulted in a far-reaching theory for the existence and stability of several types of nonlinear waves in many different models (see [13] for a useful introduction).

The first idea is to seek the profile of a traveling wave solution as the steady state of a partial differential equation. We may hope that solutions of such a partial differential equation converge to the desired steady state as time goes to infinity. This is a powerful idea in many different contexts. For the dimensionless Fitzhugh-Nagumo model (4.122), the idea may be implemented by the family (depending on $\lambda$) of new variables

$$\xi := x - \lambda t, \quad t = \tau.$$ 

With the new functions

$$\hat{V}(\xi, \tau) := V(\xi + \lambda \tau, \tau), \quad \hat{W}(\xi, \tau) := W(\xi + \lambda \tau, \tau),$$

which are simply $V$ and $W$ in the new variables, the PDE (4.122) (without the stimulus because we are interested in long-term behavior only) is recast in the form

$$\frac{\partial \hat{V}}{\partial \tau} = \frac{\partial^2 \hat{V}}{\partial \xi^2} + \hat{V} - \frac{1}{3} \hat{V}^3 - \hat{W} + \lambda \frac{\partial \hat{V}}{\partial \xi},$$

$$\frac{\partial \hat{W}}{\partial \tau} = a \hat{V} - b \hat{W} + c + \lambda \frac{\partial \hat{W}}{\partial \xi}. \quad (4.128)$$
A steady-state solution of this system, which has appropriate limits as $\xi$ approaches $\pm\infty$ as in display (4.126), is a traveling wave solution of the dimensionless Fitzhugh-Nagumo model (4.122) with wave speed $\lambda$. For each $\lambda \neq 0$, system (4.128) is equivalent to system (4.122).

We have not yet accomplished a full implementation of the first idea because our system contains the free parameter $\lambda$. A remaining problem is to determine values of this parameter so that system (4.128) has a steady state corresponding to a traveling wave solution of the original FitzHugh-Nagumo model.

The next idea is natural once we view the evolution of system (4.128) in an infinite-dimensional space of functions and consider what we are trying to achieve. The parameter $\lambda$ has to be allowed to evolve with time so that it has a chance to approach the fixed value that would correspond to an appropriate steady state. The idea is to choose $\lambda(\tau)$, for each $\tau$, so that the square of the length of the vector field (the right-hand side of the PDE (4.128)) is minimized. By minimizing this length, the velocity vector would be as close to zero as possible at each $\tau$ as this temporal variable increases toward infinity. Thus, the imposition of this constraint would push solutions toward steady states where the vector field is zero.

There is a problem: our vector field depends on the spatial variable $\xi$. The appropriate notion of length should include this dependence. One possibility is to define the length (also called the norm) of a function $\phi$ to be

$$\|\phi\| := \left( \int_{-\infty}^{\infty} (\phi(\xi))^2 \, d\xi \right)^{1/2}.$$  

The function $\phi$ is called square integrable if the integral exists. Sometimes this norm is called the root mean square (RMS). Mathematicians call this norm the $L^2$-norm; and square integrable functions $L^2$-functions. While there are many other possible norms, a major advantage of the RMS norm is that it is defined by an inner product on the vector space of all square integrable (real valued) functions. Indeed, the inner product for two such functions $\phi$ and $\psi$ is defined by

$$\langle \phi, \psi \rangle := \int_{-\infty}^{\infty} \phi(\xi)\psi(\xi) \, d\xi$$

and the norm of $\phi$ is

$$\|\phi\| = \langle \phi, \phi \rangle^{1/2}.$$
Figure 4.38: The left panel depicts a numerical approximation of the graph of a steady state for $\hat{V}$ for the differential algebraic system (4.128) and (4.130) for the finite interval $(0, 130)$, the parameters in display (4.123), and zero Neumann boundary conditions. It is an approximation of a traveling wave profile. The right panel depicts a numerical approximation of $\lambda$ versus $\tau$ for the same system and initial data. The top panel depicts the initial $\hat{V}$ used for the numerical experiment. The initial $\dot{W}$ is set to the constant steady state value -0.62426.
We may extend the $L^2$-norm to vectors of functions in several ways. Here, we will define the norm for the 2-dimensional vector $(\phi, \psi)$ of $L^2$-functions by

$$\left(\|\phi\|^2 + \|\psi\|^2\right)^{1/2}.$$ 

Returning to our traveling wave problem, let us add the condition that the $L^2$-norm of the vector field given by the right-hand side of the PDE (4.128) (viewed as an infinite-dimensional ODE) is minimized at each $\tau$ with respect to $\lambda$. To take advantage of the $L^2$ inner product, we minimize the square of this norm; that is, we minimize

$$\langle \frac{\partial^2 \hat{V}}{\partial \xi^2} + \hat{V} - \frac{1}{3} \hat{V}^3 - \hat{W} + \lambda \frac{\partial \hat{V}}{\partial \xi}, \frac{\partial^2 \hat{V}}{\partial \xi^2} + \hat{V} - \frac{1}{3} \hat{V}^3 - \hat{W} + \lambda \frac{\partial \hat{V}}{\partial \xi} \rangle$$

$$+ \langle a \hat{V} - b \hat{W} + c + \lambda \frac{\partial \hat{W}}{\partial \xi}, a \hat{V} - b \hat{W} + c + \lambda \frac{\partial \hat{W}}{\partial \xi} \rangle$$

over $\lambda$. The appropriate condition is obtained in the usual manner: differentiate with respect to $\lambda$ and set the derivative equal to zero. The resulting equation is

$$0 = \langle \frac{\partial \hat{V}}{\partial \xi}, \frac{\partial^2 \hat{V}}{\partial \xi^2} + \hat{V} - \frac{1}{3} \hat{V}^3 - \hat{W} + \lambda \frac{\partial \hat{V}}{\partial \xi} \rangle + \langle \frac{\partial \hat{W}}{\partial \xi}, a \hat{V} - b \hat{W} + c + \lambda \frac{\partial \hat{W}}{\partial \xi} \rangle.$$ 

We must impose boundary conditions at $\pm \infty$ to obtain unique solutions of the PDE (4.128) for given initial $\hat{V}$ and $\hat{W}$ and fixed $\lambda$. We would like our state variables to remain in $L^2$. Thus, at the minimum we will need to assume the state variables vanish at infinity (in both directions). By a more sophisticated analysis, it is possible to see that we should also require first derivatives with respect to the spatial variable $\xi$ to vanish at infinity. We will simply impose this (Neumann) boundary condition.

Equation (4.129) may be simplified once the Neumann boundary conditions are imposed at infinity. For example,

$$\langle \frac{\partial \hat{V}}{\partial \xi}, \frac{\partial^2 \hat{V}}{\partial \xi^2} \rangle = \int_{-\infty}^{\infty} \frac{\partial \hat{U}}{\partial \xi} \frac{\partial^2 \hat{V}}{\partial \xi^2} dx$$

$$= \lim_{x \to \infty} \frac{1}{2} \left( \frac{\partial \hat{V}}{\partial \xi} \right)^2 - \lim_{x \to -\infty} \frac{1}{2} \left( \frac{\partial \hat{V}}{\partial \xi} \right)^2$$

$$= 0.$$
Conservation of Mass

For theoretical work, we would study the existence, uniqueness, and asymptotic behavior of solutions \( \tau \mapsto (\hat{V}(\tau), \hat{W}(\tau), \lambda(\tau)) \) of the system of equations consisting of the PDEs (4.128), the constraint (4.129), and the zero Neumann boundary conditions. This system is also suitable for numerical approximation.

For numerical experiments, the dynamical system must be restricted to a finite interval, for example \((0, L)\) for some \(L > 0\). Zero Neumann boundary conditions are imposed at the ends of this interval. The \(L^2\) inner products \(\langle \frac{\partial \hat{V}}{\partial \xi}, \frac{\partial^2 \hat{V}}{\partial \xi^2} \rangle\) and \(\langle \frac{\partial \hat{W}}{\partial \xi}, \frac{\partial^2 \hat{W}}{\partial \xi^2} \rangle\), where integration is now over the interval \((0, L)\), vanish just as they do on the infinite interval. Thus, the minimization constraint (with some rearrangement) is given by

\[
0 = \lambda \int_0^L \left( \frac{\partial \hat{V}}{\partial \xi} \right)^2 + \left( \frac{\partial \hat{W}}{\partial \xi} \right)^2 \, d\xi + \int_0^L (\hat{V} - \frac{1}{3} \hat{V}^3 - \hat{W}) \frac{\partial \hat{V}}{\partial \xi} \, d\xi + \int_0^L (a \hat{V} - b \hat{W} + c) \frac{\partial \hat{W}}{\partial \xi} \, d\xi.
\]  

(4.130)

Figure 4.38 shows some of the results of a numerical experiment to approximate the solution of the differential algebraic system (4.128) and (4.130) for the finite interval \((0, 130)\), the parameters in display (4.123), and zero Neumann boundary conditions. The initial configuration depicted in the top panel evolves to a steady state in variables \(\hat{V}, \hat{W},\) and \(\lambda\). An approximation of the graph of \(\hat{V}\) at this steady state is shown. It is an approximation to a traveling wave profile for the dimensionless Fitzhugh-Nagumo system, where the wave speed for the corresponding traveling wave solution is the steady state value of \(\lambda \approx 0.8168481086\) (compare [13, p. 101]). The initial configuration for this experiment was obtained by evolving forward to \(t = 13\) the dimensionless Fitzhugh-Nagumo solution depicted in Figure 4.37.

Unfortunately, the differential algebraic system (4.128) and (4.130) has multiple (stable) steady states. In addition to the desired steady state approximated as in Figure 4.38, the system has the constant steady state given by the solutions of the algebraic equations

\[
\hat{V} - \frac{1}{3} \hat{V}^3 - \hat{W} = 0, \quad a \hat{V} - b \hat{W} + c = 0,
\]

and \(\lambda\) arbitrary. For the parameters in display (4.123) \(\hat{V} \approx -1.19941\) and \(\hat{W} \approx -0.62426\). For the modification of the numerical experiment reported
in Figure 4.38 where the initial configuration is obtained in the same manner for \( t < 10 \), the system evolves to the constant steady state.

Some basic numerical methods for differential algebraic systems are suggested in the exercises.

**Exercise 4.86.** Consider differential algebraic systems of the form

\[
\dot{x} = f(x, y), \quad 0 = g(x, y)
\]

with initial conditions \( x(0) = \xi \) and \( y(0) = \eta \) restricted so that \( g(\xi, \eta) = 0 \). Euler’s method is easily modified to be a numerical method for such a system. Indeed, we may approximate the evolution of the system with the discrete time process

\[
x^{n+1} = x^n + \Delta t f(x^n, y^n), \quad 0 = g(x^n, y^{n+1})
\]

where \( x^0 = \xi \) and \( y^0 = \eta \). (a) Analyze the proposed numerical method and write a code to implement the method. (b) Solve the DAE

\[
\dot{x} = y, \quad 0 = x + 2y
\]

with the initial condition \( x(0) = -2 \) and \( y(0) = 1 \). (c) Test your code against the exact solution of the system in part (b). Is your code accurate to first-order in \( \Delta t \)? (d) Compare the accuracy of the alternative numerical algorithm

\[
x^{n+1} = x^n + \frac{\Delta t}{2} (f(x^n, y^n) + f(x^n, y^{n+1})), \quad 0 = g(x^n, y^{n+1})
\]

to Euler’s method.

**Exercise 4.87.** (a) Repeat the numerical experiment reported in Figure 4.38. Hint: The numerical method of Exercise 4.86 is adequate provided the discretization in space is second order. The integrals that appear in the algebraic constraint may be approximated with sufficient accuracy by the trapezoidal rule. (b) Modify the experiment, by choosing the initial voltage profile so that the system evolves to a constant steady state. Determine the asymptotic value of \( \lambda \) as \( \tau \to \infty \). (c) Repeat the experiment as in part (a) with the usual second-order discretization of the second-order derivatives with respect to \( \xi \), but only first-order accurate discretizations of the first-order spatial derivatives. Does the numerical simulation reach a steady state?

**Exercise 4.88.** [Research Project] The eigenvalue problem for an \( n \times n \)-matrix \( A \) is to solve the nonlinear equation \((A - \lambda I)v = 0\) for \( \lambda \) and \( v \neq 0 \). Eigenvector-eigenvalue pairs are steady states of the ODE \( \dot{v} = (A - \lambda I)v \). We must solve for
both $\lambda$ and $v$. An appropriate constraint is obtained by minimizing the square of the length of $\dot{v}$ over $\lambda$. This leads to the constraint $\langle v, (A - \lambda I)v \rangle = 0$ and suggests integrating the differential algebraic system

$$\dot{v} = (A - \lambda I)v, \quad 0 = \langle v, (A - \lambda I)v \rangle$$

with state variables $v$ and $\lambda$. If the solution approaches a steady state whose $v$-component does not vanish, then this steady state corresponds to an eigenvector-eigenvalue pair for $A$. Analyze this approach with mathematical analysis and numerical experiments. Notes: Start with $2 \times 2$ matrices. Eigenvalues of real matrices can be complex numbers. Perhaps, for this case, an augmentation of the dynamical system to include complex state variables would be helpful.

4.6 Biology or Chemistry on an Exposed Surface

A fundamental and ubiquitous situation in biology and chemistry involves a surface exposed to an environment containing a ligand that interacts with some receptor on the surface. A typical example might be a biological cell residing in a solution that contains a molecule that may bind with some receptor on the surface of the cell, or, a metal object exposed to water vapor in the air that might oxidize molecules on the metal’s surface. In addressing such problems, the diffusion of the ligand near the surface often plays an important role simply because the concentration of the ligand will not remain constant due to the binding and unbinding with the receptors on the surface. A class of problems of this type will be discussed in this section.

4.6.1 A Generic Model: Ligand, Receptor, and First-Order Kinetics

Let us imagine, for simplicity, that our surface is the boundary $\partial \Omega$ of some object $\Omega$. Perhaps $\Omega$ is a ball whose boundary is a sphere. This surface is assumed to be covered with receptors, whose concentration at the point $\omega$ on $\partial \Omega$ is $R(\omega)$, where we will assume $R : \partial \Omega \to \mathbb{R}$ is a smooth function. The dimension of $R(\omega)$ is some measure of quantity $Q$ per area (perhaps moles per square micrometer). In the exterior of object $\Omega$, there is some ligand whose concentration at the point $p$ (in space) is $A(p)$, with units $Q$ per volume. The ligand concentration is a function $A : \mathbb{R}^3 \setminus \Omega \to \mathbb{R}$ The
ligand is assumed to be in a medium through which it diffuses (for instance, a solute in water). And, most importantly, the ligand may bind or unbind with the receptors on the surface. In the bound state, the ligand is removed from the diffusive medium.

Given some initial concentration of the ligand (that is, the initial concentration function $A$), we wish to determine the concentrations of bound and unbound receptors as a function of time.

The diffusion process is modeled by the PDE

$$A_t = K \Delta A, \quad (4.131)$$

where $K$ is a constant, as long as we assume the medium containing the ligand is homogeneous and not moving. If the medium where moving, which will often be the case (for example, when the medium-ligand mixture is a fluid fed and drained in an experiment), we would have to take into account convection. The constant $K$, called the diffusivity, has units of area per time.

There will be some interaction between the ligand and the receptors causing some of the quantity of the ligand to be deposited on the surface. We may view this process, from the point of view of the bulk concentration of ligand, to be flux of the ligand into the boundary of $\Omega$. The usual model for this is Fick’s law (see, Equation (4.23)). It states that the velocity vector field $X$ of the ligand, in units of $QA$ per area per time, goes from high to low concentrations with magnitude the gradient of the concentration times the diffusivity; that is, $X = -K \nabla A$. The amount of ligand deposited per area at a point on the surface is taken to be the negative projection of $X$ on the outer unit normal $\eta$ on the surface (that is, $-X \cdot \eta$). The total flux through (or, in this case, on) the boundary of $\Omega$ is

$$\int_{\partial \Omega} K \nabla u \cdot \eta \, dS.$$

While the model for deposition of the ligand is not due to fundamental physical laws—we have used Fick’s (constitutive) law, for most problems of the type discussed here this part of the model is universally employed. Clearly, $K \nabla A \cdot \eta$ is a quantity with the correct units: quantity per area per time. This quantity must be equated to some function $h$ (with the same units) describing the binding of ligands and receptors on the surface. In other words, our mathematical model will have a boundary condition

$$-K \nabla A \cdot \eta = h \quad (4.132)$$
on $\partial \Omega$, where the left-hand side is the amount of ligand deposited per area per time on the surface.

