Introduction to Mathematical Modeling

INTRODUCTION TO MATHEMATICAL MODELING

JOCELINE LEGA

CONTENTS

Part V. Appendices

PREFACE

What are mathematical models? How are they developed? How much should we trust them: can we be confident in their predictions, and do we know when to act upon them? These questions, routinely addressed by mathematical modeling practitioners, are also of interest to citizens and policy makers. Our modern society needs models that can be relied upon, not only to improve our understanding of a situation, but also to inform policy decisions.

This introduction to mathematical modeling was developed for an audience of college seniors pursuing an undergraduate degree in mathematics with emphasis in applied mathematics, the life sciences, or engineering. The course builds on knowledge of calculus, linear algebra, and differential equations to address the basic techniques and thought processes that are fundamental to mathematical modeling. The style is deliberately casual and the main goal is to explain how mathematics learned in core undergraduate classes may be used to understand simple phenomena that arise in physics and biology, and how the corresponding models are put together, tested, and analyzed.

The text covers all of the standard systems that are normally considered in a modeling class: the nonlinear pendulum, chaotic maps, predator-prey models, competing species, chemical reactions, and, towards the end, diffusion and spatially extended systems. None of these are complicated topics and one could argue that such models are too simple to be useful. They however form the building blocks of mathematical modeling and, in spite of their simplicity, provide the tools to tackle more elaborate and realistic models. Emphasis is placed on developing practice with simple but general methods, such as dimensional analysis, phase plane analysis, basic fixed point theory, and numerical explorations; whenever possible, connections between different systems are built by exploring similarities in the mathematical models that describe them. Although some sections involve randomness, most of the text is concerned with deterministic models based on difference or differential equations. This is a deliberate choice, in order to allow coverage of the material in a one semester course. Finally, because modelers need to be good communicators of science and should understand potential uses and abuses of mathematical models, the first chapter of the text discusses such issues, in the context of a few examples.

Many excellent texts on dynamical systems are available in the literature, some of which motivate the study of nonlinear systems through mathematical models. One may thus question the usefulness of a separate course on mathematical modeling. The point of view presented here is that mathematical modeling is the art of using one's mathematical knowledge to describe the world in mathematical terms. This requires good reasoning skills and a solid understanding of mathematical methods, as well as a type of mathematical fluency that transcends expertise in differential equations, or in any other core mathematics subject. The purpose of this text is to develop these skills and associated mindset through the practice of mathematical modeling in the context of simple, carefully chosen examples. Appendices are provided to review the basic mathematical tools needed to build and analyze the models. MATLAB GUIs are supplied with the course materials, allowing readers to explore the role of model parameters through graphical user interfaces, without requiring knowledge of numerical methods or of a particular numerical software package.

This book will give readers the background necessary to follow general scientific research articles that use mathematical modeling, such as those found in Science, Nature, and PNAS, to name a few. Successive versions of these notes have been used since 2005 as the main text for a one-semester capstone course at the University of Arizona. Students who take the mathematical modeling course also work in teams on a semester-long project, under the supervision of graduate or post-graduate mentors. Each project is based on understanding and reproducing the results of a research article. Teams write midterm and final reports on their projects and present their work in a poster session held in a public venue at the end of the semester. For the online version of the course, posters are replaced with group video presentations, judged by members of the university community. It is highly recommended that a similar model be used when teaching a class with this text. A list of recent projects and related articles is provided as an appendix.

I would like to thank all of my colleagues in the Department of Mathematics at the University of Arizona who, since 2007, have taught our mathematical modeling course with this text. I am also grateful to all of the graduate and postgraduate mentors, who for almost two decades have guided teams of undergraduates taking this course through their modeling projects. Finally, the initial development of these notes was made possible thanks to a University of Arizona TRIF (Technology and Research Initiative Fund) grant, which is acknowledged with great appreciation.

Joceline Lega
The University of Arizona
Fall 2012 & Summer 2024

PART I
INTRODUCTION

The first two chapters of these notes aim to introduce basic concepts of mathematical modeling. Particular emphasis is placed on the modeling process, which involves successive testing and refinement of a mathematical model.

Chapter 1 introduces mathematical models and the modeling process, discusses the usefulness of mathematical models, and considers the range of skills mathematical modelers should possess. The exercise section encourages the reader to reflect on how models should or should not be used.

The purpose of Chapter 2 is to build a mathematical model step by step, following the modeling process. Most readers will be familiar with the phenomenon to be modeled: a human wave in a stadium.

1.

ON THE NATURE OF MATHEMATICAL MODELING

Learning Objectives

At the end of this chapter, you will be able to do the following.

- Describe what a mathematical model is.
- Describe what it takes to be a good mathematical modeler.
- Explain how to build a mathematical model by going through the modeling cycle.
- Discuss the role of hypotheses in the model-building process and how they limit potential applications of the model.
- Reflect on the appropriate use of mathematical models.

What is a model?

The word *model* refers to a representation, often simplified, of an object or of an observation. A model is thus expected to at least mimic, but preferably explain and predict relevant aspects of a given phenomenon. A *mathematical model* typically consists of one or more equations relating *dependent* (or output) to *independent* (or input) variables. Often, a mathematical model involves *parameters*.

This definition raises the question of *reproduction* versus *explanation*. Consider for instance the formation of ripples on a sandy beach. One could imagine developing a model which would take into account how each particle of sand interacts with its neighbors, with the surrounding air, and with particles at rest on the ground. This would be a fairly complicated model, which would involve many dynamical equations, but which should, under appropriate circumstances, be able to reproduce the formation of sand ripples. Another approach could be at a more global level. One could then imagine a model that considers the elevation of the sand on the ground and shows that, in some parameter regime, this quantity oscillates as a function of space. Both models would reproduce the desired phenomenon, but would provide different types of explanation. This is due to the difference in their nature: the former is a model at the *microscopic* or particulate level, the latter is a model at the *macroscopic* level. A third model could be a sort of "black box," whose input would be relevant parameters such as the speed of the wind, and the mass and size of the particles of sand, and whose output would be a stripe pattern with the correct properties (period, height, etc). One could create such a model by conducting careful experiments, tabulating the results, and finding functions that fit the data. This model would not offer an explanation as to why sand ripples form, but could still have some predictive capabilities.

Along the same lines, it is well-known that mountains, clouds, trees or even some artistic works resemble fractals. Recognizing this fact may help graphic designers create convincing and aesthetically pleasing virtual landscapes, but does not for instance explain how mountains or clouds are formed.

Different models may thus produce similar results, and modelers need to decide which model best suits their needs. This is of course a subjective task, the outcome of which should be based on the answers to the following questions.

- *What do we want the model to do: reproduce, explain, predict, etc?*
- *Is a predictive but costly model worth the effort invested in its creation?*
- *What minimal set of criteria should the model satisfy?*

A related issue is that different phenomena may lead to similar behaviors (a little like different diseases may have similar symptoms). As a consequence, a model which reproduces the desired behaviors may not have any explanatory or predictive value (similarly, one cannot reasonably diagnose a disease based on symptoms that are not unique to this particular disease).

In this text, we will not consider curve-fitting approaches (although see the exercises for applications of the least squares method), but focus our attention on models that are *descriptive and explanatory*. Such models will often consist of a collection of dynamical equations. We will discuss how to analyze them, in order to assess their ability to reproduce and predict relevant phenomena. We will assume the reader has a working knowledge of calculus, differential equations and linear algebra. Brief reviews of these topics are provided in the appendices.

How to develop a mathematical model

We now turn to the basic steps involved in the creation of a mathematical model. We will follow these guidelines in all of the models discussed here, very explicitly at first, and then in a more implicit fashion.

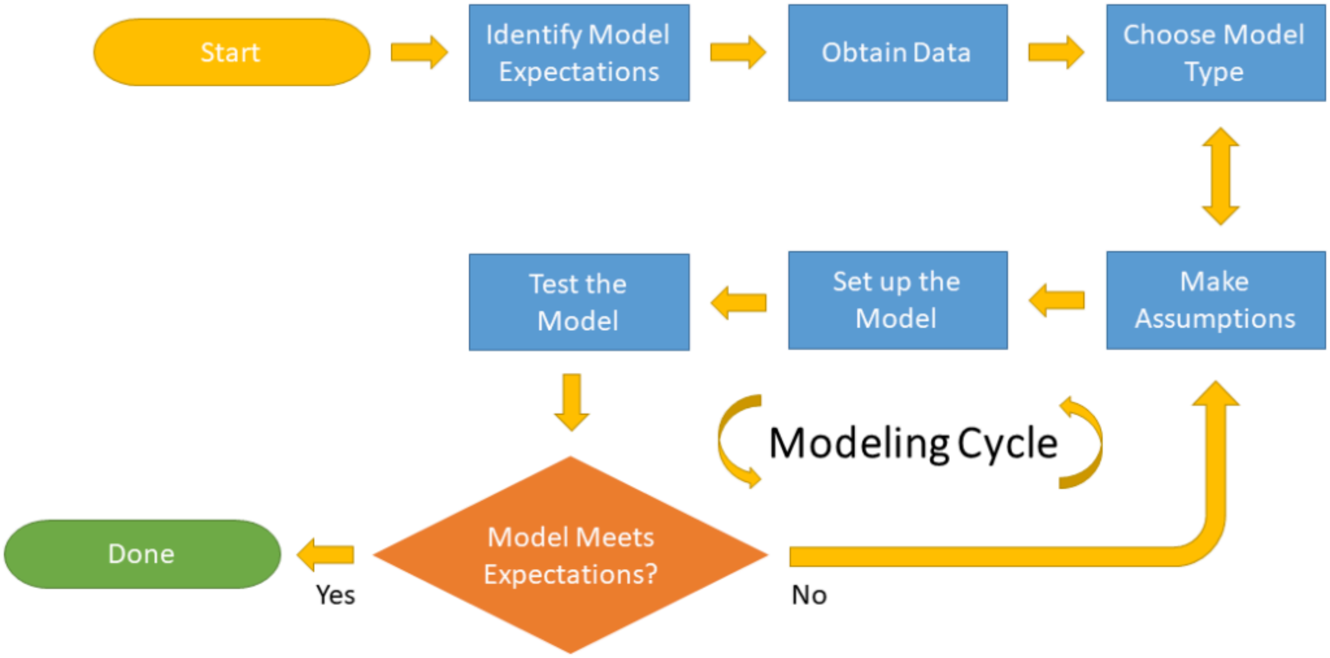


Figure 1.1. The Mathematical Modeling Process and the Modeling Cycle.

Identify what we want to model

The first step is to decide what the model should accomplish. In other words, we need to find a set of criteria that a satisfactory model should meet. These criteria typically result from balancing the desire of having a perfect model against any time and cost constraints associated with creating such a model.

Obtain data

This is often the best way to develop an intuitive understanding of the phenomenon of interest. If appropriate, we should also master the basic physical, biological or chemical principles that are responsible for the observed behavior. This may be fairly involved, and often requires finding information in the literature and/or discussing the data with experts.

Choose the level of complexity of the model

Going back to the example of the formation of sand ripples discussed above, we should for instance decide whether a microscopic model should be preferred to a macroscopic model. Such a decision may be based on personal taste, on the nature of available data, and on the current level of understanding of the phenomenon.

Make assumptions

We have to decide which facts (or parameters) are irrelevant and therefore negligible, and which ones have to be taken into account.

Set up the model

We need to define the independent and dependent variables, define the relevant parameters, and write a mathematical formulation of the model. This formulation should be as simple as possible, and we should have a clear understanding of the significance of each of the terms that appear in the equations forming the mathematical model.

Test the model

This is a crucial step in deciding whether we are satisfied with our model. The behavior of the model may be investigated analytically, or numerically, or both. We then have to ask ourselves whether our intuition generally agrees with what the model does, whether the model is qualitatively and quantitatively correct, and whether limiting cases make sense. If not, we need to find why and amend the model accordingly – in other words, we need to go back to the previous steps. This is thus *an iterative process*, illustrated in Figure 1.1, during which the model is perfected by testing the results of our analysis of the model against our intuition, against available experimental data, or against the accepted understanding of the phenomenon under study.

Why develop mathematical models

A mathematical model is a set of equations that reflects our understanding of a given phenomenon. It may look mysterious to the neophyte, but for someone who knows how to *read* a model, it is an efficient and concise means of communication. A lot of effort, thinking, testing and understanding goes into the creation of a model, but in the end, all this work is summarized by a set of equations or by the formulation of some iterative process. Such a model is a scientists' expression of their perception of the world.

Models have also more practical purposes. They may be used to describe or investigate limiting situations which cannot be reached in practice. In particular, the different ingredients of a model may be isolated from one another by setting all parameters but a few to zero. Models also allow us to save time and resources. For instance, computer simulations of car crashes are cheaper than performing actual experiments. Flight simulators are used as training devices. Models may be run as part of feasibility studies, or to test the outcome of various possible scenarios.

Because models are increasingly used to make decisions that impact society (see for instance the news articles listed in Problem 2 at the end of this chapter), it is important to know their limitations, to realize which hypotheses they are based upon, and to know the role, which may be crucial, that these hypotheses play in the model. An article by R. May, entitled *Uses and Abuses of Mathematics in Biology*, further reflects on such issues.

What does it take to be a good modeler?

Mathematical modeling combines various areas of mathematics (for instance differential equations, calculus, linear algebra, numerical analysis, probability theory, dynamical systems, statistics, uncertainty quantification) with, depending on the nature of the model, knowledge of physics, biology, chemistry, astrophysics, geology, hydrology, etc. These subjects are typically taught independently from one another, and the main instructional goal of a mathematical modeling course is to discuss how to draw on various areas of knowledge in order to build and analyze a model. As a consequence, a reader well trained in mathematics and other scientific disciplines will find that there is no “new material” in these notes. What is new is the use of a variety of mathematical tools to reach a single objective: develop, understand, and test mathematical models.

Successful mathematical modeling thus requires some sort of resourcefulness, since the modeler should be able to “think” of the right tools or methods to use. It also requires some curiosity, imagination and patience; an ability to understand problems that are not purely mathematical; an ability to simplify a given problem, in order to decide what is relevant and what is not; an ability to turn concepts into equations, to decide whether a model is good or inadequate; and finally an ability to discuss problems with others, particularly those whose expertise is different from one’s own.

Summary

The phrase *mathematical modeling* may be understood in a variety of ways. In these notes, only models that are explanatory, predictive, and have a mathematical formulation qualify as mathematical models. In particular, curve fitting techniques or statistical models are not discussed.

The modeling process involves a series of steps that modelers need to go through in a systematic fashion. These include knowing what one wants to model, getting data, deciding on the type of model to be developed, making assumptions, constructing the model, testing and amending it, and finally using the model. It is particularly important that simplifying hypotheses made in the development of the model be clearly understood, especially if the results are used for policy or decision making purposes.

Mathematical modeling is an interdisciplinary activity which requires an appreciation for the power of mathematical analysis, an interest in applied disciplines, a solid mathematical and computational background, as well as good communication skills.

Food for Thought

Problem 1

Read the article by R. May entitled *Uses and Abuses of Mathematics in Biology*. How does this article illustrate the dangers of trusting models blindly? Do you find the arguments convincing?

Problem 2

Read at least three of the news articles below.

- *Rapid Response Could Have Curbed Foot-and-Mouth Epidemic* by Martin Enserink
- *Disease control: Virtual plagues get real* by V. Gewin
- *African swine fever outbreak alarms wildlife biologists and veterinarians* by Erik Stokstad
- *U.K. expands kill zone for badgers in fight against bovine TB, sparking controversy* by Erik Stokstad
- *Italy's olive crisis intensifies as deadly tree disease spreads* by Alison Abbott
- Why computer simulations should replace animal testing for heart drugs, by Elisa Passini, Blanca Rodriguez, and Patricia Benito
- *Could computer models be the key to better COVID vaccines?* by Elie Dolgin
- *OFF THE GRID: Computer models that forecast overloaded power lines are holding back U.S. solar and wind energy projects*, by Dan Charles

Then answer the following questions; include a discussion of the articles that you read (with proper references) in your argumentation.