The biology (or chemistry) is in the construction of the function $h$. To have the potential for an impact in science, the applied mathematician should work in collaboration with a scientist who understands the biology (or chemistry) to help formulate a viable $h$.

For mathematical reasons—to ensure there is a unique solution—we will need a second boundary condition for our diffusion process on some surface that encloses the diffusive medium and the object. A natural and simple choice is to enclose the object and the medium in a volume $B$ and impose the boundary condition that the ligand concentration on the surface of $B$ is the presumed constant (bulk) concentration $A_b$ of the ligand in the medium. Mathematically, this is the Dirichlet boundary condition

$$A = A_b$$

on $\partial B$. Alternatively, if the surface is treated as a physical enclosure, we may impose a flux condition on this enclosure

$$K \nabla A \cdot N = J,$$

where $N$ is the inner normal on $\partial B$ and $J$ is a function defined on this set, which might depend on $A$.

The possibilities for constructing the surface interaction $h$ are endless. We will discuss a simple example here.

We have three species: the ligand $A$, the receptor $R$, and the bound ligand-receptor $R_b$ with interaction

$$A + R \xrightleftharpoons{\ell}{k} R_b,$$

which goes in both directions as indicated by the arrows. The symbols $k$ and $\ell$ denote the forward and backward reaction rates. In general, there are many possibilities for the corresponding rate equations that must be determined from the biology or chemistry being modeled. A simple (and often used) assumption is first-order kinetics. Using $Q_A$, $Q_R$, and $Q_{R_b}$ to denote the quantities of the corresponding substances all measured in the same unit $Q$. 

Conservation of Mass

first-order kinetics leads to the rate equations

\[
\frac{dQ_A}{dt} = kQ_{R_b} - \ell Q_A Q_R, \quad (4.135)
\]

\[
\frac{dQ_R}{dt} = kQ_{R_b} - \ell Q_A Q_R, \quad (4.136)
\]

\[
\frac{dQ_{R_b}}{dt} = \ell Q_A Q_R - kQ_{R_b}, \quad (4.137)
\]

where \( k \) is measured in units of inverse time (1/time) and \( \ell \) in inverse time per \( Q \).

To have rate equations for concentrations, we should note that the concentration of \( A \) (which is diffusing in our medium) is measured in quantity \( Q \) per volume while the concentrations of the receptor and the bound receptor species are measured in \( Q \) per area on the surface. For the surface interaction, \( A \) should be measured in quantity per area. What is the surface concentration if we know the volume concentration? The answer is not obvious. A reasonable answer is that the (numerical) value of the two concentrations is the same. To argue this alternative, note that concentration is a function of position. To recover the total quantity over some region of space, we simply integrate the concentration over the region with respect to volume. Likewise, to recover the total quantity over some surface we may integrate the concentration function restricted to the surface with respect to area on the surface. This is an abstraction: a surface has no thickness; thus, it cannot contain a total quantity of some substance. The quantity on the surface might be defined this way (compare Exercise 4.89), but the units do not work properly: the concentration function integrated over the surface simply does not have the units of quantity \( Q \) per area. Using this approach, we would simply ignore the difference between surface concentration and volume concentration. A viable alternative is to assume that the surface concentration is some function \( F \) of the concentration in the medium. In this case there is no confusion with the units; but, the choice of this function adds an additional complication to the model. A possible choice for the function \( F \) is to multiply the volume concentration times the thickness of a defined boundary layer. This produces a quantity per area as desired, but leaves open the problem of choosing the layer thickness. Continuum modeling suggests seeking a definition of \( F \) that eliminates the choice of layer thickness by taking a limit. But, the obvious continuum limit as the layer thickness decreases to zero is zero surface concentration. This is physically correct, but
Conservation of Mass

zero surface concentration does not allow interaction at the surface. The best
approach is to use a function $F$ that must be specified for each application.
The former methodology can be recovered by simply taking $F$ to be given by
$F(A) = cA$, where $c$ is one unit of length so that $F(A)$ is measured in units
of $Q$ per area.

The rate equations (4.135) are recast as rate equations for the surface
concentrations by dividing both sides of each equation by the surface area $S$
and multiplying and dividing $R$ by $S$. Thus, we obtain the system

\[
\frac{dF(A)}{dt} = kRb - \ell SF(A)R,
\]
\[
\frac{dR}{dt} = kR_b - \ell SF(A)R,
\]
\[
\frac{dR_b}{dt} = \ell SF(A)R - kRb.
\]

The sum of the concentrations of the receptor and the bound receptor
species is a given function $C$ of position on the surface and time; that is,

\[R(t, \omega) + R_b(t, \omega) = C(t, \omega).\]

A typical choice is $C = 1$ where the receptors reside and $C = 1$ elsewhere.
We may eliminate the equation for the receptor species to obtain the rate
equations

\[
\frac{dF(A)}{dt} = kRb - \ell SF(A)(C - Rb),
\]
\[
\frac{dR_b}{dt} = \ell SF(A)(C - Rb) - kRb. \tag{4.138}
\]

By balancing the rate of change of $A$ in the key boundary condition (4.132)
and using the Dirichlet boundary condition on the outer boundary, the full
(nonlinear) model is

\[
A_t = K \Delta A, \quad \text{in } \mathcal{B},
\]
\[
-K \nabla A \cdot \eta = kRb - \ell SF(A)(C - Rb), \quad \text{on } \partial \Omega,
\]
\[
R_b = \ell SF(A)(C - Rb) - kRb, \quad \text{on } \partial \Omega, \tag{4.139}
\]
\[
A = A_b, \quad \text{on } \partial \mathcal{B},
\]

where the function $F$ converts space concentration to surface concentration.
The required initial data is the initial spatial concentration $A_0$ of $A$ on $\mathcal{B}$
and the concentration $R_{b0}$ of $R_b$ on $\partial \Omega$. 
Our model has been constructed to determine the effect of diffusion on the ligand-receptor interaction. This interaction should be compared with the case where diffusion is not important; for example, the medium may be well-stirred. This regime is modeled by taking the diffusion constant $K$ to be infinite. Mathematically, we may divide by $K$ and consider the limiting system as $K$ increases without bound. The solution of the resulting subsystem

$$\Delta A = 0 \text{ in } B, \nabla A \cdot \eta = 0 \text{ on } \partial \Omega, \text{ and } A = A_b \text{ on } \partial B \quad (4.140)$$

has the constant solution $A(t,x,y,z) = A_b$ (see Exercise 4.91). Thus, the dynamics is governed by the family of linear initial value problems

$$R_b(t) = \ell SF(A_b)(C - R_b) - k R_b, \quad R_b(0,\sigma) = R_b(\sigma),$$

where the initial condition depends on the initial concentration of $R_b$ at the surface position $\sigma$. Thus, the bound receptor concentration at $\sigma$ at time $t$ is

$$R_b(t,\sigma) = R_b(\sigma)e^{-(\ell SF(A_b) + k)t} + \frac{\ell SF(A_b)C}{\ell SF(A_b)C + k}(1 - e^{-(\ell SF(A_b) + k)t}). \quad (4.141)$$

We have constructed a continuous model of a phenomenon that may be considered to be discrete: a receptor is either bound or unbound. On the other hand, measurement may approximate the bound receptor concentration, not the state of individual receptors. It is a matter of interpretation. As time goes forward, the bound receptor concentration (at every point) approaches the constant value

$$R_{b\infty} := \frac{\ell SF(A_b)C}{\ell SF(A_b)C + k}. \quad (4.142)$$

The result is controlled by $k$, the rate at which the ligand-receptor pairs dissociate. For small $k$ the concentration of bound receptors is high; for large $k$ the concentration of bound receptors is low (see Exercise 4.93).

In case $C$, $A_b$, and the parameters are not time dependent, in steady state (where the solution of the model system (4.139) does not change with time), we obtain the same boundary value problem as for the high diffusion limit (4.140). Thus, for the steady state of the PDE model, we also have the bound receptor concentration (4.142) (see Exercise 4.92).
Exercise 4.89. Recall the relation between the volume element and the area element. In three-dimensional space the volume form $\sigma$ is the determinant of $3 \times 3$ matrices viewed as a function of the (vector) columns of these matrices. In other words, $\sigma$ assigns to three vectors in space (viewed as tangent vectors at a point) the determinant of the matrix formed by taking these vectors as the columns of the matrix. The area element for a surface in three-dimensional space is obtained by the same function $\sigma$ by restriction to the surface and respect to the unit normal on the surface: the first two columns of the corresponding matrix are tangent vectors to the surface and the third column is the unit normal vector at the point where the area element is computed. Symbolically, for a surface $S$ with normal $\eta$ in the volume $V$, we have

$$\text{vol}(V) = \int_V \sigma, \quad \text{area}(S) = \int_S i_\eta \sigma,$$

where $i_\eta \sigma$ is the area element (also called the interior product of $\sigma$ by $\eta$) just described. In case there is a concentration function $A$, we have

$$\text{quantity}(V) = \int_V A \sigma, \quad \text{quantity}(S) = \int_S A i_\eta \sigma.$$

(a) Make this description rigorous. (b) Compute the total quantity of some substance on the unit square in the $(x, y)$-plane if its concentration in space is given by $A(x, y, z) = y + \sin(x) + z$. (c) Compute the total quantity of some substance on the part of the plane with equation $2x + y + z = 0$ that lies inside the unit sphere if its concentration in space is given by $A(x, y, z) = 3x + 2y - z$.

Exercise 4.90. Show that the IVP (4.141) has the stated solution, determine the concentrations of the unbound receptors, and discuss your result.

Exercise 4.91. Show that the BVP (4.140) has the stated solution and prove that this solution is unique.

Exercise 4.92. In case $C$, $A_b$, and the parameters are not time dependent, determine the steady state solution of the model system (4.139) and show that the bound receptor concentration is given by formula (4.142).

Exercise 4.93. Show that the IVP (4.141) has the stated solution, determine the concentrations of the unbound receptors, and discuss your result.

Exercise 4.94. The model system (4.139) is written for the Dirichlet boundary condition at the outer boundary. Rewrite the model for the no flux Neumann boundary condition on the outer boundary and discuss the system in steady state. Determine the steady state solutions and compare your results with those obtained for the Dirichlet boundary condition.
4.6.2 Model With Rectangular Geometry

The geometry of Ω and $B$ in the model (4.139) plays an important role in the analysis of the model and the numerical approximation of solutions. For simplicity, let us consider Ω to be the square bottom with side length $\omega$ of a rectangular box $B$ with height $b$. For definiteness, we may assume that Ω is in the horizontal coordinate plane in its first quadrant with sides along the $x$ and $y$ coordinate axes, and the box lies above this plane with a vertical edge along the $z$ coordinate axis.

Let us make our dynamical model dimensionless with the preliminary scaling

$$t = as, \quad x = b\xi, \quad y = b\nu, \quad z = b\zeta, \quad A = \alpha L, \quad R_b = \beta P$$

where $a$ has dimension time, $b$ has dimension length, $\alpha$ has dimension $Q/V$, and $\beta$ has dimension $Q/S$. This change of variables produces the new model equations

$$L_s = \lambda \Delta L,$$
$$\nabla L \cdot \eta = G(L)(\gamma - P) - \delta P,$$
$$P_s = \mu G(L)(\gamma - P) - \tau P,$$
$$L = L_b,$$

where the dimensionless parameters are

$$\lambda := \frac{aK}{b^2}, \quad G(L) := \frac{\ell \beta S b}{K \alpha} F(\alpha L), \quad \mu := \frac{aK\alpha}{\beta b},$$
$$\tau := ka, \quad \delta := \frac{b k \beta}{K \alpha}, \quad \gamma := \frac{C \beta}{\beta}.$$

One possible choice for the free parameters $a$, $\alpha$, and $\beta$ is

$$a := \frac{1}{k}, \quad \alpha = \frac{K}{k \ell ab^3}, \quad \beta = \frac{K}{k \ell ab^4}.$$

It yields the dimensionless model

$$L_s = \lambda \Delta L, \quad \nabla L \cdot \eta = G(L)(\gamma - P) - P, \quad P_s = \gamma - P, \quad L = L_b,$$

where the scaled-time coordinate is $s$, the scaled space coordinate is $(\xi, \nu, \zeta)$,

$$\lambda = \frac{K}{kb^2}, \quad \gamma(\xi, \nu) = \frac{ab^4 k \ell}{K} C(b\xi, b\nu), \quad G(L) = \frac{\ell S}{k} F(\alpha L);$$

in $B_d$,

on $\partial \Omega_d$,

on $\partial \Omega_d$,

on $\partial B_d$. 

(4.146)
and, $B_d$ is the rectangular box with square bottom, unit height, and a square bottom whose side dimension is $\omega/b$.

Several computational problems are suggested; many variations are possible.

**Problem 4.95.** For definiteness, use the parameter values

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>box height $b$</td>
<td>0.04 m,</td>
</tr>
<tr>
<td>box bottom length $\omega$</td>
<td>0.04 m,</td>
</tr>
<tr>
<td>bulk ligand concentration $A_b$</td>
<td>$10^{-5}$ mol/m$^3$,</td>
</tr>
<tr>
<td>diffusion constant $K$</td>
<td>$1.0 \times 10^{-3}$ m$^2$/sec,</td>
</tr>
<tr>
<td>binding conversion rate $k$</td>
<td>$1.0 \times 10^{-4}$/sec,</td>
</tr>
<tr>
<td>ligand-receptor dissociation rate $\ell$</td>
<td>$1.5 \times 10^{-4}$/sec mol).</td>
</tr>
</tbody>
</table>

Let $C$, the total concentration of receptors and unbound receptors, be given by $C(x,y) = 1$ and the initial state of all receptors be unbound. Take $F$ to be $F(A) = cA$, where $c = 1$ unit of length, and assume the bulk concentration (Dirichlet) boundary condition at the outer boundary. Assume the initial ligand concentration in the medium is uniformly $A_b$. Determine the time-dependent bound receptor concentration and compare the result with formula (4.141).

**Problem 4.96.** Use the parameters (4.147). Suppose the initial ligand concentration is the bulk concentration in the rectangular layer above the plane 0.03 m parallel and above the bottom of the box. Let $C = 1$ and suppose the receptors are initially unbound. Take $F(A) = cA$, where $c = 1$ unit of length, and use the Dirichlet boundary condition on the outer boundary. When is the bound receptor concentration half of its maximum value?

**Problem 4.97.** Use the parameters (4.147). Suppose the initial ligand concentration is the bulk concentration in the rectangular layer above the plane 0.03 m parallel and above the bottom of the box. Let $C = 1$ and the receptors be initially unbound, and take $F$ to be given by $F(A) = cA$, where $c = 1$ unit of length. Set the boundary condition on the enclosure to be no flux on the lateral boundary of the box and evaporation at the rate of $10^{-6}$ mol/m$^2$ lost through the top of the box. Determine the bound receptor concentration as a function of time. When (if ever) are 20% of the receptors bound. Determine the critical value of the evaporation rate below which 20% of the receptors are eventually bound and graph the time when this state occurs versus the evaporation rate.

**Problem 4.98.** Use the parameters (4.147). Suppose the initial ligand concentration is the bulk concentration in the rectangular layer above the plane 0.03 m parallel and above the bottom of the box. Let $C = 1$ and assume that the receptors
are initially unbound. Choose the boundary condition on the enclosure to be no flux on the top and lateral boundary of the box. Take $F(A) = cA$, where $c$ is a constant with the dimensions of length. Determine the bound receptor concentration after 10 sec as a function of $c$ (measured in meters) on the range $0 \leq c \leq 1$.

4.7 Splitting Methods

This section introduces an interesting and not completely understood algorithm for approximating solutions of evolution equations based on the idea of separating differential equations into sums of simpler equations.

4.7.1 A Product Formula

Let $A$ be a matrix and define the matrix exponential of $A$ by

$$e^A = I + \sum_{j=1}^{\infty} \frac{A^j}{j!},$$

(4.148)

where $I$ is the identity matrix. Also, let $\mathcal{L}(E)$ denote the finite-dimensional vector space of all $N \times N$-matrices. The infinite sum in display (4.148) converges (absolutely with respect to every norm on $\mathcal{L}(E)$) and the matrix exponential satisfies the usual rules of exponents except that

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

if and only if $AB = BA$ (that is, the matrices $A$ and $B$ commute). Also,

$$\frac{d}{dt} e^{tA} = Ae^{tA} = e^{tA} A;$$

or, in other words, the function $t \mapsto e^{tA}$ is the matrix solution of the initial value problem

$$\dot{x} = Ax, \quad x(0) = I.$$  

These results are proved in textbooks on ordinary differential equations (see, for example, [15]).

The next theorem is a special case of the Lie–Trotter product formula for the exponential of a sum of two $N \times N$-matrices when the matrices do not necessarily commute (see [15] for a detailed proof).
**Theorem 4.99.** If \( \gamma : \mathbb{R} \to \mathcal{L}(E) \) is a continuously differentiable function with \( \gamma(0) = I \) and \( \dot{\gamma}(0) = A \), then the sequence \( \{\gamma^n(t/n)\}_{n=1}^{\infty} \) converges to \( \exp(tA) \). In particular, if \( A \) and \( B \) are \( k \times k \)-matrices and \( \gamma(t) := e^{tA}e^{tB} \), then

\[
e^{t(A+B)} = \lim_{n \to \infty} \left(e^{\frac{t}{n}A}e^{\frac{t}{n}B}\right)^n.
\]

The solution of the initial value problem

\[
\dot{x} = Ax + Bx, \quad x(0) = x_0. \tag{4.149}
\]

is

\[
x(t) = e^{t(A+B)}x_0.
\]

But, this compact notation hides the infinite sum that is used to define the exponential. To obtain numerical values, we must approximate this sum. A natural and viable way to proceed is to approximate the exponential with a partial sum of its Taylor series (see [48]).

Let us consider a variant of this problem. Suppose that we know how to compute \( \exp tA \) and \( \exp tB \), and we wish to compute \( \exp(t(A + B)) \). There are many possible ways to accomplish this. We will discuss some methods based on the product formula that will lead to a numerical method for solving some PDEs.

Take \( n = 1 \) in the product formula to obtain the approximation

\[
e^{t(A+B)} \approx e^{tA}e^{tB}.
\]

How much error does this introduce? To answer this question, simply expand in Taylor series (with respect to \( t \) at \( t = 0 \)) and subtract:

\[
\|e^{t(A+B)} - e^{tA}e^{tB}\| = \|I + t(A + B) + \frac{t^2}{2!}(A + B)^2
- (I + t(A + B) + \frac{t^2}{2!}(A^2 + 2AB + B^2))\| + O(t^3).
\]

In the generic case where \( A \) and \( B \) do not commute, the second-order terms do not cancel and the estimate reduces to

\[
\|e^{t(A+B)} - e^{tA}e^{tB}\| = O(t^2).
\]
Conservation of Mass

Viewing this as a numerical algorithm to solve system (4.149), we have a first-order method given by

$$x^0 = x_0, \quad x^{k+1} = e^{\Delta t A} e^{\Delta t B} x^k.$$  

To obtain a second-order method, we must cancel the second-order terms in the Taylor series that are used in our error estimates. Note that for

$$\gamma(t) = e^{t/2A} e^{tB} e^{t/2A}$$

we have $\gamma(0) = I$ and $\dot{\gamma}(0) = A + B$. Thus, this choice satisfies the hypothesis of Theorem 4.99; moreover,

$$\ddot{\gamma}(0) = A^2 + AB + BA + B^2 = (A + B)^2.$$  

Thus, with this choice for $\gamma$, we have the estimate

$$\|e^{t(A+B)} - e^{tA} e^{tB}\| = O(t^3)$$

and the second-order numerical method

$$x^0 = x_0, \quad x^{k+1} = e^{\Delta t/2A} e^{\Delta t B} e^{\Delta t/2A} x^k.$$  

(4.150)

A fourth-order method can be obtained using Richardson extrapolation. To accomplish this, note that one step of our second-order procedure is given by

$$U(h) = e^{h/2A} e^{hB} e^{h/2A} x_0;$$  

(4.151)

it is the approximation of the solution of our ODE at $t = h$. With half the step size, the solution is

$$U\left(\frac{h}{2}\right) = e^{h/4A} e^{h/2B} e^{h/4A} e^{h/4A} e^{h/2B} e^{h/4A} x_0.$$  

Be careful here; the notation is confusing. To use Richardson extrapolation, we must compare approximations of the same value (for instance, the solution of our initial value problem at time $t = h$). Here, $U(h)$ denotes the value at $t = h$ of our procedure, not a function with name $U$ defined by the expression in display (4.151). An application of Richardson extrapolation yields the fourth-order methods

$$U^1(h) = \frac{1}{3} (4U\left(\frac{h}{2}\right) - U(h)).$$  

(4.152)
Exercise 4.100. (a) Implement the numerical method (4.150) for the matrices
\[ A := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad B := \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix} \]

(b) Show by computation of this example that the method is second-order. (c) Show by computation that the Richardson formula (4.152) leads to a fourth-order method.

4.7.2 Products for Nonlinear Systems

To generalize the numerical method (4.150) to nonlinear systems we will take several leaps of faith, which is not uncommon in applied mathematics. The idea is simple: replace the matrix exponential solution \( t \mapsto \exp(tA) \) of \( \dot{x} = Ax \) by the flow of the system \( \dot{x} = f(x) \).

Recall that the flow of an ODE is defined to be the function \( \phi(t, \xi) \) such that \( t \mapsto \phi(t, \xi) \) is the solution of the initial value problem
\[ \dot{x} = f(x), \quad x(0) = \xi. \]

Solutions of differential equations may not be defined for all time, but it is easy to prove that
\[ \phi(0, \xi) = \xi, \quad \phi(t, \phi(t, \xi)) = \phi(t + s, \xi) \]
whenever both sides are defined (see [15]). Of course, the flow of the linear system \( \dot{x} = Ax \) is given by
\[ \phi(t, \xi) = e^{tA} \xi. \]

It is convenient to view flows as one-parameter groups of transformations of the state space. To emphasize this interpretation, let us write \( \phi_t(x) = \phi(t, x) \) so that
\[ \phi_0 = I, \quad \phi_t \circ \phi_s = \phi_{t+s}. \]

We can approximate the solutions of the system \( \dot{x} = f(x) + g(x) \), where we know the flow \( \phi \) of \( \dot{x} = f(x) \) and the flow \( \psi \) of \( \dot{x} = g(x) \) with the numerical method
\[ x^0 = \xi, \quad x^{k+1} = \phi_{\Delta t/2} \circ \psi_{\Delta t} \circ \phi_{\Delta t/2}(x^k), \tag{4.153} \]
where compositions of maps replace the products in display (4.150).

More generally, the solution of a well-posed parabolic PDE is given by a semiflow; that is, a semigroup of transformations of the underlying infinite-dimensional state space. While flows are defined for positive and negative
time (perhaps restricted to some interval of zero), semiflows are defined only
for some interval of the form \([0, T]\) on the real line for \(T > 0\) or \(T = \infty\). Fortunately, the formula in (4.153) makes sense for semiflows.

Let us return to the Gray-Scott model:

\[
\begin{align*}
\dot{u} &= \lambda (u_{xx} + u_{yy}) + F (1 - u) - uv^2, \\
\dot{v} &= \mu (v_{xx} + v_{yy}) - (F + K)v + uv^2 \\
\end{align*}
\]

(4.154)

with our zero Neumann boundary conditions. The obvious splitting corresponds to the reaction and diffusion systems

\[
\begin{align*}
\dot{u} &= F (1 - u) - uv^2, \\
\dot{v} &= -(F + K)v + uv^2 \\
\end{align*}
\]

(4.155)

and

\[
\begin{align*}
\dot{u} &= \lambda (u_{xx} + u_{yy}), \\
\dot{v} &= \mu (v_{xx} + v_{yy}). \\
\end{align*}
\]

(4.156)

The system of linear partial differential equations (4.156) with (almost any choice of) boundary data has a semiflow \(L_t\). It is the operator on initial data (concentrations \(u\) and \(v\) defined on the spatial domain \([0, L] \times [0, L]\) and satisfying the boundary conditions) that evolves the data forward from time zero to time \(t\) by solving the system of partial differential equations (4.156) including boundary conditions. Also, let \(\phi_t\) denote the flow of the system of ODEs (4.155). A useful numerical method (which incorporates Richardson extrapolation as in Section 4.7.1) is given schematically by

\[
\begin{align*}
\frac{w^{k+1}}{3} &= 4L_{\Delta t} \circ \phi_{\Delta t} \circ L_{\Delta t} \circ \phi_{\Delta t} \circ L_{\Delta t} (w^k) \\
&
\end{align*}
\]

(4.157)

where \(w = (u, v)\).

There is a hidden danger in method (4.157): Perhaps the numerical method does not preserve the boundary data. While, by definition, the semiflow \(L_t\) preserves the boundary conditions, this is not true in general for the flow \(\phi_t\) (see Exercise 4.101). Fortunately, for applications to Problem 4.47,
the flow does preserve the (no flux) zero Neumann boundary conditions. Indeed, suppose that \( u(x,y) \) and \( v(x,y) \) are concentrations that satisfy the Neumann boundary conditions. These concentrations are changed by the flow \( \phi \) to new concentrations given by the first and second components of the function

\[
(x, y) \mapsto \phi_t(u(x, y), v(x, y)).
\]

By the chain rule, we have that

\[
D\phi_t(u(x, y), v(x, y)) \begin{pmatrix} u_x(x, y) & u_y(x, y) \\ v_x(x, y) & v_y(x, y) \end{pmatrix}.
\]

The partial derivatives with respect to the first variable of the new concentrations vanish at \( x = 0 \) and \( x = L \) because the corresponding partial derivatives of \( u \) and \( v \) both vanish on these lines. Similarly, the partial derivatives with respect to the second variable of the new concentrations vanish at \( y = 0 \) and \( y = L \).

Implementation of the algorithm (4.157) leaves open many choices for numerical approximations of the semiflow \( L \) and the flow \( \phi \).

A viable method is to use the Crank-Nicolson scheme to approximate \( L \) and the trapezoidal method to approximate \( \phi \). These are, as mentioned previously, essentially the same second-order methods. The order estimates, for a code written for this algorithm and applied to Problem 4.48, are listed in Figure 4.39. Note that this implementation performs at nearly the theoretical fourth-order estimate using the larger step sizes. Accuracy degrades for smaller step-sizes, where there is a more likely accumulation of roundoff errors. The computed averages \((u + v)/2\) are consistently 0.453021 over the
Figure 4.40: The average concentrations in this figure are for the product method using the data of Problem 4.47. Gauss-Seidel iteration is used for the diffusion terms with the stopping criterion: successive iterations that differ in Euclidean norm by less than $10^{-6}$. Newton’s method is used in conjunction with the trapezoidal method for the reaction terms. The floating point numbers are the computed averages of the concentrations $(u + v)/2$.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$32 \times 32$</th>
<th>$64 \times 64$</th>
<th>$128 \times 128$</th>
<th>$256 \times 256$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta t = 1/4$</td>
<td>0.485018</td>
<td>0.463834</td>
<td>0.352093</td>
<td>0.335836</td>
</tr>
<tr>
<td>$\Delta t = 1/8$</td>
<td>0.485018</td>
<td>0.463835</td>
<td>0.352096</td>
<td>0.335841</td>
</tr>
<tr>
<td>$\Delta t = 1/16$</td>
<td>0.485018</td>
<td>0.463835</td>
<td>0.352097</td>
<td>0.335842</td>
</tr>
<tr>
<td>$\Delta t = 1/32$</td>
<td>0.485018</td>
<td>0.463835</td>
<td>0.352097</td>
<td>0.335842</td>
</tr>
<tr>
<td>$\Delta t = 1/64$</td>
<td>0.485018</td>
<td>0.463835</td>
<td>0.352097</td>
<td>0.335842</td>
</tr>
<tr>
<td>$\Delta t = 1/128$</td>
<td>0.485018</td>
<td>0.463835</td>
<td>0.352097</td>
<td>0.335842</td>
</tr>
</tbody>
</table>

Exercise 4.101. (a) Show that zero Dirichlet boundary conditions are usually not preserved by the numerical scheme (4.157). (b) Show that periodic boundary conditions are preserved.

Exercise 4.102. An alternative splitting that might be used to approximate solutions of the Gray-Scott model is obtained by incorporating the linear part of the reaction equations into the linear diffusion system. The remaining ”reaction” system (that is, the two equations $\dot{0} = -uv^2$ and $\dot{v} = uv^2$) might be solved analytically. Implement algorithm (4.157) for (4.47) using this new splitting. Compare your results with the original splitting.
Exercise 4.103. Use extrapolation across rows in Table 4.40 to estimate \( (u+v)/2 \) at \( T = 1024 \). Discuss the result.

4.8 Feedback Control

4.8.1 A Mathematical Model for Heat Control of a Chamber

We will discuss the problem of maintaining a desired temperature in a chamber equipped with a controlled heater/cooler, which for simplicity will be called the heater. Basic modeling with heat transfer and PID control will be considered. This problem is encountered in industrial process control (for example, in furnaces and tank reactors) and in laboratory work where it is desirable to control the temperature of a sample. Similar control systems are ubiquitous in biology, automobile cruise controls, steering of ships, and many other applications.

Let us represent the heater and chamber as open subsets \( \hat{\Omega} \) and \( \Omega \) of \( \mathbb{R}^3 \) respectively; and, let the wall separating them be represented by \( \Sigma := \partial \hat{\Omega} \cap \partial \Omega \subseteq \mathbb{R}^2 \). The temperature in the heater at \( x \in \hat{\Omega} \) and time \( t \) is denoted...
by $\hat{u}(x, t)$ and the temperature in the chamber at $x \in \Omega$ and time $t$ is denoted by $u(x, t)$.

Heat transport through the interface wall $\Sigma$ will be modeled under the assumption that the wall is thin and the heat flux through the wall is proportional to the difference in temperature on the two sides of the wall (Newton’s law of cooling). The temperature of the heater is actuated by a PID controller whose input is the continuous measurement of temperature by a thermometer placed at a fixed point $x_0$ in the interior of the chamber and whose action is to change the temperature in some region (for example, a heating element coil) in the heater box. For simplicity, let us assume that the temperature may be raised or lowered by the heater instantaneously. For definiteness and simplicity convection is ignored and the heater and chamber are assumed to be insulated at all exterior walls. Heat loss due to all causes (radiation, conduction or leakage through walls, opening and closing access to the chamber, etc.) not specifically modeled is included in a simplistic manner by modeling all such causes as if the chamber as a whole was subject to Newton’s law of cooling. The cases where the temperature can be changed in only one direction or where there is a delay in the time required to obtain the heat change called for by the controller are left to the student after reading this section.

The physical parameters are the heat transfer coefficient $\lambda$ across the wall, the heat transfer rate $\alpha$ from all causes that are not specifically modeled, the desired (set point) temperature $T_s$, the diffusion constants $\hat{K}$ and $K$, the initial heater temperature $\hat{u}_0$ and chamber temperature $u_0$, and the ambient temperature $T_a$.