- Do you believe mathematical models should be used to make decisions that impact people and society? Why or why not?

- What criteria do you think such models should satisfy, if they are to be used for policy-making purposes?

Problem 3

Assume you are an important business or government person, and that you have to make a decision that will influence the future of many people. Your decision relies on simulations of a mathematical model. What kind of questions would you ask the developers of the model, in order to help you in your decision-making process?

Problem 4

Read the article by N. Goldenfeld and L. Kadanoff, entitled *Simple Lessons from Complexity*. Summarize the main points of this paper and indicate which of the authors' conclusions you think one should especially keep in mind when developing a model.

Problem 5

Consider a set of data points in the plane $D = \{(X_i, Y_i),\ i = 1, \ldots n\}$, and a straight line of equation $y = ax + b$. Define the distance δ between the data set D and the set $F = \{(X_i, f(X_i)),\ i = 1, \ldots n\}$ as $\delta = \sum_{i=1}^{n} (f(X_i) - Y_i)^2$.

Show that one can find a pair (a, b) which minimizes the value of δ. This particular choice of parameters provides a linear fit, $y = ax + b$, of the data set D. This fitting technique is called the *least squares method*.

Problem 6

Using the least squares method (see Problem 5), find the straight line that best fits the following data points: (1,3.3), (2.5, 8), (4,13), (7,22), (8,25.5). Plot the line and the data points on the same graph. Is the fit satisfactory? Why or why not?

Problem 7

Browse the MATLAB documentation on curve fitting, and familiarize yourself with MATLAB's basic fitting interface. Create a set of data points close to the graphs of the following functions, and see if the fit proposed by MATLAB is satisfactory. What do you conclude?

1. $f(x) = 3x + 5.$
2. $f(x) = 7x.$
3. $f(x) = \exp(3x + 7).$
4. $f(x) = \exp(7x).$

2.

FIRST STEPS: MODELING THE WAVE

Learning Objectives

At the end of this chapter, you will be able to do the following.

- Apply the modeling process by developing a model from the ground up.
- Translate word statements into mathematical formulations.
- Test the behavior of a model with a numerical simulation.
- Assess whether a mathematical model is behaving as expected.

We will start our exploration of mathematical modeling with a simple problem, which does not require any particular *a priori* knowledge, but which allows us to illustrate the modeling process discussed in the previous chapter.

Formulation of the problem

We will consider the problem of the *human wave*, which occurs in a stadium when spectators stand up, raise their hands and sit down, in order to form a wave which propagates around the bleachers. According to Farkas *et al.*, such a phenomenon is often called the Mexican Wave (La Ola), a term that was first used during the broadcasting of the 1986 Soccer World Cup held in Mexico. The discussion below is based on two articles by Farkas *et al.* entitled *Mexican waves in an excitable medium* and *Human waves in stadiums*.

Our goal is to develop a model that reproduces and explains the propagation of the wave. The model will have to depend on parameters that can be described in everyday terms, and the behavior of the model when these parameters are changed should correspond to what is actually observed in a stadium.

Obtain data and choose the level of complexity of the model

The two articles by Farkas *et al.* referenced above provide some information, based on 14 video recordings of waves in stadiums containing more than 50,000 seats. Their observations indicate that waves propagate on average at a speed of 12 m s^{-1} (or equivalently 22 seats per second), and have a width between 6 and 12 meters (or equivalently 11 to 22 seats), with an average of about 15 seats. Moreover, approximately 3 out of 4 waves propagate clockwise, when one looks at the stadium from above.

Our model will be as simple as possible. In particular, it will be one-dimensional, and discrete. We will however carefully define our hypotheses so that we will know what to modify if other effects need to be included.

Make assumptions

We will use periodic boundary conditions, since each row in a stadium forms a closed loop. We will not take into account the presence of stairs, and thus assume that all of the seats are equally spaced. We will also consider that all seats are occupied. Each spectator will be either seated or *doing the wave*. This latter state corresponds to consecutively standing up, waving one's hands, and sitting down. Once a person has started doing the wave, they will not stop until they are seated again. At that time, the person may decide to stand up and do the wave again.

Set up the model

First, the wave needs to be initiated: some people must be standing up for the wave to start. Then, for the wave to propagate, spectators must be influenced by what their neighbors do. Since most waves propagate in a clockwise direction, people must be more sensitive to what is happening on their right than on their left.

It is clear that the mood of the spectators is important: if the game is too boring, no one will feel like cheering. On the other hand, if the game is too exciting, there may be people standing up all the time, and many waves could be initiated at once. Based on this, we can describe the wave as *a wave of enthusiasm*. People standing up are *excited* and their behavior influences the level of enthusiasm of their neighbors. If this level is above a person's *threshold of enthusiasm*, then the person will become *excited* and start doing the wave as well.

We will therefore need a parameter, s_{th}, that describes the mood, or the threshold of enthusiasm, of each spectator. This parameter is likely to vary from one spectator to the next, and we will simply assume that it is uniformly distributed over an interval $[c-\delta, c+\delta]$. More complicated distributions could be used. The parameter c describes the average responsiveness of the crowd watching the game, and $\delta/\sqrt{3}$ measures the spread or the standard deviation of this responsiveness about its mean.

We will assume that a person's level of enthusiasm varies as they do the wave. This is reasonable since it seems legitimate to consider that a person who is standing and waving is more excited than a person who is just getting up or sitting down. We will thus introduce a function $f(t)$ that describes the level of enthusiasm of a person doing the wave, as a function of time.

Finally, we will define an asymmetric function, w which describes how a spectator is influenced by the behavior – in our case the level of enthusiasm – of their neighbors. We now discuss how to convert the above statements into equations.

Let N be the number of seats in one row of the stadium (recall that this is a one-dimensional model. It can however easily be extended to two dimensions, as is the case for the models discussed in the articles by Farkas *et al.*). Since we assumed that all seats are occupied, we can define a function $x(i)$ describing the level of enthusiasm of spectator i, for $i = 1 \cdots N$. The variable x will be either 0, if the person is seated, or larger than some activity threshold $\mathcal{A}$. If $x(i) = 0$, then spectator i will start doing the wave only if the combined enthusiasm of their neighbors, denoted by $w(i)$, exceeds $s_{th}(i)$, i.e. exceeds their threshold of enthusiasm. Once $x(i)$ becomes non-zero, this variable will evolve as a function of time as follows:

$$x(i,t) = \begin{cases} f(t) & \text{if } f(t) > \mathcal{A} \\ 0 & \text{if } f(t) \leq \mathcal{A} \end{cases}. \qquad (2.1)$$

The function $f(t)$ should be bell-shaped, its maximum corresponding to the moment when the person is standing up with their hands up. Many functional forms can be chosen; here, we will simply take $f(t) = 1/\cosh(b(t - t_0))$, where t_0 is the maximizer of f and b is a parameter which selects the duration τ of the wave, that is how long it takes for one person to stand up, raise their hands, and sit down. More precisely, since the function $1/\cosh(t)$ is of order one on an interval centered at the origin and of width 10, τ can be approximated by

$$\tau \simeq \frac{10}{b}.$$

We will choose t_0 such that $t_0 = t_\mathcal{I} + \Delta\tau$, where $t_\mathcal{I}$ is the time at which a person becomes excited, i.e. at which $w(i)$ exceeds $s_{th}(i)$, and $\Delta\tau$ is a parameter that we can adjust. One can think of $\Delta\tau$ as a reaction time, which measures the time elapsed between the moment a person notices their neighbors' activity and the moment they stand up and do the wave.