The model equations, under our assumptions and with the notation just described, is given by a system of PDEs, boundary conditions, and initial conditions. The PDEs are

\[
\begin{align*}
\frac{\partial \hat{u}}{\partial t} &= \hat{K} \Delta \hat{u} + h(x) \text{PID}(T_s - u(x_0, t)), \quad (x, t) \in \hat{\Omega} \times (0, \infty), \\
\frac{\partial u}{\partial t} &= K \Delta u + \alpha(T_a - u), \quad (x, t) \in \Omega \times (0, \infty),
\end{align*}
\]  

where the function $h$ is nonzero at points in the heating element and zero otherwise, and PID denotes the output of the PID controller. A model for heat flow through the interface $\Sigma$ between the heater and the chamber is determined from Newton’s law of cooling: The heat flux through a surface is proportional to the difference in the temperatures across the surface; and,
Conservation of Mass

Fourier’s law: The vector field that determines the heat flow is proportional to the negative gradient of the temperature. (Note: At a slightly deeper level Newton’s law is a special case of Fourier’s law.) There are related boundary conditions for $\hat{u}$ and $u$ on $\Sigma$. For the outer unit normal $\hat{\eta}$ on $\Sigma$, which points into the chamber, the boundary condition is

$$\int_{\Sigma} -\hat{K} \text{grad} \hat{u} \cdot \hat{\eta} \, dS = \lambda (\hat{u} - u),$$

where $\lambda$ is the (positive) constant of proportionality in Newton’s law and $\hat{K}$ is the (positive) constant of proportionality in Fourier’s law (the thermal diffusivity). For example, suppose that $\hat{u} > u$. In this case, the gradient of $\hat{u}$ points into the heater, $-\hat{K} \text{grad} \hat{u}$ points into the chamber, the dot product is positive and the positive integral agrees in sign with the difference $\hat{u} - u$.

The constant $\text{lambda}$ depends on the material properties of the interface. For the boundary of the chamber,

$$\int_{\Sigma} -K \text{grad} u \cdot \eta \, dS = \lambda (u - \hat{u})$$

with the same $\lambda$ because $\Sigma$ is the boundary for the heater and the chamber.

For the insulated exterior walls, the assumption of zero flux reduces to several conditions, one for each wall $W$, of the form

$$\int_{W} \text{grad} v \cdot N \, dS = 0$$

where $v$ is either $\hat{u}$ or $u$ and $N$ is the outer normal on a exterior boundary wall of the heater (corresponding to the boundary conditions for $\hat{u}$) or the outer normal on a boundary of the chamber (corresponding to $u$). The initial conditions are

$$\hat{u}(x, 0) = \hat{u}_0 (x), \quad x \in \hat{\Omega},$$
$$u(x, 0) = u_0 (x), \quad x \in \Omega,$$  \hspace{1cm} (4.159)

for $\hat{u}_0$ a given temperature distribution in the heater and $u_0$ a temperature distribution in the chamber.

The PID control—which is written in the PDEs with an abuse of notation—is given by

$$\text{PID}(f)(t) = K_p f(t) + K_i \int_{0}^{t} f(\sigma) \, d\sigma + K_d f'(t),$$  \hspace{1cm} (4.160)
where $K_p$, $K_i$ and $K_d$ are parameters that determine the proportional, integral, and derivative controller gains. The control input in the model equations may be rewritten in the explicit form

$$\text{PID}(T_s - u(x_0, t)) = K_p(T_s - u(x_0, t)) + K_i \int_0^t (T_s - u(x_0, \sigma)) d\sigma - K_d \frac{\partial u}{\partial t}(x_0, t).$$

Note that the integral part of the PID control depends on the entire past history of the control process. In practice, the controller should be influenced by only part of the past history; that is, integration should be over some interval $[t - \tau, t]$, where $\tau$ is some appropriately chosen positive number.

### 4.8.2 A One-dimensional Dimensionless Heat Control Model With PID Control

We may collapse the geometry of our controlled heater-chamber model to one dimension under the assumption that the axial direction from the heater to the thermometer carries the essential information. In other words, we may assume that the temperature in each slice perpendicular to this axis is constant. The heater reduces to a rod of length $\hat{L}$ parameterized by the interval $\hat{\Omega} := (0, \hat{L})$ (which represents the interior of the heater) and a rod of length $L$ parameterized on the interval $\Omega := (\hat{L}, \hat{L} + L)$ (which represents the interior of the chamber). With this change of notation, $\Sigma := \{\hat{L}\}, \partial \Omega_1 \setminus \Sigma = \{0\}$, and $\partial \Omega \setminus \Sigma = \{\hat{L} + L\}$. The original model (4.158)-(4.159) reduces to the coupled system of boundary value problems

\begin{align*}
\hat{u}_t &= \hat{K}\hat{u}_{xx} + h(x)(K_p(T_s - u(x_0, t)) + K_i \int_0^t (T_s - u(x_0, \sigma)) d\sigma - K_d \frac{\partial u}{\partial t}(x_0, t)), \quad (x, t) \in \Omega_1 \times (0, \infty), \\
\hat{u}_x(0, t) &= 0, \\
-A\hat{K}\hat{u}_x(\hat{L}, t) &= \lambda(\hat{u}(\hat{L}, t) - u(\hat{L}, t)) \quad (4.161)
\end{align*}

and

\begin{align*}
u_t &= Ku_{xx} + \alpha(T_a - u), \quad (x, t) \in \Omega \times (0, \infty), \\
u_x(\hat{L} + L, t) &= 0, \\
AKu_x(\hat{L}, t) &= \lambda(u(\hat{L}, t) - \hat{u}(\hat{L}, t)) \quad (4.162)
\end{align*}
with the initial conditions

\[ \hat{u}(x, 0) = \hat{u}_0(x), \quad x \in \hat{\Omega}, \]
\[ u(x, 0) = u_0(x), \quad x \in \Omega. \quad (4.163) \]

The change in sign of the left-hand side of the last interface boundary condition is due to the use of the partial derivative \( \partial u / \partial x \) to replace the dot product of the gradient and the normal. The outer normal on the boundary of the chamber points in the negative direction of the spatial coordinate.

The parameters in our model are \( \hat{K}, K, \lambda, T_s, \hat{L}, L, \) and \( A \) (the area of the interface \( \Sigma \)). More specifically, \( \hat{K} \) and \( K \) are thermal diffusivities with units of area per time (with values in SI units of square meters/second), \( \lambda \) is diffusivity per length, \( T_s \) is temperature (measured in SI units in kelvins), and the length and areas (which are measured in meters and square meters respectively). There are many thermal quantities defined in the literature; thus care is required to make sure named measured quantities are compatible with a given model. For example, standard thermal diffusivities for various materials are available; they can also be computed using thermal diffusivity is equal to \( \kappa / (\rho c_s) \), where \( \kappa \) is the thermal conductivity, \( \rho \) is the material density, and \( c_s \) is the specific heat. Similarly, \( \lambda \) is thermal diffusivity times length, where the appropriate length for our heat control problem is the thickness of the interface wall.

The natural temperature scale for our model is \( T_s \), a natural length scale is \( \hat{L} \), and a natural time scale is \( AL/\lambda \). There are many other possibilities. Using these scales, define the dimensionless variables

\[ \xi := \frac{x}{\hat{L}}, \quad \tau = \frac{\lambda t}{AL}, \quad T_s \hat{U}(\xi, \tau) := \hat{u}(x, t), \quad T_s U(\xi, \tau) := u(x, t) \]

and dimensionless groups

\[ \ell := \frac{L}{\hat{L}}, \quad T_r := \frac{T_a}{T_s}, \]
\[ \hat{k} := \frac{KA}{\lambda \hat{L}}, \quad k := \frac{KA}{\lambda L}, \quad \hat{\mu} := \frac{\lambda \hat{L}}{AK}, \quad \mu := \frac{\lambda L}{AK}, \quad a := \frac{LA\alpha}{\lambda}, \]
\[ k_p := \frac{LAK_p}{\lambda}, \quad k_i := \frac{LAK_i}{\lambda}, \quad k_d := \frac{LAK_d}{\lambda} \]
to obtain the dimensionless model
\[ \hat{U}_\tau = \hat{k}\hat{U}_{\xi\xi} + H(\xi)(k_p(1 - U(\xi_0, \tau)) + k_i \int_0^\tau (1 - U(\xi_0, \sigma)) d\sigma \\
- k_dU_\tau(\xi_0, \tau)), \]
\[ U_\tau = kU_{\xi\xi} + a(T_r - U), \]
\[ \hat{U}_\xi(0, \tau) = 0, \]
\[ -\hat{U}_\xi(1, \tau) = \hat{\mu}(\hat{U}(1, \tau) - U(1, \tau)), \]
\[ U_\xi(\ell + 1, \tau) = 0, \]
\[ \hat{U}(\xi, 0) = \hat{U}_0(\xi), \]
\[ U(\xi, 0) = U_0(\xi), \] (4.164)

where \( H(\xi) = h(\hat{L}\xi), \hat{U}_0(\xi) = \hat{u}_0(\hat{L}\xi)/T_s, \) and \( U_0(\xi) = u_0(\hat{L}\xi)/T_s. \)

In an ideal world we would like to look in tables in the literature to find values for the physical parameters \( \hat{K}, K \) and \( \lambda \), substitute these into our model with appropriate lengths and areas for the problem at hand and use the model to predict the behavior of the system in open loop (without the control) and in closed-loop (with the control). This should be possible if our model were a close fit to reality. In the real world, the effect of the assumptions used to make the model is to render the model to be less than exact. Thus, the model with table-book physical parameters may not produce accurate results. Perhaps a model is too simplistic or simply wrong. In case the model is physically reasonable (like the model (4.161)–(4.163)), we may try to calibrate the model by modifying the coefficients to fit experimental data gathered from the physical problem at hand (the open loop heater-chamber system).

**Problem 4.104.** Suppose the physical control system discussed in this section has a heating/cooling element located in the center of an air filled cubic box with axial length \( \hat{L} = 20 \) centimeters; the function \( h \) in the model has value \( h = 1 \) on this set and zero otherwise; the interface is a steel plate with area 400 square centimeters and thickness 0.5 centimeters (which in our model is a portion of a plane with zero thickness); the chamber is a rectangular box with axial length 60 centimeters filled with air; and the thermometer is placed in the center of the chamber. In reality, the order of magnitude of the thermal diffusivity of air is \( 10^{-5} \) square meters per second and the same
for (some types of) steel. For definiteness, use SI units (meters, seconds, kilograms, and kelvins) and assume the physical model parameters are

\[ \hat{K} = K = 10^{-5}, \quad \lambda = 5 \times 10^{-8}, \quad \hat{L} = 0.2, \quad L = 0.6, \quad A = 0.04. \]

The corresponding dimensionless groups are approximated by

\[ \hat{k} = k = 40.0, \quad \hat{\mu} = \mu = 0.025. \]

Also, assume that the rate of heat loss \( \alpha \) due to causes not specifically modeled is \( 6.86 \times 10^{-6} \), which is about a drop of 5 degrees Kelvin in one hour when the chamber temperature is 500 kelvins and the ambient temperature is 295 kelvins. The corresponding dimensionless variable has value \( a \approx 1.097 \).

(a) Discuss tuning the control gains to achieve the shortest time startup from system temperature 295 kelvins to the set-point temperature 500 kelvins without overshoot that would exceed 550 kelvins as the system continues to run with the same control gains. To make the problem somewhat realistic, limit the dimensionless control gains (which would be related to the limits of the heater/cooler) to be in the range \([0, 1000]\).

(b) Discuss tuning the controller gains that would keep the chamber temperature within five kelvins of the set point when subjected to a fluctuation of ambient temperature of at most 15 kelvins.

To simulate our dimensionless heat control problem, we may seek an analytic solution or approximation of our model or discretize the equations and use a numerical method to approximate solutions.

The method of lines (recall, Exercise 4.75) is well suited to numerical simulation for our model equations. The idea is simple: discretize in space (but not in time) to obtain a system of ODEs that incorporates the boundary conditions and approximate its solutions using a numerical method for ODEs. In principle, Euler’s method applied to PDEs is a special case of the method of lines. Thus, the new method is simply a change in point of view. To be clear, let us consider the heat equation in one space dimension with zero Neumann boundary conditions and initial data:

\[ u_t = u_{xx}, \quad u_x(0, t) = 0, \quad u_x(L, t) = 0, \quad u(x, 0) = f(x). \quad (4.165) \]

Discretize the spatial domain \((0, L)\) into \( n \) parts of length \( \Delta x = L/n \) and define \( n - 1 \) variables \( \{u_i\}_{i=1}^{n-1} \). The spatial derivative may be discretized
Conservation of Mass

For simplicity and in keeping with usual practice for a first approximation, the second-order centered difference approximation (with \( u_i := u(i\Delta x, t) \))

\[
    u_{xx}(i\Delta x, t) \approx \frac{u_{i-1} - 2u_i + u_{i+1}}{\Delta x^2}
\]

is a reasonable choice. The Neumann boundary conditions may be approximated by taking \( u_0 := u_1 \) and \( u_n := u_{n-1} \). This choice is simple to program, but not second-order accurate. Higher-order accuracy can be achieved by using multi-point approximations of the first derivative. For these choices we obtain a coupled system of \( n - 1 \) ODEs:

\[
    \begin{align*}
    \dot{u}_1 &= \frac{1}{\Delta x^2}(-u_1 + u_2), \\
    \dot{u}_i &= \frac{1}{\Delta x^2}(u_{i-1} - 2u_i + u_{i+1}), \quad i = 2, 3, 4, \ldots, n - 2 \\
    \dot{u}_{n-1} &= \frac{1}{\Delta x^2}(u_{n-2} - u_{n-1}).
    \end{align*}
\]

(4.166)

Initial data for the system is defined in the obvious manner from the initial data for the PDE. Note the presence of the factor \( 1/(\Delta x)^2 \), which will be large whenever \( \Delta x \) is small. This could cause numerical difficulties. One possible remedy is the change of time scale \( t = \Delta x^2 s \). It removes the undesirable factor, but forces the computation of a large number of time steps. These considerations should remind you of the Courant-Lewy-Friedricks condition discussed in in Section 4.4.4. There is no obvious resolution of these difficulties, which are inherent in most numerical methods. They may be partially resolved by further numerical analysis that is beyond the scope of this book.

For a simple implementation of the method of lines for system (4.164), we may use second-order accurate discretizations of the second-order spatial derivatives and first-order accurate discretizations of the zero Neumann boundary conditions. To describe a viable discretization of the boundary conditions at the interface \( (\xi = 1) \), consider the discretization of the interval \((0, \ell + 1)\) given by choosing positive integers \( m \) and \( n \), defining \( \Delta \xi := 1/m \) and \( \Delta \xi = \ell/n \), and using \( \dot{u}_i := \dot{U}(i\Delta \xi, \tau) \) and \( u_i := U(i\Delta \xi, \tau) \). The temperature at the thermometer (placed in the middle of the chamber) may be approximated by \( (u_{[n/2]_+} + u_{[n/2]_+})/2 \), where \([j]_+\) is the smallest integer larger than or equal to \( j \) and \([j]_-\) is the largest integer less than or equal to
This approximation may be used to code the proportional control. The integral control may be implemented by adding to the method-of-lines system of ODEs the differential equation and initial value
\[
\frac{dz}{d\tau} = 1 - U(\xi_0, \tau) \approx 1 - \frac{u_{[n/2]_-} + u_{[n/2]_+}}{2}, \quad z(0) = 0.
\]

The variable \( z \) is the desired integral. The derivative control may be added by simply using the right-hand sides of the method-of-lines ODEs at \([n/2]_-\) and \([n/2]_+\). The boundary conditions at the interface couple the system. Using first-order approximations (which are adequate, but not very accurate) and the proposed numbering scheme, the boundary conditions are
\[
\hat{u}_m - \hat{u}_{m-1} \left/ \Delta \xi \right. = \hat{\mu}(u_0 - \hat{u}_m),
\]
\[
\frac{u_1 - u_0}{\Delta \xi} = \mu(u_0 - \hat{u}_m). \tag{4.167}
\]

The state variables for the method of lines are \( \hat{u}_i \) for \( i = 1, 2, 3, \ldots, m - 1 \) and \( u_i \) for \( i = 1, 2, 3, \ldots, n - 1 \); the system (4.167) is two equations for the two unknowns \( \hat{u}_m \) and \( u_0 \); that is, the (scaled) temperatures of the heater and the chamber at the interface. By solving this system, we have that
\[
\hat{u}_m = \frac{1}{1 + \hat{\mu} \Delta \xi + \mu \Delta \xi}((1 + \mu \Delta \xi)\hat{u}_{m-1} + \hat{\mu} \Delta \hat{\xi}u_1),
\]
\[
u_0 = \frac{1}{1 + \hat{\mu} \Delta \xi + \mu \Delta \xi}(\mu \Delta \xi \hat{u}_{m-1} + (1 + \hat{\mu} \Delta \hat{\xi})u_1). \tag{4.168}
\]

These formulas are used whenever \( \hat{u}_m \) or \( u_0 \) are needed for computation of the discretized second derivatives.

All the ingredients are now in place to simulate the system using the method of lines. The system of \( m + n - 2 \) ODEs may be approximated with a variety of methods. Euler’s method is a possibility. To maintain the second-order accuracy for the spatial discretizations of second derivatives, implicit or explicit trapezoidal time-stepping is preferable. In keeping with the method of lines point of view, we may also use a (perhaps more sophisticated) black-box ODE solver.

**Exercise 4.105.** Implement the method of lines to approximate system (4.165) using the ODEs (4.166). Discuss and compare the efficiency and accuracy of the
method for the ODEs in their given form and after scaling time with the change of variables \( t = \Delta x^2 s \). Compare your numerical results with the exact solution for your choice of initial data.

**Exercise 4.106.** Implement the method of lines to approximate system (4.165) using the ODEs (4.166). Discuss the accuracy of the numerics for the ODEs and for their modification using a second-order accurate approximation for the boundary conditions. State explicitly your second-order accurate discrete approximation of the second derivative. Compare your numerical results with the exact solution for your choice of initial data.

**Exercise 4.107.** Consider the following dimensionless model for heat diffusion with proportional boundary control:

\[
\begin{align*}
  u_t &= u_{xx}, \\
  u_x(0, t) &= u(0, t) - k_p (1 - u(1/2, t)), \\
  u(0, 1) &= 3, \\
  u(x, 0) &= 0.
\end{align*}
\]

(a) Show that it is not possible to choose the control gain \( k_p \) so that the system will maintain the set point \( u(1/2, t) = 1 \)? (b) What is the behavior of the function \( t \to u(1/2, t) \) as \( t \) increases—that is, when the system is allowed to run—for a fixed control gain. (c) Show that the set point can be maintained using a PI control.

**Exercise 4.108.** Problem (4.104) is the main problem for this section. (1) Implement a numerical method and solve its part (a). (2) Solve part (b).

**Exercise 4.109.** Modify Problem (4.104) for the case where no active cooling is available; that is, a heater is available subject to PID control but there is no active cooling. The heater may be turned off. Using the given parameters solve part (a).

**Exercise 4.110.** [Modeling Project] Consider a cylindrical tank with base radius \( a \) and height \( h \). Suppose water flows into the tank intermittently. The tank has a circular drain in the center of its base with radius \( r \) and a controllable valve that is designed to change the radius of the drain from zero (fully closed) to \( r \) fully open. A sensor measures the depth of the water in the tank and is connected to a PID controller and a servo mechanism that can continuously activate the valve control. Suppose that the inflow never exceeds the outflow capacity with the valve fully open and the flow velocity in the drain is \( \sqrt{2gz} \), where \( g \) is the acceleration due to gravity and \( z \) is the depth of water in the tank (see Exercise ??).

(a) Make a mathematical model of the open loop system (no control); that is, a fixed drain radius.

For the following parts of this problem assume that the system parameters are

\[
\begin{align*}
  a &= 3 \text{ m}, \\
  h &= 6 \text{ m}, \\
  r &= 0.12 \text{ m}, \\
  g &= 9.8 \text{ m/sec}^2.
\end{align*}
\]
the desired depth is \( h_{\text{set}} = 5 \text{ m} \), and the density of water is \( \rho = 10^3 \text{ kg/m}^3 \).
(b) Discuss controller gain tuning for P control used to maintain the depth of the water in the tank at the set point for a constant inflow of 300 kg/sec.
(c) Discuss controller gain tuning for PI control with the set point depth as in part (b) for the system startup from an empty tank.
(d) Discuss controller gain tuning for PID control with the set point depth as in part (b) for the system startup from an empty tank.
(e) Discuss controller gain tuning in case the inflow rate fluctuates between 80 and 120 kg/sec on a one minute cycle. Note: There is usually no reason to tune control gains to meet all possible situations. For example, a good tuning for startup might not be suitable to maintain a desired set point during later operation. Setting different gains for different regimes is called gain scheduling.

4.8.3 Problems and Projects: Concentration Gradients, Chemotaxis, and Cruise Control.

Exercise 4.111. [Concentration Gradients] (a) Make a mathematical model to describe the diffusion of a solute (call it’s concentration \( u \)) in water through a permeable membrane idealized as a cross section of a tube (perhaps an idealized blood vessel) with one closed end. Suppose the tube radius is \( a \), its length is \( L \), and the membrane is placed at a distance \( pL \) (for some \( 0 < p < 1 \)) from the open end of the tube. The tube is filled with pure water and its open end is connected to a reservoir with a large supply of solution containing the mentioned solute with concentration \( c \). This solute diffuses in the tube. The solute moves across the membrane according to a form of Fick’s law: the solute flux across the membrane is proportional to the concentration difference across the membrane and in the direction from higher to lower concentration. Let \( k \) denote the diffusion constant for the solute in water and \( \lambda \) the diffusivity constant (the constant of proportionality across the membrane).
(b) Using the same geometry, make a model for the situation where the solute is secreted through the closed end of the tube and leaves the open end of the tube that is immersed in pure water. Assume that the flux across the closed end of the tube is known. (c) Choose numerical values for all the parameters in this problem and determine the corresponding steady state solute concentration in the tube as a function of position along the tube.

Exercise 4.112. [Chemotaxis] Let \( u \) denote the density of a population of bacteria, cells, insects, or other organism whose motion is influenced by the presence of a chemical in their environment with concentration \( c \) and recall the basic con-
Conservation of Mass

Equation (4.22)
\[ u_t = - \text{div}(X) + f, \]
where \( X \) is the diffusive flux of \( u \) and \( f \) is the amount of substance generated per volume per time, which is here taken to be zero. Our organisms are assumed to diffuse from higher concentration to lower concentration independent of the presence of the chemical. Use the usual constitutive law \( X = -K \text{grad} u \) to model this process and note that \( X \) is measured in mass per area per time. In the present case mass is essentially the number of organisms at a point (the units \( u \) times volume). The orientation of organisms with respect to the presence of the chemical (the chemotaxis) is modeled by modifying the flux term to read
\[ X = -K \text{grad} u + u\chi(c) \text{grad}(c), \]
where \( \chi \) is the chemotactic sensitivity measured in units of inverse chemical concentration times length per time. This assumption leads to the chemotaxis equation
\[ u_t = \text{div}(K \text{grad} u - u\chi(c) \text{grad}(c)). \]

In case the chemical concentration is not affected by the presence of the organisms, it may be specified. Or, it’s concentration may be modeled by a reaction diffusion equation
\[ c_t = k\Delta c + g(c, u), \]
where \( g \) models the creation and consumption of the chemical in the environment. (a) Suppose that the organisms do not diffuse \((K = 0)\), the chemical concentration in a two-dimensional environment is given by a function of the form \( e^{-x^2-y^2} \) and the chemotactic sensitivity is constant (either \( +1 \) or \( -1 \)) discuss the change in concentration of the organisms with respect to time under the assumption that the organisms are initially uniformly distributed over the environment. (b) Repeat part (a) with diffusion \((K > 0)\). (c) Suppose the organisms secrete the chemical at a rate inversely proportional to the presence of the chemical. Write the model under this assumption and discuss the effect on the concentration of organisms. Use a two-dimensional environment on a finite part of the plane and discuss your choice of boundary conditions.

Exercise 4.113. [Chemotaxis for Individuals-Agent Based Modeling] Imagine a hypothetical insect that has three antennae: one on each side of its head and one pointing forward along the insect’s axis of symmetry. The insect can sense the concentration of a chemical with each of its antennae. Also the insect has short-term memory: It can remember the output of its sensors over some fixed period of time. The insect feeds on substances the excrete the chemical it can sense. (a) Design a controller (that would mimic the controller that would have been
designed through evolution for the insect) so that the insect will follow a gradient of increasing concentration of the chemical to its source, where the food is likely to reside. Your model might be continuous or discrete and the controller might not be PID. (b) Test your control strategy by simulation using chemical trails of your own design. (c) Suppose there are many insects. Augment your control rules so that two insects cannot occupy the same space at the same time. Simulate the motion of several insects using your control strategy. Your model is likely to be an example of an agent based model: a set of agents, an environment in which they reside, and a set of rules that determine how each agent behaves.

Exercise 4.114. [Cruise Control] A basic model for an automobile cruise control starts with a model of the motion of an automobile over a road. For simplicity, assume that the controller will be tested on a straight road over hilly terrain. Choose an inertial coordinate system—in this case, a coordinate system fixed to the Earth is a reasonable approximation—such that the positive horizontal axis is in the direction of the road and the vertical axis points away from the surface of the Earth. The road may then be idealized as the graph of a function $f$ that gives the elevation $y$ over each horizontal position $x$ via $y = f(x)$. This reduces the problem to two-dimensions. As usual, let the position of the automobile as a function of time $t$ be denoted by $(x(t), y(t))$. There are several forces acting on the automobile: the forward force provided by the automobile’s engine, gravity, aerodynamic drag, rolling resistance due to deformation of the tires, and the force that keeps the automobile from dropping through the road. Assume that the latter force $F_N$ is everywhere normal to the road and lump the sum of all other forces except gravity into one force denoted by $G$. According to Newton’s second law and assuming the mass $m$ of the automobile does not change with time due to fuel consumption or other reasons) the equation of motion is

$$ m \begin{pmatrix} \ddot{x} \\ \ddot{y} \end{pmatrix} = -mg \begin{pmatrix} 0 \\ 1 \end{pmatrix} + F_N + G. \quad (4.169) $$

The unit tangent to the road is the vector

$$ T := \frac{1}{\sqrt{1 + (f'(x))^2}} \begin{pmatrix} 1 \\ f'(x) \end{pmatrix}. $$

By computing the inner product of both sides of the differential equation (4.169) with $T$, the model is reduced to

$$ \frac{m}{\sqrt{1 + (f'(x))^2}} (\ddot{x} + f'(x)\ddot{y}) = -mg \frac{f'(x)}{\sqrt{1 + (f'(x))^2}} + G \cdot T. $$
Conservation of Mass

The system is reduced to one-dimension by using the relation $y = f(x)$ to obtain

$$m\left(\ddot{x}\sqrt{1 + (f'(x))^2} + \frac{f'(x)f''(x)}{\sqrt{1 + (f'(x))^2}}\right) = -mg\frac{f'(x)}{\sqrt{1 + (f'(x))^2}} + G \cdot T. \quad (4.170)$$

While the equation of motion (4.170) with $G$ specified is a model of the automobile motion, it is not well suited to designing and simulating a cruise control in case the sensor determines the vehicle speed along the road; that is, with respect to arc length along the graph of $f$, which is given by

$$\ell := \int_{x_0}^{x} \sqrt{1 + (f'((\xi))^2)} \, d\xi.$$ 

For this reason, it is advantageous to recast the model (4.170) with the dependent variable being speed along the road; that is,

$$s := \frac{d\ell}{dt} = \frac{d\ell}{dx} \frac{dx}{dt} = \dot{x} \sqrt{1 + (f'(x))^2}.$$ 

Note that

$$\dot{s} = \dot{x} \sqrt{1 + (f'(x))^2} + \frac{f'(x)f''(x)}{\sqrt{1 + (f'(x))^2}}.$$ 

Thus, we may consider the model equation as the system

$$m\dot{s} = -mg\frac{f'(x)}{\sqrt{1 + (f'(x))^2}} + G \cdot T,$$

$$\dot{x} = \frac{s}{\sqrt{1 + (f'(x))^2}}. \quad (4.171)$$

The drag force acts in the direction opposite to the motion and its magnitude may be approximated by

$$\frac{1}{2} \rho C_d A s^2,$$

where $\rho$ is the density of the air, $C_d$ is the dimensionless drag coefficient for the particular automobile under consideration and $A$ is the cross sectional area presented by the projection of the automobile onto a plane perpendicular to the direction of its motion.

A good model for rolling resistance requires some knowledge of the physics of tires. A crude approximation is made by simply regarding this force as a resistance to the motion proportional to the speed of the vehicle; that is,

$$F_{rr} := C_{rr}s,$$

where $C_{rr}$ is a constant with units of mass per time.
The force due to the automobile engine may be modeled in several different ways. A simple approach requires the tire radius \( r \) and the angle \( \theta \) (measured in radians) that the vector from axis of rotation to a point on the tread of a tire makes with the horizontal. In the ideal situation where there are no forces acting, the speed \( s \) of forward motion is given by \( s = 2\pi r \dot{\theta} \). The automobile mass times its acceleration is

\[
m \ddot{s} = 2\pi r \ddot{\theta} = \frac{2\pi m r}{I} |\tau|,
\]

where \( I \) is the moment of inertia of the wheel and \( \tau \) is the torque on the wheel provided by the engine.

By inserting the forces into system (4.171), the automobile motion model is

\[
m \ddot{s} = -mg \frac{f'(x)}{\sqrt{1 + (f'(x))^2}} - \frac{1}{2} \rho C_d As^2 - C_{rr} s + \frac{2\pi m}{I} |\tau|,
\]

\[
\dot{x} = \frac{s}{\sqrt{1 + (f'(x))^2}}.
\] (4.172)

The cruise control actuator may be taken to be a linkage to a variable valve that changes the amount of fuel flowing to the engine. In turn, the fuel supply to the engine determines the torque supplied to the wheel through the drive train of the automobile. The valve, fuel pump, fuel line, injection system, transmission, differential, etc. is a complicated mechanical system. Is it necessary to have a precise model of this system to design a useful cruise control? This question can only be answered by building and testing cruise control systems. The purpose of the model is preliminary design. It seems that a detailed model of the fuel supply to the engine and the drive train would lead to unnecessary complication. A reasonable way to proceed is to simply assume the torque \(|\tau|\) supplied to the wheel is proportional to the fuel supply. For example, we may assume that

\[
|\tau| = \lambda Q,
\]

where \( Q \) is the volumetric flow rate \( (dV/dt) \) of fuel to the engine. In this case, the constant of proportionality \( \lambda \) has units of mass per length per time. Perhaps the constant of proportionality would be easier to measure by experiments with the automobile if this coefficient were dimensionless. To achieve this we might redefine our relationship to be

\[
|\tau| = \lambda m r Q,
\]

where \( m \) is the mass of the automobile and \( r \) is the tire radius so that the new \( \lambda \) is a dimensionless coefficient of proportionality between volumetric fuel consumption of the engine and the torque supplied to the wheels. The volumetric flow rate \( Q \) may be subjected to a PID control with desired (set-point) speed \( s_d \). In this application,
the volumetric fuel consumption is limited by the design of the automobile engine. There are two positive rates $\alpha$ and $\omega$ such that

$$\alpha \leq Q \leq \omega.$$ 

The lower limit is positive to keep the engine running even when the speed exceeds the set point. Thus, the PID control must be composed with an appropriate function $H$ to impose the saturation limits; that is, the controlled volumetric flow rate is

$$Q = H(k_P(s_d - s) + k_I \int_{t-a}^{t} (s_d - s(\sigma)) d\sigma - k_D \dot{s}),$$

where $a \geq 0$ is the amount of time kept in the controller memory.