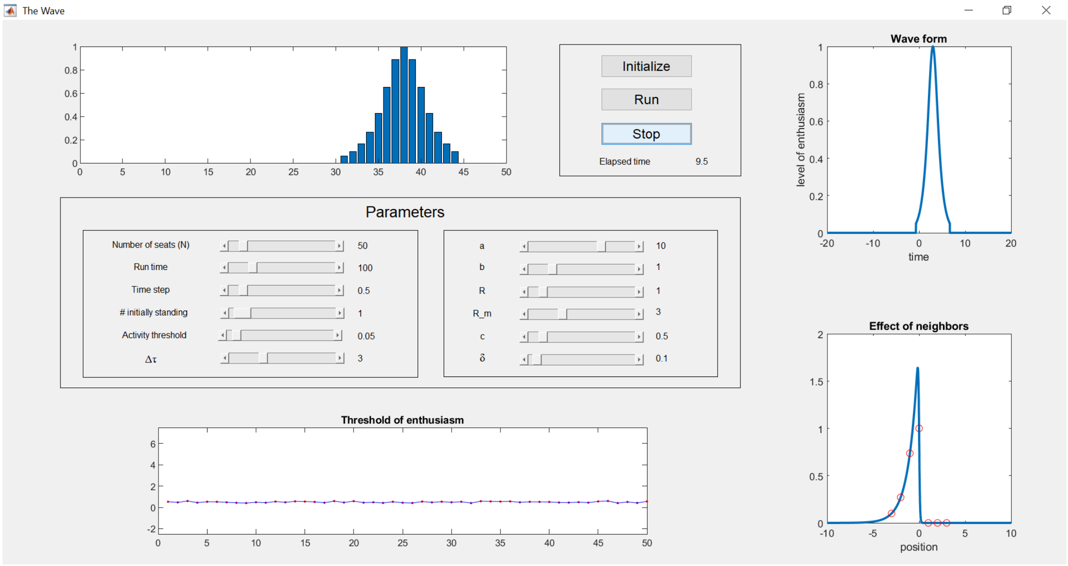
Since observation of human waves indicate that most waves propagate clockwise, we will use an asymmetric function to describe the combined level of enthusiasm of a person's neighbors. Again, many choices are possible. Here, we use

$$w(i) = \sum_{\substack{j,\ |i-j| \le R_m \\ x(j) > \mathcal{A}}} e^{-|i-j|/R} \ (1 - \tanh(a(j-i)))\,,$$

where a, R and R_m are parameters. This function is such that spectator i is affected by excited neighbors seating less than $R_m + 1$ seats away. The exponential term indicates that the most important influence comes from nearest neighbors, and the hyperbolic tangent term is such that neighbors on the right of person i are more influential than neighbors on the left (for positive values of a). When looking at the one-dimensional row from the center of the stadium, the more influential neighbors are thus seated left of person i, leading to a wave propagating clockwise.

Test the model

The best way to test this model is by numerical simulation. Once we have chosen the number of seats, the number of people initially standing and the enthusiasm threshold of each of the spectators, and we have defined the functions f and w, we can iteratively apply the rule given by Equation (2.1) to each spectator. Whether the wave propagates or dies will depend on the parameters of the model, and we can explore different situations by changing these parameters.



Elapsed time 9.5

Number of seats (N) 50
Run time 100
Time step 0.5
# initially standing 1
Activity threshold 0.05
$\Delta\tau$ 3

a 10
b 1
R 1
R_m 3
c 0.5
δ 0.1

Figure 2.1. MATLAB Graphic User Interface (GUI) for the Mexican Wave model.

The MATLAB code The_Wave.m simulates the above model. Figure 2.1 shows the corresponding MATLAB interface. The various parameters may be modified by adjusting sliders. When the simulation runs, a plot of $x(i)$ is shown in the top left window, and is updated every time step, t_r. The bottom left window displays a spectator's threshold of enthusiasm, $s_{th}(i)$, as a function of their seat number i. The top right panel shows the graph, as a function of t, of the level of enthusiasm $x(i,t)$ (see Equation (2.1)) of a person doing the wave. The bottom right panel illustrates how neighbors affect the behavior of a person. More precisely, it shows the graph of

$$e^{-|i-j|/R}\ (1-\tanh(a(j-i)))$$

as a function of $j-i$; the red circles indicate the contribution of spectator j sitting within a distance R_m of spectator i. All of these windows are updated when the parameters are changed by the user. Default parameter values are such that a wave about 15 seats wide propagates clockwise. Below are a few suggestions for exploring the model.

- Change the number of seats and check that the size of the wave is not affected, provided N is larger than the width $\mathcal{W}$ of the wave. What happens if N is less than $\mathcal{W}$?
- Change the threshold of enthusiasm of the spectators, by modifying the mean c and the range $[c-\delta, c+\delta]$ of s_{th}.
- Change the values of R and R_m. What do you observe?
- Is there a relation between the minimum value of c and the minimum value of R for a wave to form? Why or why not? What is the significance of such a relation?
- What is the speed of the wave?
- Change the time step t_r. How does it affect the size of the wave? Can you modify the value of b to compensate? Why or why not? (Hint: what other time scale is there in the model?) What does t_r actually select? Why?
- Change the activity threshold $\mathcal{A}$. Do you understand why the wave dies when $\mathcal{A}$ is too large? What other parameter can you modify to prevent this from happening?
- Change the number of people initially standing. What do you observe?
- What can you change to make the wave propagate to the left (i.e. anti-clockwise)?

The above numerical exploration shows that the model gives a reasonable representation of the formation and propagation of a wave in a stadium. But the simulation can also help identify some of the limitations of the model. For instance, stairs or aisles could block the propagation of the wave. Another limitation is that if there is more than one person initially standing up, all of these spectators sit down in exactly the same fashion. This is because we use the same function $f(t)$ for each spectator. In practice, different people do the wave differently, and one could think of letting the function f depend on i.

Even the simple version of the model presented here has a large number of parameters. A numerical exploration of the model indicates that it is not the parameters themselves, but combinations thereof, which are in fact relevant. For instance, we could decide to measure time in units of the time step, t_r. Then, we would be left with only two dimensionless parameters, τ/t_r and $\Delta\tau/t_r$. Identifying relevant combinations of parameters is essential for a thorough exploration of the properties of any model. We will discuss this at length in the next chapter, when we introduce dimensional analysis and scalings.

Summary

This chapter illustrates the various steps involved in the modeling process, using the example of a human wave in a stadium. These steps consist in formulating the problem, obtaining data, identifying the level of complexity of the desired model, making simplifying assumptions, setting up the model and finally testing it. The tools we used were elements of calculus, to choose the functional forms of f and w, and numerical simulation. The latter helped us test the model, but simulations are also useful to build an intuitive understanding of the properties of any model. This knowledge can then be used to identify the ingredients responsible for each particular property of a model, and possibly simplify or modify the model accordingly. The phenomenon discussed in this chapter is an example of wave propagation in an *excitable medium*.

A system is said to be excitable if perturbations of large enough amplitude can trigger a fast and large response, followed by a slower relaxation of the system back to its resting state. Other examples of excitable media include neurons (nerve impulses are excitable waves – called action potentials – propagating along axons) and cardiac tissue (a heartbeat corresponds to an electric wave propagating through the heart, leading to contraction and relaxation of the heart muscle).

Food for Thought

Problem 1

Find three different *bell-shaped* or *pulse-like* functions. How do you adjust their height? How do you adjust their width?

Problem 2

Find a monotone, differentiable function $f(x)$ defined on the real line, and such that $\lim_{x\to-\infty} f(x) = -6, \lim_{x\to\infty} f(x) = 3.$

Plot the function you found and check that it is monotonically increasing from -6 (as $x \to -\infty$) to 3 (as $x \to +\infty$).

Problem 3

Plot the function $f(t) = 1/\cosh(bt)$ for various values of b. Describe in words the role of this parameter.

Problem 4

Plot the function $w(y) = e^{-|y|/R}\,(1 - \tanh(ay))$ for different values of the parameters R and a. Describe in words the role of each parameter.

Problem 5

How would you modify the model described in this chapter in order to take into account the presence of aisles or stairs in a stadium?

Problem 6

Read the articles by Farkas *et al.* entitled *Mexican waves in an excitable medium* and *Human waves in stadiums*.

1. Do you find their model convincing?
2. Is all of the needed information actually included in the articles? Why or why not?
3. Are the hypotheses made by the authors essential? Explain.

Problem 7

What are the similarities and differences between the model discussed in this chapter and a one-dimensional version of the detailed n-state model of Farkas *et al.* entitled *Mexican waves in an excitable medium* and *Human waves in stadiums*? Justify your answer.

PART II
MODELS FROM CLASSICAL MECHANICS

The goal of this section is to familiarize ourselves with the use of differential equations as models, to discuss scalings and dimensional analysis, and to introduce phase plane techniques. This will be done on examples of application of classical mechanics.

In the first chapter, we will go over the derivation of the equation of motion of the nonlinear pendulum, check the dimension of each term in this equation, discuss the limit of small oscillations, and investigate the nonlinear dynamics of the pendulum by means of phase plane analysis.

We will then consider the problem of stone-skipping, and introduce Euler's equation for the dynamics of a rigid body. This discussion will also provide a good illustration of the power of dimensional analysis.

3. THE NONLINEAR PENDULUM

Learning Objectives

At the end of this chapter, you will be able to do the following.

- Apply the modeling process to a simple mechanical system, the nonlinear pendulum.
- Use Newton's law to derive a differential equation for the dynamics of the pendulum.
- Combine variables and parameters into dimensionless quantities.
- Assess whether a mathematical model is dimensionally correct.
- Synthesize the dynamics of an autonomous second order differential equation by means of energy methods or phase plane analysis.

Nature of the problem, assumptions and model equations

We consider the motion of a pendulum, that is of a mass suspended at one end of a string, the other end being attached to a fixed point (see Figure 3.1). It is known that the period of the pendulum depends on its amplitude. We refer the reader to Figure 4 of a 2005 article by Lima and Arun, and references therein.

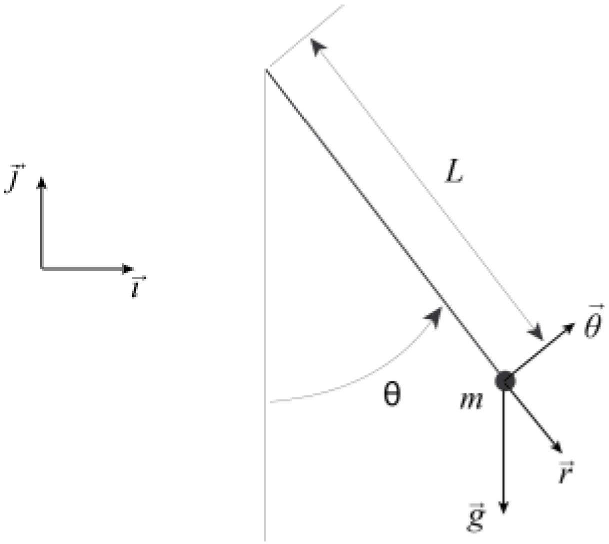


Figure 3.1. Sketch of a pendulum.

In what follows, we will assume that the object attached to the string is a *point mass*, that the string is massless and that neither the mass of the object nor the length of the string change with time. We will consider that motion takes place in a plane (note this is an approximation since the pendulum tends to undergo a slow precession motion). We will use *Newton's law* to describe the dynamics of the pendulum. The forces acting on the mass are the force of gravity $\vec{F}_g$, which we will consider constant (see exercises for a justification), the tension $\vec{F}_t$ exerted by the string, and the friction force $\vec{F}_f$ exerted on the mass by the surrounding air. We will assume that this force is proportional to the negative of the velocity of the mass. This too is an approximation; for more information, see for instance *Pendulum Damping* by P. Squire (1986).

In order to set up the problem, we define a basis of vectors $(\vec{\imath}, \vec{\jmath})$ for the plane in which the motion is taking place. We call l the length of the string and m the mass of the object attached to the string. We measure the position of the point mass in terms of the angle θ between the string and the vertical (see Figure 3.1). The position vector of the point mass $\vec{x}$ is given by

$$\vec{x} = l\sin(\theta)\,\vec{\imath} - l\cos(\theta)\,\vec{\jmath} \equiv l\vec{r},$$

where $\vec{r} = \sin(\theta)\vec{\imath} - \cos(\theta)\vec{\jmath}$. The unit vector orthogonal to $\vec{r}$ and such that $(\vec{r}, \vec{\theta})$ forms a direct basis of the (x, y) plane is $\vec{\theta} = \cos(\theta)\,\vec{\imath} + \sin(\theta)\,\vec{\jmath}$.

We can now use this information to express the first and second derivatives of the position vector $\vec{x}$ in terms of the vectors $\vec{r}$ and $\vec{\theta}$, as well as of the derivatives of the angle θ. Indeed, we have

$$\frac{d\vec{x}}{dt} = (l\cos(\theta)\,\vec{\imath} + l\sin(\theta)\,\vec{\jmath}\,)\,\frac{d\theta}{dt} = l\frac{d\theta}{dt}\vec{\theta} \tag{3.1}$$

$$\begin{aligned} \frac{d^2\vec{x}}{dt^2} &= l\frac{d^2\theta}{dt^2}\vec{\theta} + (-l\sin(\theta)\,\vec{\imath} + l\cos(\theta)\,\vec{\jmath}\,)\left(\frac{d\theta}{dt}\right)^2 \\ &= l\frac{d^2\theta}{dt^2}\vec{\theta} - l\left(\frac{d\theta}{dt}\right)^2\vec{r}. \end{aligned} \tag{3.2}$$

Newton's law tells us that the acceleration of the point mass is equal to the sum of the forces applied to the mass. In other words, we have

$$m\frac{d^2\vec{x}}{dt^2} = \vec{F}_g + \vec{F}_t + \vec{F}_f, \qquad (3.3)$$

where

- $\vec{F}_g = -mg\,\vec{\jmath} = mg\cos(\theta)\,\vec{r} - mg\sin(\theta)\,\vec{\theta}$ is the force of gravity,
- $\vec{F}_t = -T\,\vec{r}$ is the tension exerted by the string,
- $\vec{F}_f = -c\dfrac{d\vec{x}}{dt} = -cl\dfrac{d\theta}{dt}\,\vec{\theta}$, with $c > 0$, is the friction force (we used Equation (3.1) to obtain the last equality).

Note that we have expressed the forces in terms of the vectors $\vec{r}$ and $\vec{\theta}$, and not $\vec{\imath}$ and $\vec{\jmath}$. The reason is that the expression for $d^2\vec{x}/dt^2$ is simple if written in terms of $\vec{r}$ and $\vec{\theta}$ (see Equation (3.2)). Putting all of this information together, we obtain

$$ml\frac{d^2\theta}{dt^2}\,\vec{\theta} - ml\left(\frac{d\theta}{dt}\right)^2\vec{r} = mg\cos(\theta)\,\vec{r} - mg\sin(\theta)\,\vec{\theta} - T\,\vec{r} - cl\frac{d\theta}{dt}\,\vec{\theta},$$

which, by projecting onto the $\vec{r}$ and $\vec{\theta}$ directions, gives the system of equations

$$-ml\left(\frac{d\theta}{dt}\right)^2 = mg\cos(\theta) - T \qquad (3.4)$$

$$ml\frac{d^2\theta}{dt^2} = -mg\sin(\theta) - cl\frac{d\theta}{dt}. \qquad (3.5)$$

Equation (3.4) gives an expression for the tension T in terms of the angle θ and its derivative, and Equation (3.5), which does not involve T, is a *nonlinear* ordinary differential equation for the angle θ. This equation describes the motion of the *nonlinear pendulum*. It is typically complemented with *initial conditions*, which give the angle and angular velocity of the pendulum at a particular time, say $t = 0$. These initial conditions read

$$\theta(0) = \theta_0, \qquad \frac{d\theta}{dt}(0) = \Omega_0, \qquad (3.6)$$

where θ_0 and Ω_0 are known.