A closed loop control model is

$$m \ddot{s} = -mg \frac{f'(x)}{\sqrt{1 + (f'(x))^2}} - \frac{1}{2} \rho C_d A s^2 - C_{rr} s$$

$$+ \frac{2\pi m^2}{I} H(k_P(s_d - s) + k_I \int_{t-a}^{t} (s_d - s(\sigma)) d\sigma - k_D \dot{s}),$$

$$\dot{x} = \frac{s}{\sqrt{1 + (f'(x))^2}}.$$  \hspace{1cm} (4.173)

(a) Criticize the closed loop control model.

(b) Set all system parameters equal to unity. Suppose that the road elevation is given by $f(x) = 2 \exp(-x^2)$ and the automobile starts at position $x = 0$ at time $t = 0$. Optimize the control parameters $k_P$, $k_I$, $k_D$, and $a$ over the interval $0 \leq x \leq 5$.

(c) Using the parameters for part (b), define a typical road terrain $f$ that might be used to test cruise control systems, defend your choice, and optimize the control parameters over a finite distance on your terrain. Define two new terrains (one of which might have some atypical feature) and test your optimized control system over the new terrains. Write a report on your findings.

(d) Consult the literature on automotive engineering for approximations of the system parameters for a particular automobile. Define a typical terrain $f$ and optimize the control parameters for a set-point speed of 70 miles per hour over a finite distance on your terrain. Define two new terrains (which perhaps have some atypical but physically reasonable features) and test your optimized control system over the new terrains. Write a report on your findings.
Appendix A

Mathematical Notes

A.1 Arzela-Ascoli Theorem

A set $S$ of continuous functions defined on a compact set $\Omega$ is uniformly bounded if there is some number $M > 0$ such that $\sup_{x \in \Omega}|f(x)| \leq M$ for every $f \in S$. The set is equicontinuous if for every $\varepsilon > 0$ there is a $\delta > 0$ such that $|f(x) - g(y)| < \varepsilon$ whenever $f$ and $g$ are in $S$ and $|x - y| < \delta$. For a proof see, for example, [1].

Theorem A.1. The closure of a uniformly bounded and equicontinuous set is compact. In particular, every sequence has a uniformly convergent subsequence.

The limit of the uniformly convergent subsequence may not be in the set $S$.

A.2 $C^1$ Convergence

Suppose that $\{f_n\}_{n=1}^{\infty}$ is a sequence of continuous functions defined on a compact set $\Omega$. If $\{f_n\}_{n=1}^{\infty}$ converges uniformly to $f$ and the sequence of its derivatives $\{Df_n\}_{n=1}^{\infty}$ is uniformly convergent to $g$, then $f$ is differentiable and $Df = g$. This theorem is proved in books on advanced calculus.
A.3 Existence, Uniqueness, and Continuous Dependence

If $f : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^k \rightarrow \mathbb{R}^n$ is a continuously differentiable function, $x_0 \in \mathbb{R}^n$, $t_0 \in \mathbb{R}$, and $\lambda_0 \in \mathbb{R}^k$, then the initial value problem

$$\frac{dx}{dt} = f(x, t, \lambda_0), \quad x(t_0) = x_0$$

has a unique solution defined for $t$ in some open interval containing $t_0$. More generally, there is a continuously differentiable function $\phi(t, \xi, \lambda)$ defined in some product neighborhood of $t_0$, $x_0$, and $\lambda_0$ such that $t \mapsto \phi(t, \tau, \xi, \lambda)$ is the solution of the initial value problem

$$\frac{dx}{dt} = f(x, t, \lambda), \quad x(\tau) = \xi$$

defined in some open interval containing $\tau$.

In short, smooth ordinary differential equations always have solutions that depend continuously on initial data and parameters. The solutions are as smooth as the function $f$. Solutions may be defined only for a short time. On the other hand, solutions exist until they reach the boundary of the (spatial) domain of definition of $f$, which might not be defined on all of $\mathbb{R}^n$ or until they blow up to infinity.

A.4 Green’s Theorem and Integration by Parts

**Theorem A.2** (Green’s Theorem). Let $\Omega$ be a bounded open subset of $\mathbb{R}^n$ whose boundary $\partial \Omega$ is a piecewise smooth hypersurface. If $F$ is a continuously differentiable function $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ defined on the closure of $\Omega$ (that is, $\Omega$ together with $\partial \Omega$), and $\eta$ is the outer unit normal on $\partial \Omega$, then

$$\int_{\Omega} \text{div } F \, dV = \int_{\partial \Omega} F \cdot \eta \, dS,$$

where $dV$ is the Euclidean volume and $dS$ is the surface area element induced on the hypersurface via the outer unit normal.

**Proof.** See a book on advanced calculus. \qed
Corollary A.3. With the same notation as in Green’s theorem, suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and $i$ is an integer in the range $1 \leq i \leq n$. Then,

$$\int_{\Omega} f x_i dV = \int_{\partial \Omega} f \eta_i dS.$$ 

Proof. Define a vector valued function $F$ whose component functions are all zero except for the $i$th component that is defined to be $f$ and apply Green’s theorem. □

Corollary A.4 (Integration by Parts). With the same notation as in Green’s theorem and the first corollary, suppose that $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable. Then,

$$\int_{\Omega} g x_i h dV = -\int_{\Omega} gh x_i dV + \int_{\partial \Omega} gh \eta_i dS.$$ 

Proof. Define $f$ to be the product of $g$ and $h$ and apply the first corollary. □

A.5 Gerschgorin’s Theorem

Let $A = a_{ij}$ be an $n \times n$ matrix and let $\Gamma_i$ denote the closed disk in the complex plane with center at $a_{ii}$ and radius $\sum_{j=1, j \neq i}^{n} |a_{ij}|$. The eigenvalues of $A$ are contained in the set $\bigcup_{i=1}^{n} \Gamma_i$. See [12] for a proof of this result and additional information.

A.6 Gram-Schmidt Procedure

Let $\{v_i\}_{i=1}^{N}$ be a basis for an inner product space $H$ whose inner product is denoted by angle brackets. An orthogonal basis consists of the set of vectors $\{g_i\}_{i=1}^{N}$ given by

$$g_1 := v_1, \quad g_i := v_i - \sum_{j=1}^{i-1} \frac{\langle v_i, g_j \rangle}{\langle g_j, g_j \rangle} g_j.$$ 

An orthonormal basis is given by $\{\frac{1}{|g_i|}g_i\}_{i=1}^{N}$, where

$$|g_i| = \sqrt{\langle g_j, g_j \rangle}.$$
A.7 Grobman-Hartman Theorem

A smooth system of (autonomous) differential equations is locally conjugate to its linearization at a hyperbolic rest point (all eigenvalues have nonzero real parts). See [15] for a proof.

A.8 Order Notation

Given two functions $f$ and $g$, we say that $f(x) = g(x) + O(x^n)$ if

$$\frac{|f(x) - g(x)|}{|x^n|}$$

is bounded by a positive constant for $|x|$ sufficiently close to zero. We say that $f(x) = g(x) + o(x^n)$ if

$$\lim_{x \to 0} \frac{|f(x) - g(x)|}{|x^n|} = 0.$$

For example, we have that

$$x - \sin x = O(x^3), \quad x - \sin x = o(x^2).$$

A.9 Taylor’s Formula

Taylor’s formula is an essential tool in mathematical analysis for approximating the values of functions. Suppose that $f : \mathbb{R}^m \to \mathbb{R}^n$ is smooth (at least $C^2$ for this appendix). Let $x$ and $h$ be elements of $\mathbb{R}^m$ and $t \in \mathbb{R}$. We have that

$$\frac{d}{dt} f(x + th) = Df(x + th)h.$$

After integration on the interval $0 \leq t \leq 1$, we have the identity

$$f(x + h) = f(x) + \int_0^1 Df(x + th)h \, dt;$$

or, equivalently,

$$f(x) = f(a) + \int_0^1 Df(x + t(x-a))(x-a) \, dt.$$
It follows immediately that
\[ f(x + h) = f(x) + Df(x)h + \int_0^1 (Df(x + th)h - Df(x)h) \, dt; \]
or, equivalently,
\[ f(x) = f(a) + Df(a)(x - a) + \int_0^1 (Df(x + t(x - a))(x - a) - Df(a)(x - a)) \, dt. \]

## A.10 Liouville’s Theorem

Suppose that \( t \mapsto \Phi(t) \) is a matrix solution of the homogeneous linear system \( \dot{x} = A(t)x \) on the open interval \( J \). If \( t_0 \in J \), then
\[
\det \Phi(t) = \det \Phi(t_0) e^{\int_{t_0}^t \text{tr} A(s) \, ds}
\]
where \( \det \) denotes determinant and \( \text{tr} \) denotes trace.

## A.11 Transport Theorem

Let \( \phi_t \) denote the flow of the system \( \dot{x} = f(x), \ x \in \mathbb{R}^n \), and let \( \Omega \) be a bounded region in \( \mathbb{R}^n \). Define
\[
V(t) = \int_{\phi_t(\Omega)} dx_1 dx_2 \cdots dx_n
\]
and recall that the divergence of a vector field \( f = (f_1, f_2, \ldots, f_n) \) on \( \mathbb{R}^n \) with the usual Euclidean structure is
\[
\text{div} \ f = \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}.
\]
Liouville’s theorem and the change of variables formula for multiple integrals can be used to prove that
\[
\dot{V}(t) = \int_{\phi_t(\Omega)} \text{div} \ f(x) dx_1 dx_2 \cdots dx_n.
\]
In particular, the flow of a vector field whose divergence is everywhere negative contracts volume.

Suppose that $g : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$ is smooth and, for notational convenience, let $dx = dx_1dx_2 \cdots dx_n$. The (Reynolds) transport theorem states that

$$\frac{d}{dt} \int_{\phi_t(\Omega)} g(x,t) \, dx = \int_{\phi_t(\Omega)} g_t(x,t) + \text{div}(gf)(x,t) \, dx. \quad (A.1)$$

A standard proof has three steps: The change of variables formula is used to freeze the integration over the set $\Omega$ so that the time-derivative may be taken across the integral sign; differentiation with respect to $t$ is carried out under the integral sign, Abel’s theorem is used to write the time-derivative of the Jacobian as the product of the trace of the derivative of the flow field (which is the divergence of $f$) and the Jacobian; and, the change of variables formula is applied to obtain the form given in the statement of the theorem.

A useful alternate form of the transport theorem is obtained by applying the divergence theorem to the second term in the integrand on the right-hand side of formula (A.1):

$$\frac{d}{dt} \int_{\phi_t(\Omega)} g \, dV = \int_{\phi_t(\Omega)} g_t \, dV + \int_{\partial \phi_t(\Omega)} g f \cdot \eta \, dS, \quad (A.2)$$

where $\eta$ is the outer unit normal on the boundary $\partial \phi_t(\Omega)$.

## A.12 Least Squares and Singular Value Decomposition

The basic problem of linear algebra is to solve for the unknown vector $x$ in the system of linear equations $Ax = b$, where $A$ is a matrix and $b$ is a vector. In case $A$ is a square matrix that is nonsingular (its determinant is not zero or its columns are linearly independent), there is a unique solution $x = A^{-1}b$. In general, the worst possible way to compute the solution $x$ is to compute the matrix inverse. Two central principles of numerical linear algebra state: never compute a determinant and never compute the inverse of a matrix. The best solution methods for linear systems are based on Gaussian elimination or iteration. Some iterative methods are discussed in this book.

There are important problems in applied mathematics that require a solution of $Ax = b$ in case $A$ is not a square matrix or $A$ is singular. The prime
example is linear regression where the matrix is generally not square: We are given a finite set of points in the plane \((x_i, y_i)\) for \(i = 1, 2, 3, \ldots, N\) and asked to find the best fitting line. More precisely, the problem is to find the line such that the sum of the squares of the deviations at the \(x_i\) from the line to \(y_i\) is minimized, that is,

\[
\min_{(m, b) \in \mathbb{R}^2} \sum_{i=1}^{N} |mx_i + b - y_i|^2.
\]

This problem may be recast into the abstract form

\[
\min_{x \in \mathbb{R}} |Ax - b|^2,
\]

where the vertical bars denote the Euclidean norm; \(A\) is the \(N \times 2\)-matrix with first column the transpose of the row vector \((x_1, x_2, x_3, \ldots, x_N)\) and second column the \(N\)-vector all of whose entries are one; \(b\) is the transpose of the row vector \((y_1, y_2, y_3, \ldots, y_N)\); and \(x\) is the column vector \((m, b)\). The corresponding matrix equation \(Ax = b\) is a prototypical example of an overdetermined system of linear equations (more equations than unknowns). This equation has a solution exactly when every coordinate pair \((x_i, y_i)\) lies on the same line. The purpose of linear regression is to find a line that best fits the data when the data points do not all lie on the same line.

The reason for the squares is to simplify the mathematics. For instance, the problem \(\sum_{i=1}^{N} \min_{m, b} |mx_i + b - y_i|\) is more difficult. The key point is that the Euclidean norm is defined by an inner product. In fact, using the usual inner (dot) product \(\langle v, w \rangle := \sum v_i w_i\), the length of a vector \(v\) is defined to be \(|v| = \sqrt{\langle v, v \rangle}\). The square of the length is just the inner product.

Assume that \(A\) is \(m \times n\) with \(m \geq n\) and define \(f : \mathbb{R}^n \to \mathbb{R}\) by \(f(x) = |Ax - b|^2 = \langle Ax - b, Ax - b \rangle\). We wish to minimize \(f\) over all of \(\mathbb{R}^n\). From calculus we know that the derivative of \(f\) must vanish at a minimum. To compute the derivative \(Df(x)\) abstractly, which is best in this case, recall that \(Df(x)\) is a linear transformation from \(\mathbb{R}^n\) to \(\mathbb{R}\). Let \(v\) denote an arbitrary vector in \(\mathbb{R}^n\). By the chain rule,

\[
\frac{d}{dt} f(x + tv) \bigg|_{t=0} = Df(x)v.
\]
Thus, we have that
\[
Df(x)v = \frac{d}{dt} \langle A(x + tv) - b, A(x + tv) - b \rangle \bigg|_{t=0}
\]
\[
= \langle Av, Ax - b \rangle + \langle Ax - b, Av \rangle
\]
\[
= 2\langle Ax - b, Av \rangle.
\]

Let us denote the matrix transpose (interchanging rows and columns) of \( A \) by \( A^T \). An important property (which is easy to check by a computation in components) is that \( \langle w, Az \rangle = \langle A^Tw, z \rangle \) for every pair of vectors \( w \) and \( z \). (Alternatively, we may define the matrix transpose of \( A \) to be the unique matrix \( A^T \) that satisfies the inner product identity and then prove that in coordinates \( A^T \) is obtained from \( A \) by interchanging its rows and columns). Using the transpose, it follows that
\[
Df(x)v = 2\langle A^TAx - A^Tb, v \rangle.
\]

Suppose that \( \langle u, v \rangle = 0 \) for every \( v \). Then, in particular, \( |u|^2 = \langle u, u \rangle = 0 \) and \( u = 0 \). Hence, if the derivative \( Df(x) \) is the zero matrix, then
\[
A^TAx - A^Tb = 0.
\]

This latter equation is called the normal equation for the least squares problem. Since \( A \) is \( m \times n \), the matrix \( A^TA \) is \( n \times n \).

For simplicity, let us make an additional assumption: The columns of \( A \) are linearly independent. In this case, \( A^TA \) is invertible. To prove this fact, suppose that \( v \) is a vector and \( A^TAv = 0 \). Taking the inner product with respect to \( v \), we have that
\[
0 = \langle A^TAv, v \rangle = \langle Av, Av \rangle = |Av|^2.
\]
Since the columns of \( A \) are linearly independent and \( Av \) is a linear combination of these columns, \( Av = 0 \) only if \( v = 0 \). This means that \( A^TA \) has linearly independent columns and hence is invertible.