Analysis in the absence of friction

In the absence of friction, Equation (3.5) can be re-written as

$$\frac{d^2\theta}{dt^2} = -\frac{g}{l}\sin(\theta) = -\omega_0^2 \sin(\theta), \qquad (3.7)$$

where we set $\omega_0 = \sqrt{g/l}$.

Dimensional analysis

The first step in analyzing a model is to perform some *dimensional analysis*, in order to check that the various terms appearing in the equation(s) that form the model have the same dimension. In what follows, we will use the following standard notation:

- T (not to be confused with the tension above) will represent quantities that have the *dimension* of a time. Note that such quantities may be expressed in different *units*, such as seconds, minutes, days, years, etc.
- M will represent quantities that have the dimension of a mass. The corresponding units may be grams, kilograms, etc.
- L will represent quantities that have the dimension of a length. Here again, such quantities my have different units, such as meters, feet, yards, kilometers, etc.

It is thus important to distinguish between *dimension* and *units*. Consider for instance the terms appearing in Equation (3.7). The angle θ is dimensionless (note that this does not mean it has no units, since typically angles are measured in radians or degrees). Quantities which appear as arguments of functions (such as θ in the $\sin(\theta)$ term of Equation (3.7)) must be dimensionless. This is something that we always have to check, each time we are faced with a mathematical expression. Second, the left-hand-side of Equation (3.7) has the dimension of inverse time square. We thus write (the bracket notation indicating "dimension")

$$\left[\frac{d^2\theta}{dt^2}\right] = T^{-2}.$$

The right-hand-side of Equation (3.7) must have the same dimension. Since functions, such as $\sin(\theta)$ are dimensionless, we conclude that

$$\left[\frac{g}{l}\right] = T^{-2},$$

which we must now check. To do so, we need to find the dimension of the acceleration of gravity g. As indicated by its name, g has the dimension of an acceleration, i.e. $[g] = LT^{-2}$. Since $[l] = L$, we see that indeed, $[g/l] = T^{-2}$. As a consequence, $[\omega_0] = T^{-1}$, i.e. ω_0 is a frequency, as expected. Equation (3.7) is therefore dimensionally correct.

Scalings

Before starting to analyze a differential equation, it is often useful to rescale it, that is rewrite it in a form that has as few parameters as possible. The reason is that, quite often, quantities that control the dynamics of the problem are not the parameters themselves, but combinations thereof. For instance, Equation (3.7) tells us that the combination g/l is the relevant quantity to describe the motion of a frictionless pendulum. As a consequence, there is only one relevant parameter, ω_0, instead of two (g and l). We can even go one step further. Since ω_0 has the dimension of an inverse time, we can define a characteristic time

$$t_0 = \sqrt{\frac{l}{g}},$$

and define a *dimensionless* time variable τ as

$$\tau = \frac{t}{t_0}.$$

By substituting these relations in Equations (3.7) and (3.6), we obtain

$$\frac{d^2\theta}{d\tau^2} = -\sin(\theta), \qquad (3.8)$$

together with the initial conditions

$$\theta(0) = \theta_0, \qquad \frac{d\theta}{d\tau}(0) = t_0\Omega_0. \qquad (3.9)$$

In other words, the motion of the frictionless nonlinear pendulum can be described by an ordinary differential equation, (3.8), which has no parameters! This is a very important result since it will dramatically simplify our investigation of the nonlinear pendulum: if we can completely characterize the dynamics of Equation (3.8), then we are done with our analysis of Equation (3.7); there is no need to vary parameters!

Small oscillations: the harmonic oscillator

If the pendulum undergoes small oscillations, that is if θ is small, we can write a Taylor expansion of the expression $\sin(\theta)$ in powers of θ and obtain a simplified equation. In particular, if we keep only the first term in the expansion, Equation (3.8) becomes the equation for the harmonic oscillator. Indeed, at lowest order, $\sin(\theta) \simeq \theta$, and substitution into Equation (3.8) gives

$$\frac{d^2\theta}{d\tau^2} + \theta = 0. \qquad (3.10)$$

The general solution of this equation is

$$\theta(\tau) = A\cos(\tau) + B\sin(\tau) \qquad (3.11)$$

or equivalently (see exercises)

$$\theta(\tau) = C\cos(\tau + \phi), \qquad (3.12)$$

where A, B, C and ϕ are constants. If we now impose the initial conditions (3.9), we get

$$\theta_0 = \theta(0) = A, \qquad t_0\Omega_0 = \frac{d\theta}{d\tau}(0) = B,$$

or

$$\theta_0 = \theta(0) = C\cos(\phi), \qquad t_0\Omega_0 = \frac{d\theta}{d\tau}(0) = -C\sin(\phi).$$

The first system of equations gives A and B directly, whereas one needs to solve the second system for C and ϕ (see exercises) in order to have an expression for $\theta(\tau)$. Writing the solution as (3.11) makes it easier to impose the initial conditions, but Equation (3.12) makes it easier to describe the dynamics of the solution. We indeed see that the angle θ oscillates as a function of time between the values $C = \sqrt{\theta_0^2 + t_0^2\Omega_0^2}$ and $-C$, with a period equal to 2π (in the scaled variable τ). The pendulum oscillates indefinitely, which is as expected since we assumed that the dynamics is frictionless.

Nonlinear dynamics

Equation (3.8) can be solved explicitly in terms of elliptic functions, but writing such an expression would not necessarily help us describe the motion of the pendulum. We will thus not try to solve this equation, but instead qualitatively describe its dynamics in the corresponding phase space. This may seem surprising to readers who did not take a course on dynamical systems. My hope is to convince you, on a few examples throughout this text, that this is a most general and efficient way of dealing with nonlinear differential equations.

Our goal will be to describe the solutions of (3.8) in the *phase plane* $(\theta, d\theta/d\tau)$. Each initial condition $(\theta_0, t_0\Omega_0)$ will be associated with a curve in this plane. [1] If we know the arrangement of these solution curves, we can give a complete qualitative description of the dynamics of the nonlinear pendulum. An expression for the solution curves can be obtained as follows. If we multiply Equation (3.8) by $d\theta/d\tau$ and integrate, we obtain

$$\frac{1}{2}\left(\frac{d\theta}{d\tau}\right)^2 = \cos(\theta) + E,$$

where the constant E is arbitrary. It represents the energy of the dimensionless pendulum, and is conserved by the dynamics. The set of solution curves is thus described by

$$\frac{d\theta}{d\tau} = \pm\sqrt{2\cos(\theta) + 2E}, \qquad \frac{d\theta}{dt} \in \mathbb{R}, \qquad E \in [-1, \infty). \qquad (3.13)$$

If $E > 1$, then the left hand side of Equation (3.13) is real for all values of θ, and the corresponding solution curves are not closed. On the other hand, if $-1 \leq E \leq 1$, $d\theta/dt$ is real only for values of θ in intervals of the form $[-\arccos(E) + 2m\pi, \arccos(E) + 2m\pi]$, $m \in \mathbb{Z}$, and the corresponding solution curves are closed orbits. Finally, because of the $\pm$ sign in the right hand side of (3.13), trajectories are symmetric with respect to the horizontal axis of the phase plane. We can use software such as Maple or MATLAB to plot these curves, paying particular attention to the special cases $E = -1$ and $E = 1$. Figure 3.2 shows some of these trajectories. The description of the dynamics is completed by adding arrows to the curves, indicating the direction in which each trajectory is followed. This is not a difficult task since by definition $d\theta/d\tau$ is positive if θ increases as a function of τ and negative otherwise. As a consequence, the arrows point to the right in the upper part of the phase plane, and to the left in the lower part.

[1] See the appendix on ordinary differential equations for the existence and uniqueness of solutions to initial value problems, and for a brief description of phase plane analysis.