Under our assumption that the columns of \( A \) are independent, the normal equation has a unique solution
\[
x = (A^TA)^{-1}A^Tb.
\]
Thus, our function \( f \) has a unique critical point, which must be a minimum because \( f(x) \geq 0 \) for every \( x \) and \( f(x) \) grows to infinity as \( |x| \) grows without bound.
The quantity \((A^T A)^{-1} A^T\) is called the pseudo inverse of \(A\). Using the pseudo inverse, the linear regression problem is solved: the transpose of the vector \((m, b)\) is exactly \((A^T A)^{-1} A^T b\) for the given \(N \times 2\) matrix \(A\) and the \(N\)-vector \(b\).

In practice, the pseudo inverse is not computed directly. The normal equations are solved by elimination or an iterative method; or, better yet, the pseudo inverse is computed using the singular value decomposition of \(A\). The reason for not simply solving the normal equations is that these equations may be ill conditioned, for example the matrix \(A^T A\) may be nearly singular.

The singular value decomposition of an \(m \times n\) real matrix \(A\), which always exists, has the form \(A = U \Sigma V^T\), where \(U\) is an \(m \times m\) orthogonal matrix (that is, \(U^T U = UU^T = I\)), \(\Sigma\) is an \(m \times n\) diagonal matrix, and \(V\) is an \(n \times n\) orthogonal matrix. The square roots of the diagonal elements of \(\Sigma\) are called the singular values of \(A\).

Using the singular value decomposition and the properties of its factors, the normal equation may be written in the form

\[
V \Sigma^T U^T U \Sigma V^T x = V \Sigma^T U^T b,
\]

and simplified to

\[
\Sigma^T \Sigma V^T x = \Sigma^T U^T b.
\]

In case the matrix \(A\) has linearly independent columns, its singular values are all positive. The matrix \(\Sigma^T \Sigma\) is square and its diagonal elements are the squares of the singular values of \(A\). Thus, the inverse of \(\Sigma^T \Sigma\) is diagonal with diagonal elements the reciprocals of its diagonal elements. An easy calculation shows that \(\Sigma^+ := (\Sigma^T \Sigma)^{-1}\Sigma^T\) is diagonal with diagonal elements the reciprocals of the singular values of \(A\). It follows that the least squares minimum is achieved at

\[
x = V \Sigma^+ U^T b.
\]

The matrix \(V \Sigma^+ U^T\) is the singular value decomposition pseudo inverse of \(A\). The ease of the inversion of \(\Sigma^T \Sigma\) and the efficiency of the numerical algorithms available to calculate the singular value decomposition make the method presented here the most used numerical method for computation of least squares problems. Of course, the singular value decomposition has many other applications.

Algorithms for the numerical computation of the singular value decomposition are presented in all books on numerical linear algebra.
A.13 The Morse Lemma

If a class $C^\infty$ function $f : \mathbb{R}^n \to \mathbb{R}$ has a nondegenerate critical point $a$ (that is, $Df(a) = 0$ and zero is not an eigenvalue of the symmetric linear transformation representing the quadratic form $D^2f(a)(x, x)$), then there is a $C^\infty$ change of coordinates defined in a neighborhood of $a$ that transforms $f$ to the function $\xi \mapsto f(a) + D^2f(a)(\xi, \xi)$.

A general proof of Morse’s lemma is given in [1]. The proof for the one-dimensional case is elementary. Reduce to the case where the function $f$ is given by $f(x) = x^2h(x)$ and $h(0) > 0$. Define $g(x) = x\sqrt{h(x)}$ and prove that there is a function $k$ such that $g(k(x)) = x$. The desired change of coordinates is given by $x = k(z)$.

A.14 Newton’s Method

Newton’s method can be used to solve systems of equations of the form

\[
\begin{align*}
  f_1(x_1, x_2, \ldots, x_n) &= 0, \\
  f_2(x_1, x_2, \ldots, x_n) &= 0, \\
  \vdots \\
  f_n(x_1, x_2, \ldots, x_n) &= 0.
\end{align*}
\]

Let $F : \mathbb{R}^n \to \mathbb{R}^n$ denote the function whose components are $(f_1, f_2, \ldots, f_n)$ and suppose that $r \in \mathbb{R}^n$ is a root (that is, $F(r) = 0$). The idea underlying Newton’s method is to guess an approximation $a \in \mathbb{R}^n$ of $r$ and refine this approximation using linearization. We must assume that $F$ is smooth (at least class $C^1$ to implement the method and at least $C^2$ for the method to perform as it should).

Note that (using the notation of Appendix A.8)

\[
0 = F(r) = F(a + (r - a)) = F(a) + DF(a)(r - a) + O((r - a)^2).
\]

From this equation, we would expect that $F(a) + DF(a)(r-a) \approx 0$. Since this expression is zero for $r$ replaced by $a - [DF(a)]^{-1}F(a)$, we might expect that this number is a better approximation to $r$ than $a$. Thus, we can iteratively improve our initial guess $x_0 := a$ by Newton’s Method:

\[
x_{n+1} = x_n - [DF(x_n)]^{-1}F(x_n).
\]
Newton’s method has a special feature that makes it very effective for an initial guess that is sufficiently close to a root of $F$: it is quadratically convergent. To see what this means, note first that an alternative way to view Newton’s method is to define the function $G : \mathbb{R}^n \to \mathbb{R}^n$ by

$$G(x) = x - [DF(x)]^{-1}F(x)$$

and observe that the initial guess is refined by iteration of $G$. Indeed, we have $x_{n+1} = G(x_n)$. If Newton’s method converges to the root $r$, it follows that $r = G(r)$; that is, $r$ is a fixed point of $G$. Now, suppose that $a$ is an approximation to $r$. If $G$ were simply a function with a fixed point at $r$, we always have (using Taylor’s formula in Appendix A.9) the identity

$$|G(a) - r| = |G(a) - G(r)| = \left| \int_0^1 DG(r + t(r-a))(r-a) \, dt \right|;$$

and, in case the derivative of $G$ is bounded by $M$, the estimate

$$|G(a) - r| \leq M|r-a|.$$

If we are lucky and $M < 1$, then $G(a)$ is closer to $r$ than $a$. In fact,

$$|G(G(a)) - r| \leq M^2|r-a|,$$

and so on. Hence, the iterates of $a$ by $G$ do indeed converge to $r$. The convergence rate is called linear because the ratio

$$\frac{|a_{n+1} - a_n|}{|a_n - a_{n-1}|}$$

is asymptotically constant (equal to $M$). For Newton’s method, the important observation is that $DG(r) = 0$, a fact that is easy to check. Thus, we have (using Taylor’s formula again)

$$|G(a) - r| = |G(a) - G(r)| = \left| \int_0^1 DG(r + t(r-a))(r-a) - DG(r)(r-a) \, dt \right|. \tag{A.3}$$

Under our smoothness assumption, we can apply Taylor’s formula to $DG$. If the second derivative of $G$ is bounded by $M$, then we have

$$|DG(r + t(r-a)) - DG(r)| \leq M|r-a|.$$
Using this estimate in equation (A.3), it follows that

\[ |G(a) - r| \leq M|a - r|^2. \]

In this case, the convergence is much faster: the ratio

\[ \frac{|a_{n+1} - a_n|}{|a_n - a_{n-1}|} \]

goes to zero (which means the numerator is much smaller than the denominator). The ratio

\[ \frac{|a_{n+1} - a_n|}{|a_n - a_{n-1}|^2} \]

is asymptotically constant.

For example, \( \{ \frac{1}{2^n} \}_{n=1}^{\infty} \) is linearly convergent and \( \{ \frac{1}{2^{2n}} \}_{n=1}^{\infty} \) is quadratically convergent.

An important result that gives sufficient conditions for the convergence of Newton’s method is the Newton-Kantorovich theorem. This result has several variants; two of them are stated here. A proof of the first version is given in [35]; the second version is proved in [38].

**Theorem A.5.** Suppose that \( F : \Omega \to \mathbb{R}^n \) is a differentiable function from the open subset \( \Omega \) in \( \mathbb{R}^n \) to \( \mathbb{R}^n \), \( \omega \in \Omega \), and the derivative \( DF(\omega) : \mathbb{R}^n \to \mathbb{R}^n \) is invertible. Define \( v = -[DF(\omega)]^{-1}(\omega)F(\omega) \) and \( U \) the ball of radius \( |v| \) centered at \( \omega + v \in \mathbb{R}^n \). If \( U \subset \Omega \), there is a positive number \( M \) such that

\[ \|DF(x) - DF(y)\| \leq M|x - y| \]

for all \( x \) and \( y \) in \( U \).

and

\[ M|F(\omega)|\|[DF(\omega)]^{-1}(\omega)\|^2 \leq \frac{1}{2}, \]

then \( F \) has a unique zero in \( U \) and for every starting point in \( U \) Newton’s method converges to this zero.

**Theorem A.6.** Suppose that

1) \( F : \Omega_r \to Y \) is a twice continuously differentiable function from the open ball \( \Omega_r \) of radius \( r > 0 \) of the Banach space \( X \) to the Banach space \( Y \);

2) \( A \) is a bounded linear operator \( A : X \to Y \);
3) \( \omega \in \Omega_r \) and there are positive constants \( \eta, \delta, \) and \( C \) such that

a) \( \|AF(\omega)\| \leq \eta, \)
b) \( \|ADF(\omega) - I\| \leq \delta < 1, \)
c) \( \|AD^2 F(x)\| \leq C \) for all \( x \in \Omega_r. \)

Define

\[
h = \frac{C \eta}{(1-\delta)^2}, \quad \alpha = \frac{(1-\sqrt{1-2h})\eta}{h(1-\delta)}, \quad \beta = \frac{(1+\sqrt{1-2h})\eta}{h(1-\delta)}.
\]

If \( \alpha \leq r < \beta \) and \( h < 1/2 \), or \( \alpha \leq r \leq \beta \) and \( h = 1/2 \), then \( F \) has a unique zero \( \omega_{\infty} \in \Omega_r \), the sequence \( \{\omega_n\}_{n=1}^{\infty} \) of Newton approximates converges to \( \omega_{\infty} \), and

\[
\|\omega_{\infty} - \omega_n\| \leq \frac{(2h)^{2^n} \eta}{2^n(1-\delta)}.
\]

In the second version of the theorem, \( A \) is usually taken to be \([DF(\omega)]^{-1}\) in case this operator is bounded.

### A.15 Variation of Parameters Formula

**Proposition A.7** (Variation of Parameters Formula). Consider the initial value problem

\[
\dot{x} = A(t)x + g(x, t), \quad x(t_0) = x_0 \quad (A.4)
\]

and let \( t \mapsto \Phi(t) \) be a fundamental matrix solution for the homogeneous linear system \( \dot{x} = A(t)x \) that is defined on some interval \( J_0 \) containing \( t_0 \). If \( t \mapsto \phi(t) \) is the solution of the initial value problem defined on some subinterval of \( J_0 \), then we have (the variation of parameters formula)

\[
\phi(t) = \Phi(t)\Phi^{-1}(t_0)x_0 + \Phi(t) \int_{t_0}^{t} \Phi^{-1}(s)g(\phi(s), s) \, ds. \quad (A.5)
\]

**Proof.** Define a new function \( z \) by \( z(t) = \Phi^{-1}(t)\phi(t) \). We have

\[
\dot{z}(t) = A(t)\phi(t) + \Phi(t)\dot{\phi}(t).
\]

Thus,

\[
A(t)\phi(t) + g(\phi(t), t) = A(t)\phi(t) + \Phi(t)\dot{\phi}(t)
\]

Consider the initial value problem

\[
\dot{x} = A(t)x + g(x, t), \quad x(t_0) = x_0 \quad (A.4)
\]

and let \( t \mapsto \Phi(t) \) be a fundamental matrix solution for the homogeneous linear system \( \dot{x} = A(t)x \) that is defined on some interval \( J_0 \) containing \( t_0 \). If \( t \mapsto \phi(t) \) is the solution of the initial value problem defined on some subinterval of \( J_0 \), then we have (the variation of parameters formula)

\[
\phi(t) = \Phi(t)\Phi^{-1}(t_0)x_0 + \Phi(t) \int_{t_0}^{t} \Phi^{-1}(s)g(\phi(s), s) \, ds. \quad (A.5)
\]

**Proof.** Define a new function \( z \) by \( z(t) = \Phi^{-1}(t)\phi(t) \). We have

\[
\dot{z}(t) = A(t)\phi(t) + \Phi(t)\dot{\phi}(t).
\]

Thus,

\[
A(t)\phi(t) + g(\phi(t), t) = A(t)\phi(t) + \Phi(t)\dot{\phi}(t)
\]
and

\[ \dot{z}(t) = \Phi^{-1}(t)g(\phi(t), t). \]

Also note that \( z(t_0) = \Phi^{-1}(t_0)x_0. \)

By integration,

\[ z(t) - z(t_0) = \int_{t_0}^{t} \Phi^{-1}(s)g(\phi(s), s) \, ds, \]

or, in other words,

\[ \phi(t) = \Phi(t)\Phi^{-1}(t_0)x_0 + \Phi(t)\int_{t_0}^{t} \Phi^{-1}(s)g(\phi(s), s) \, ds. \]

\[ \square \]

**A.16 The Variational Equation**

Let \( t \mapsto \phi(t, \xi, \lambda) \) denote the solution of the differential equation

\[ \dot{x} = f(x, t, \lambda) \]  \hspace{1cm} (A.6)

such that \( \phi(0, \xi, \lambda) = \xi \), where \( \lambda \) is a parameter in \( \mathbb{R}^k \), \( x \in \mathbb{R}^n \) and \( \xi \in \mathbb{R}^n \). The first (matrix) variational equation (also called the linearization) along the solution \( \phi \) is given by

\[ \dot{W} = D_xf(\phi(t, \xi, \lambda), t, \lambda)W \]

with initial condition \( W(0) = I \), where \( I \) is the \( n \times n \)-identity matrix and \( D_x \) denotes the derivative of \( f \) with respect to \( x \). The second variational equation is

\[ \dot{U} = Df(\phi(t, \xi, \lambda), t, \lambda)U + D\lambda(\phi(t, \xi, \lambda)) \]

with initial condition \( U(0) = 0 \), where \( D\lambda \) denotes the derivative of \( f \) with respect to \( \lambda \).

These equations are important for several reasons. The first variational equation is the differential equation for the derivative of \( \phi \) with respect to the initial condition; in fact,

\[ W(t) = D_\xi\phi(t, \xi, \lambda). \]

The second variational equation is the differential equation for \( \phi \) with respect to the parameter; in fact,

\[ U(t) = D_\lambda\phi(t, \xi, \lambda). \]
These facts follow immediately by differentiating both sides of the differential equation (A.6) with \( x \) replaced by \( \phi \). The key observations are that

\[
D_\xi D_t = D_t D_\xi \quad \text{and} \quad D_\lambda D_t = D_t D_\lambda,
\]

under the assumption that \( f \) is class \( C^1 \).

The first variational equation also arises as a linearization. Indeed, let \( \psi \) denote a solution of the differential equation (A.6) and consider its deviation \( \eta \) from the given solution \( \phi \); that is,

\[
\psi = \phi + \eta.
\]

We have that

\[
\dot{\eta} = \dot{\phi} - \dot{\psi} = f(\phi, t, \lambda) - f(\psi, t, \lambda) = f(\phi, t, \lambda) - f(\phi + \eta, t, \lambda) = D_x f(\phi, t, \lambda)\eta + O(\eta^2),
\]

where the last equation is obtained by expanding the function \( x \mapsto f(\phi + x, t, \lambda) \) in its Taylor series at \( x = 0 \). This suggests that the solution of the differential equation

\[
\dot{\eta} = D_x f(\phi(t, \xi, \lambda), t, \lambda)\eta
\]

provides a good approximation to the deviation \( \eta \) when \( \psi \) is close to \( \phi \). The initial condition is \( \eta(0) = \xi - \psi(0) \).