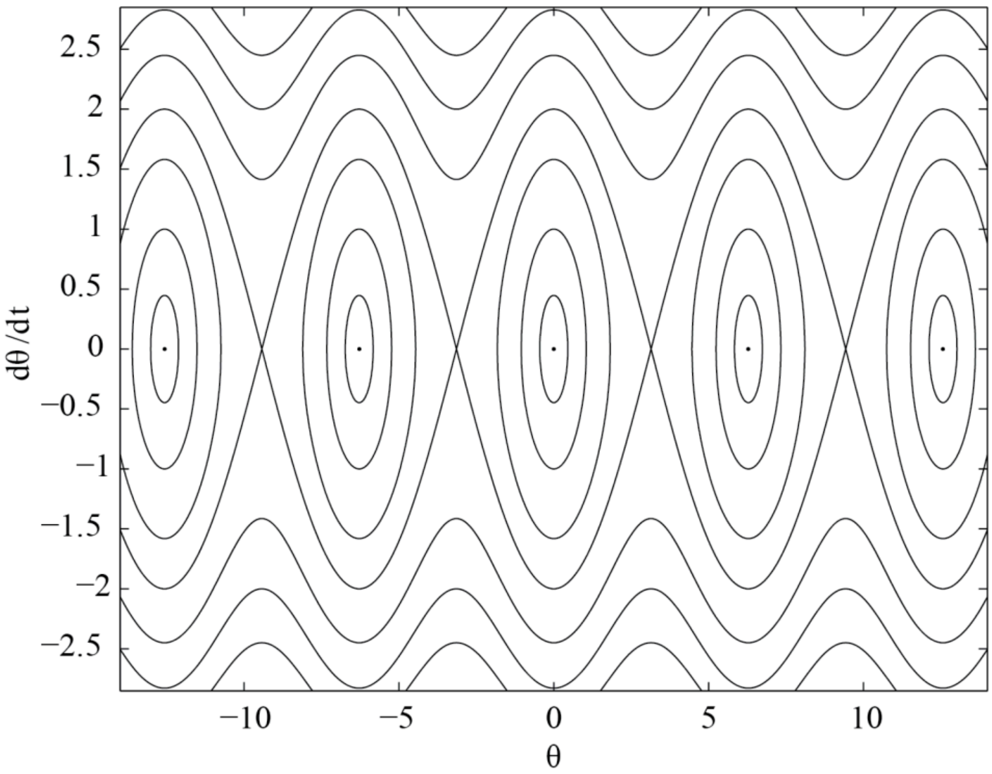


Figure 3.2. Solution curves of the frictionless nonlinear pendulum.

These results can be confirmed by using a phase plane analysis software such as the Phase Plane App. This program can be run in MATLAB. It allows the user to enter Equation (3.8) written as a first order system,

$$\frac{d\theta}{d\tau} = \Lambda, \qquad \frac{d\Lambda}{d\tau} = -\sin(\theta),$$

and to plot the corresponding direction field and trajectories in the phase plane. The direction field for this system is obtained by plotting vectors with components $(1, -\sin(\theta))$ (i.e. the right-hand-sides of the above equations) at points of coordinates (θ, Λ) in the phase plane. The solution curve of (3.8) that goes through the point (θ, Λ) is tangent to the direction field at that point. Figure 3.3 shows the phase plane for Equation (3.8), as obtained with PPLANE (developed by John C. Polking at Rice University). The solution curves found numerically are, as expected, in excellent agreement with those shown in Figure 3.2, which were obtained from Equation (3.13).

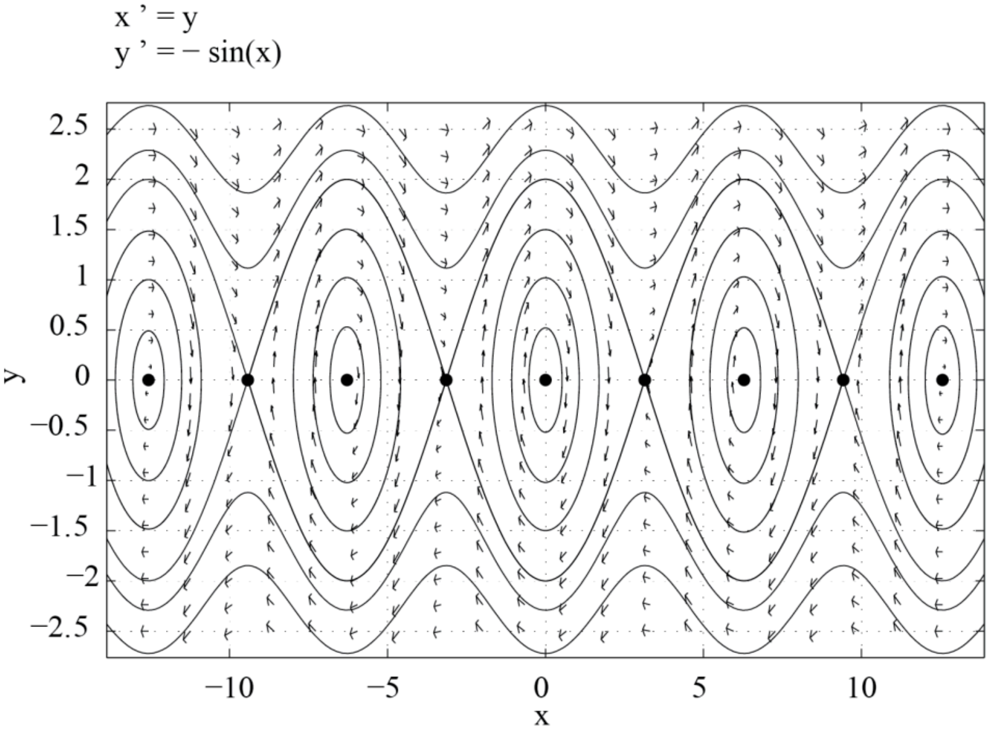


Figure 3.3. Phase plane of the frictionless nonlinear pendulum, as obtained with the software PPLANE.

More generally, any dynamical system of the form

$$\frac{d^2x}{dt^2} + \frac{dV}{dx} = 0 \quad (3.14)$$

can be analyzed in this manner. Indeed, multiplying Equation (3.14) by dx/dt and integrating once gives

$$\frac{1}{2}\left(\frac{dx}{dt}\right)^2 + V(x) = E, \quad (3.15)$$

where the energy E is a constant of motion. By letting $\frac{dx}{dt} = \Lambda$, we obtain an equation for the solution curves in the (x,Λ) plane:

$$\Lambda = \pm\sqrt{2\,(E - V(x))}, \qquad \Lambda \in \mathbb{R}.$$

The corresponding phase portrait can be sketched by noticing that solution curves of energy E only exist for values of x such that $E \geq V(x)$. In particular, *fixed points* of the dynamics, which are such that $x(t) = x_0$ is constant in time, are extrema of $V(x)$ (from Equation (3.14)), and the corresponding energy is given by $E = V(x_0)$ (from Equation (3.15)). One can check that minima of V correspond to *centers* and maxima of V to *saddle points* (see exercises). Moreover, trajectories of energy E that cross the horizontal axis $(\Lambda = 0)$

do so at values of x such that $V(x) = E$. These trajectories have a vertical tangent at the points of crossing (see exercises). Finally, trajectories which emanate from or converge to saddle points are tangent to the eigenvectors of the linearization of (3.14) written as a first order system, at these equilibrium points (see exercises). Figure 3.4 illustrates how the previous statements may be used to plot the phase portrait of Equation (3.14) with $V(x) = -\cos(x)$. The result should be compared to Figures 3.2 and 3.3.