### A.17 Linearization and Stability

For a system on \( \mathbb{R}^n \) of the form

\[
\dot{u} = Au + g(u)
\]

(A.7)

where \( A \) is an \( n \times n \) matrix and \( g : \mathbb{R}^n \to \mathbb{R}^n \) is a smooth function such that \( g(0) = Dg(0) = 0 \), the origin \( u = 0 \) is a rest point whose linearization is \( \dot{w} = Aw \). The system matrix \( A \) of the linearized system can be used to determine the stability of the rest point.

**Theorem A.8.** If all the real parts of the eigenvalues of the system matrix of the linearization at a rest point of an autonomous system have negative real parts, then the rest point is (locally) asymptotically stable.
Theorem A.8 is a basic result. One method used to prove it is based on Lyapunov’s direct approach, which is the content of another basic result in stability theory for ordinary differential equations.

Consider a rest point \( x_0 \) for the autonomous differential equation

\[
\dot{x} = f(x), \quad x \in \mathbb{R}^n.
\]  

(A.8)

A continuous function \( V : U \to \mathbb{R} \), where \( U \subseteq \mathbb{R}^n \) is an open set with \( x_0 \in U \), is called a Lyapunov function for the differential equation (A.8) at \( x_0 \) if

(i) \( V(x_0) = 0 \),

(ii) \( V(x) > 0 \) for \( x \in U \setminus \{x_0\} \),

(iii) the function \( V \) is continuously differentiable on the set \( U \setminus \{x_0\} \), and, on this set, \( \dot{V}(x) := \nabla V(x) \cdot f(x) \leq 0 \).

The function \( V \) is called a strict Lyapunov function if, in addition,

(iv) \( \dot{V}(x) < 0 \) for \( x \in U \setminus \{x_0\} \).

**Theorem A.9 (Lyapunov’s Stability Theorem).** If there is a Lyapunov function defined in an open neighborhood of a rest point of the differential equation (A.8), then the rest point is stable. If, in addition, the Lyapunov function is a strict Lyapunov function, then the rest point is asymptotically stable.

For system (A.7), assume that every eigenvalue of \( A \) has negative real part, and for some \( a > 0 \), there is a constant \( k > 0 \) such that (using the usual norm on \( \mathbb{R}^n \))

\[
|g(x)| \leq k|x|^2
\]

whenever \( |x| < a \). To show that the origin is an asymptotically stable rest point, it suffices to construct a quadratic Lyapunov function. To do this, let \( \langle \cdot, \cdot \rangle \) denote the usual inner product on \( \mathbb{R}^n \), and \( A^* \) the transpose of the real matrix \( A \). Suppose that there is a real symmetric positive definite \( n \times n \) matrix that also satisfies Lyapunov’s equation

\[
A^*B + BA = -I
\]

and define \( V : \mathbb{R}^n \to \mathbb{R} \) by

\[
V(x) = \langle x, Bx \rangle.
\]
The restriction of $V$ to a sufficiently small neighborhood of the origin is a strict Lyapunov function. To see this, make an estimate using the Schwarz inequality. To complete the proof show that

$$B := \int_0^\infty e^{tA^*} e^{tA} dt$$

is a symmetric positive definite $n \times n$ matrix which satisfies Lyapunov’s equation. This claim is proved with a few observations. First, note that $A^*$ and $A$ have the same eigenvalues, all of which are in the open left half of the complex plane. There is some number $\lambda > 0$ such that all eigenvalues of both matrices have real parts less than $-\lambda$. In this case,

$$\|e^{tA} x\| \leq C e^{-\lambda t} \|x\|$$

for all $t \geq 0$ and all $x \in \mathbb{R}^n$. This estimate is used to prove that the integral converges.

Alternatively, we can solve Lyapunov’s equation using the following outline: Lyapunov’s equation in the form $A^* B + B A = S$, where $A$ is diagonal, $S$ is symmetric and positive definite, and all pairs of eigenvalues of $A$ have nonzero sums, has a symmetric positive-definite solution $B$. In particular, under these hypotheses, the operator $B \mapsto A^* B + B A$ is invertible. The same result is true without the hypothesis that $A$ is diagonal. This fact can be proved using the density of the diagonalizable matrices and the continuity of the eigenvalues of a matrix with respect to its components.

**Theorem A.10 (Routh-Hurwitz Criterion).** Suppose that the characteristic polynomial of the real matrix $A$ is written in the form

$$\lambda^n + a_1 \lambda^{n-1} + \cdots + a_{n-1} \lambda + a_n,$$

let $a_m = 0$ for $m > n$, and define the determinants $\Delta_k$ for $k = 1, 2, \ldots, n$ by

$$\Delta_k := \det \begin{pmatrix} a_1 & 1 & 0 & 0 & 0 & \cdots & 0 \\ a_3 & a_2 & a_1 & 1 & 0 & \cdots & 0 \\ a_5 & a_4 & a_3 & a_2 & a_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{2k-1} & a_{2k-2} & a_{2k-3} & a_{2k-4} & a_{2k-5} & a_{2k-6} & \cdots & a_k \end{pmatrix}.$$ 

If $\Delta_k > 0$ for $k = 1, 2, \ldots, n$, then all roots of the characteristic polynomial have negative real parts.
A.18 Poincaré-Bendixson Theorem

Two versions of the Poincaré-Bendixson and an important corollary are stated here (see, for example, [15]).

**Theorem A.11.** If a closed and bounded positively invariant set for a $C^1$ autonomous system in the plane contains no rest points, then it contains at least one limit cycle.

**Theorem A.12.** If a closed bounded positively invariant set in the plane for a $C^1$ autonomous system contains no periodic orbits and exactly one hyperbolic asymptotically stable rest point (the real parts of the eigenvalues of the system matrix of the linearization at the rest point are negative), then every solution starting in the invariant set is asymptotic to the rest point.

**Theorem A.13.** A periodic orbit for a $C^1$ autonomous system in the plane surrounds at least one rest point.
Appendix B

Answers to Selected Exercises

Page 8

Ex. 2.11 (g) The answer is probably not known. But, with a high level of confidence, it seems that the fate of the solution with initial data $x(0) = 0$ and $\dot{x}(0) = 1$ is convergence to the constant solution $x(t) = 1$.

Ex. 2.10 $f(\phi, \theta) = (A \sin \theta + B \cos \theta)(\frac{1-\cos\theta}{\sin\theta})$. After separation of variables we find the ODE $\sin \phi p'' + \cos \phi p' - p = 0$, which can be written in the form $(\sin \phi p')' = \frac{p}{\sin \phi}$. This suggests the substitution $u = \sin \phi p'$, etc.

Page ??

Ex. ?? $\phi \approx 5.8629$, $e \approx 0.41279$, $a \approx 1.65967$ AU, and the initial speed is $35.23465$ km/sec.

Page 79

Ex. 4.39 The bifurcation diagram is depicted in Figure B.1.

Page 122
Ex. 4.55 Suppose that it takes $N$ steps to reach $L$. The global error is proportional to $Nh^{n+1}$. This quantity is equal to

$$N \left( \frac{L}{N} \right) \left( \frac{N}{L} \right) h^{n+1} = L \frac{1}{h} h^{n+1} = L h^n.$$ 

Ex. 4.69 The data was generated using $a = 0.001$ and $b = 1.23$ by rounding off the values of $f$ to two decimal places.

Page 126

Ex. 4.69 The data was generated using $a = 0.44$ and $b = 888$ by rounding off the values of $f$ to three decimal places.

Page 126

Ex. 4.110 The dynamical system is given by the mass balance: rate of change of mass equals mass in minus mass out. For $f(t)$ the mass flow rate into the tank and $\rho A \sqrt{2g} z$ the mass flow rate through the drain (where $A$ is the area of the drain cross section), the mass balance differential equation is

$$\frac{d}{dt} (\rho \pi a^2 z) = f(t) - \rho A \sqrt{2g} z.$$
The valve control mechanism actuates the control by changing the area \( A \) of the drain cross section. There does not seem to be an obvious way to implement the control; that is, to define the radius of the drain cross section as a function of the proportional error \( k(z_{\text{set}} - z) \). This function would be used to run the actuator that moves the valve mechanism. Control design requires experience and ingenuity. One possible actuation function is

\[
A = \pi \left( \frac{r}{2\pi} \left( \pi - \arctan(k(z_{\text{set}} - z)) \right) \right)^2,
\]

where the controller gain is assumed to be positive. When the water level exceeds the set-point depth, the argument of the arctangent is negative and approaches \(-\pi\) as the depth of the water in the tank increases; thus, the area of the drain cross section increases toward \( r \). The opening rate of the orifice increases as the control gain \( k \) increases or the water depth increases. Likewise, the area decreases as the depth or control gain decreases. Note: We have implemented the inherently linear \( P \) control as the argument of a nonlinear function! By assumption, the mass inflow rate is less than \( \rho \pi r^2 \sqrt{2gh} \), which with the given choice of parameters is approximately \( 491 \text{ kg} / \text{m} \).

Numerical experiments show that it is not possible to maintain the set point depth for an inflow of 300 kg/sec. Why not? In practice, it is usually better to analyze the system before doing numerical experiments. But, the results of computation might suggest that something in the control design is not working as expected.

The problem is apparent by looking at the steady state(s) of the control system. The steady state of the model is not at the set point. For the given data, the steady state is \( z_{\text{SS}} \approx 7.23 \). Thus, there is good reason to incorporate an \( I \) control. Because \( I \) controls are explicitly time-dependent their presence makes the control process non autonomous; no steady states exist unless the depth is at the set point over the entire time the integral controller is functioning.
Ex. ?? Using the predictor-corrector scheme (??), compute the local truncation error in three steps: Apply Taylor series expansions to obtain

\[ y(t + \Delta t) = y(t) + y'(t)\Delta t + \frac{y''(t)}{2}\Delta t^2 + O(\Delta t^3) \]

\[ = y(t) + f(y(t))\Delta + \frac{1}{2}f'(y(t))f(y(t))\Delta t^2 + O(\Delta t^3) \]

and

\[ y_{n+1} = y_n + \Delta t f(y_n + \frac{\Delta t}{2} f(y_{n-1/2})) \]

\[ = y_n + \Delta t f(y_n) + \frac{1}{2}f'(y_n)f(y_{n-1/2})\Delta t + O(\Delta t^3), \]

estimate the error by

\[ |y(t + \Delta t) - y_{n+1}| = \frac{1}{2}|f'(y_n)f(y_n) - f'(y_n)f(y_{n-1/2})|\Delta t^2 + O(\Delta t^3), \]

and, use the mean value theorem, to estimate the important difference

\[ |f(y_n) - f(y_{n-1/2})| = |f'(\xi)||y_n - y_{n-1/2}| \leq |f'(\xi)|(\|y_n - y_{n-1}\| + \|y_{n-1} - y_{n-1/2}\|). \]

The predictor-corrector formulas for the previous step can be used to show that \( |y_n - y_{n-1}| \) and \( |y_{n-1} - y_{n-1/2}| \) are \( O(\Delta t) \).

Page ??

Ex. ?? The push forward of the vector field \((c, 0)\) by \(Q\) is

\[ \left( \frac{c\sigma}{2(\sigma^2 + \tau^2)} , -\frac{c\tau}{2(\sigma^2 + \tau^2)} \right). \]

This vector field is not divergence free; therefore it is not a steady state solution of the incompressible Euler’s equations.
We have that
\[ \rho \dot{v}_i = (\lambda \varepsilon_{kk} \delta_{ij} + 2\mu \varepsilon_{ij}) \frac{1}{2}(u_{k,kj} + u_{k,kj}) \delta_{ij} + \mu(u_{i,ij} + u_{j,ij}) + \rho b_i \]
\[ = \lambda u_{k,kj} \delta_{ij} + \mu(u_{i,ij} + u_{j,ij}) + \rho b_i \]
\[ = \lambda u_{k,ki} + \mu(u_{i,ij} + u_{j,ij}) + \rho b_i \]
\[ = \lambda u_{j,ij} + \mu(u_{i,ij} + u_{j,ij}) + \rho b_i. \]

The quantity $\nabla u$ is a $3 \times 3$-matrix. Also recall that $\nabla \dot{u} = \nabla^T u$. These two facts are all that is needed to show that $\nabla \cdot (\nabla u) = u_{j,ij}$.

**Ex. ??** To prove that the set $\mathcal{L}$ of square integrable functions on the interval $[0, L]$ is not finite-dimensional, let us suppose that $\mathcal{L}$ has a generating set consisting of $n \geq 1$ elements. Recall that $e^k := \sin \frac{k \pi x}{L}$ is in $\mathcal{L}$ for each positive integer $k$. For the case $n = 1$, suppose there is a function $g^1$ that generates every element in $\mathcal{L}$. Then there are two nonzero numbers $a$ and $b$ such that $e^1 = ag$ and $e^2 = bg$. The functions $e^1$ and $e^2$ are orthogonal. Thus, $(ag, bg) = 0$. Equivalently, $ab\|g\|^2 = 0$. But this is impossible because $\|g\| > 0$.

Suppose there is a generating set with $n > 1$ elements, say $\{g^i\}_{i=1}^n$. There are real numbers $\{b_\ell\}_{\ell=1}^n$ such that $e^{n+1} = \sum_{\ell=1}^n b_\ell g^\ell$. By renumbering the elements in the generating set if necessary and noting that $e^{n+1} \neq 0$, we may assume that $b_1 \neq 0$. Also, there is an $n \times n$ array of numbers $a_{ij}$ such that $e^i = \sum_{j=1}^n a_{ij} g^j$, for $i = 1, 2, 3, \ldots n$. The function $e^{n+1}$ is orthogonal to each such $e^i$.

Hence, we have that
\[
0 = \left( \sum_{\ell=1}^n b_\ell g^\ell, \sum_{j=1}^n a_{ij} g^j \right) = \sum_{\ell=1}^n \sum_{j=1}^n b_\ell a_{i,j}(g^\ell, g^j) = \sum_{j=1}^n a_{i,j} b_j.
\]

By rearranging, dividing by $b_1 \neq 0$, and defining $c_j = -b_j/b_1$, for $j = 2, 3, 4, \ldots, n$, we obtain the identities
\[
a_{i1} = \sum_{j=2}^n c_j a_{ij}, \quad i = 1, 2, 3, \ldots, n.
\]
Note that, by substitution for $a_{i1}$,

$$e^i = \sum_{j=1}^{n} a_{ij}g^j = (\sum_{j=2}^{n} c_j a_{ij})g^1 + \sum_{j=2}^{n} a_{ij}g^j = \sum_{j=2}^{n} a_{ij}(c_j g^1 + g^j).$$

Thus, the set of $n$ independent vectors $\{e^i\}_{i=1}^{n}$ is generated by the set $G_{n-1}$ of $n-1$ functions

$$c_2 g^1 + g^2, \quad c_3 g^1 + g^3, \quad c_4 g^1 + g^4, \ldots, c_n g^1 + g^n.$$  

The element $e^n$ is orthogonal to each $e^i$, for $i = 1, 2, 3, \ldots, n-1$. There must be at least one nonzero function in $G_{n-1}$. Thus, by renumbering if necessary, we may assume that the first coefficient of the linear combination of functions in $G_{n-1}$ equal to $e^n$ is not zero. The same argument used to obtain $G_{n-1}$ can be used to show that the set of $n-1$ independent vectors $\{e^i\}_{i=1}^{n-1}$ is generated by a set $G_{n-2}$ of $n-2$ functions. Continuing in this manner, we may conclude that $e^1$ and $e^2$ are generated by one function, which we have shown to be impossible. This contradiction completes the proof.

Page ?? Ex. ?? A graph of the solution of the boundary value problem $u_{xx} + \sin(x)u = 1$ with $u(0) = 0$ and $u(2\pi) = 0$ is depicted in Figure B.2.
Ex. ?? The miss distance is the same as the miss distance in the lead angle plane, as it should be.
Bibliography


207