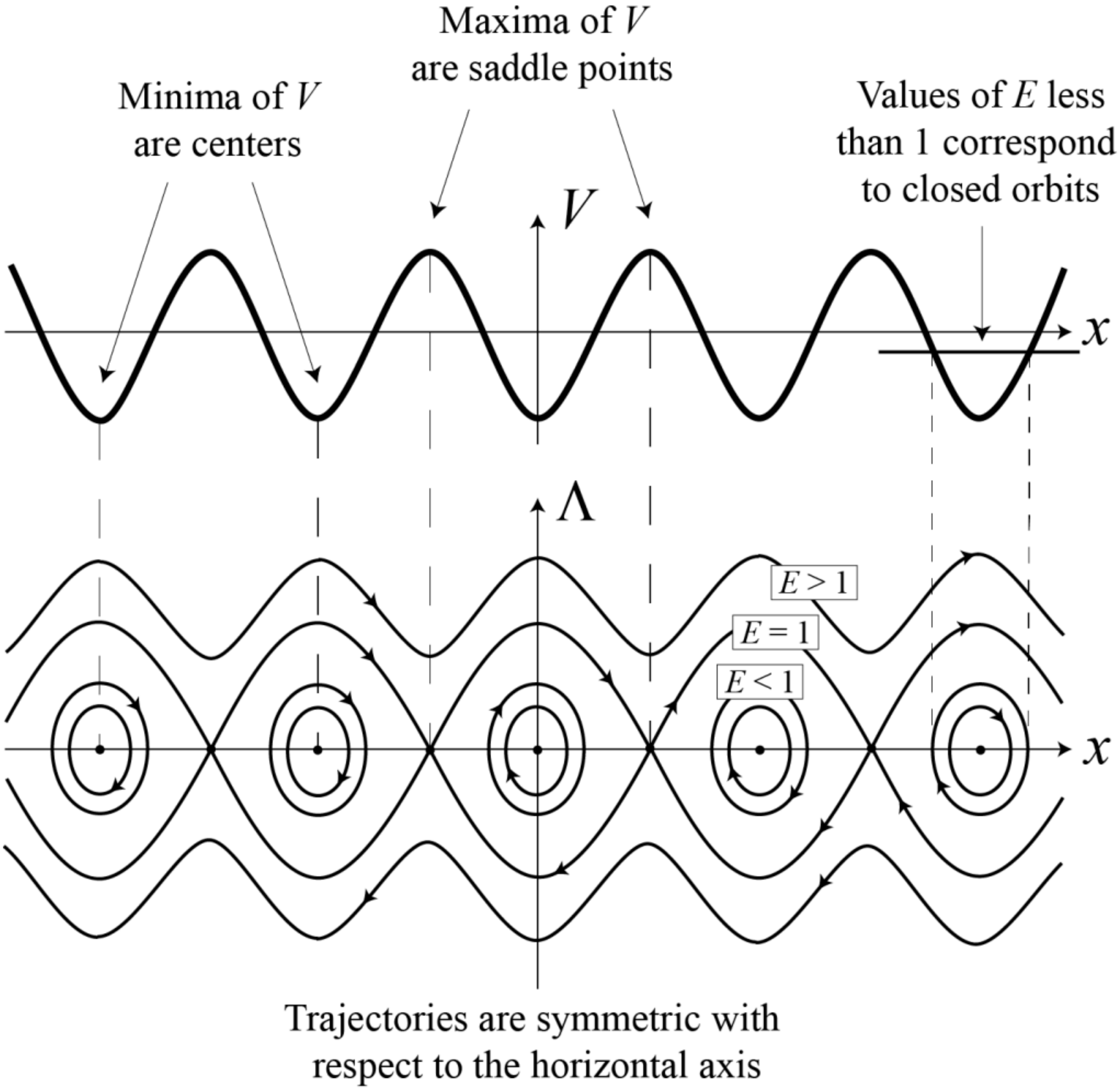

Maxima of V are saddle points
Minima of V are centers
Values of E less than 1 correspond to closed orbits
V
x
Λ
$E > 1$
$E = 1$
$E < 1$
x
Trajectories are symmetric with respect to the horizontal axis

Figure 3.4. Construction of the phase portrait of Equation (3.14), with $V(x) = -\cos(x)$.

In Figure 3.4, the graph of V as a function of x is shown at the top. For each value of the energy E, the quantity $E - V$ is proportional to Λ^2. This information allows us to infer the behavior of the various solution curves in the (x, Λ) plane. The resulting phase portrait is plotted at the bottom.

Analysis in the presence of friction

In the presence of friction, the equation of motion for the nonlinear pendulum reads

$$\frac{d^2\theta}{dt^2} = -\frac{g}{l}\sin(\theta) - \frac{c}{m}\frac{d\theta}{dt}. \qquad (3.16)$$

As before, we first need to check that the dimension of the last term is the correct one. We have

$$\left[\frac{c}{m}\frac{d\theta}{dt}\right] = \left[\frac{1}{l\,m}\,c\,l\,\frac{d\theta}{dt}\right] = L^{-1}M^{-1}[\text{force}] = L^{-1}M^{-1}MLT^{-2} = T^{-2},$$

where we have used the fact that

$$\vec{F}_f = -c\,l\,\frac{d\theta}{dt}\,\vec{\theta}.$$

We now proceed to simplify Equation (3.16) by making the change of variable $\tau = t/t_0$. We find

$$\frac{d^2\theta}{d\tau^2} = -\sin(\theta) - \alpha\frac{d\theta}{d\tau}, \qquad \alpha = \frac{c}{m}\sqrt{\frac{l}{g}}, \qquad (3.17)$$

and we can check that the parameter α is dimensionless, since

$$[\alpha] = [c]M^{-1}T = [\text{force}]L^{-1}TM^{-1}\,T = MLT^{-2}\,L^{-1}\,T^2\,M^{-1} = 1.$$

If $c = 0$, then $\alpha = 0$, and we recover Equation (3.8). The initial conditions for Equation (3.17) are, as before, given by Equation (3.9). Our next step in this analysis is to describe the phase portrait of Equation (3.17). We cannot use the same method as before, since the energy E is no longer conserved. Indeed, we can calculate

$$\begin{aligned}\frac{dE}{d\tau} &= \frac{d}{d\tau}\left(\frac{1}{2}\left(\frac{d\theta}{d\tau}\right)^2 - \cos(\theta)\right) = \frac{d\theta}{d\tau}\frac{d^2\theta}{d\tau^2} + \sin(\theta)\,\frac{d\theta}{d\tau}\\ &= \left(\frac{d^2\theta}{d\tau^2} + \sin(\theta)\right)\frac{d\theta}{d\tau}\end{aligned}$$

which gives

$$\frac{dE}{d\tau} = -\alpha\left(\frac{d\theta}{d\tau}\right)^2.$$

Thus, since $\alpha > 0$, the energy E is either constant (if $d\theta/d\tau = 0$) or decreasing as a function of time. Since E is bounded from below by -1, one can expect that trajectories for which $d\theta/d\tau \neq 0$ will converge towards solutions of energy $E = -1$. Such solutions are given by $\cos(\theta) = 1$, i.e. $\theta = 2n\pi, n \in \mathbb{Z}$; they are fixed points of the dynamics. Pictorially, we can imagine a particle being trapped in one of the valleys of the potential $V(x) = -\cos(x)$, losing energy as it moves, and therefore converging to the local minimum of the potential.

We now turn to the description of the phase plane for Equation (3.17). The corresponding first order system reads

$$\begin{aligned} \frac{d\theta}{d\tau} &= \Lambda, \\ \frac{d\Lambda}{d\tau} &= -\sin(\theta) - \alpha\Lambda. \end{aligned} \tag{3.18}$$

Fixed points are obtained by setting the left-hand-sides of the above equations to zero, that is solving $\Lambda = 0$ and $\sin(\theta) = 0$, which corresponds to points of coordinates $(n\pi, 0)$, $n \in \mathbb{Z}$ in the (θ, Λ) plane. Information on the local dynamics may be obtained by *linearizing* system (3.18) about these fixed points. To this end, we set

$$\theta = n\pi + \mu, \qquad \Lambda = 0 + \nu,$$

where μ and ν are small, and substitute in Equations (3.18). We find

$$\begin{aligned} \frac{d\mu}{d\tau} &= \nu, \\ \frac{d\nu}{d\tau} &= -\sin(n\pi + \mu) - \alpha\nu \\ &= -\cos(n\pi)\mu - \alpha\nu + O(\mu^3), \end{aligned} \tag{3.19}$$

which, if μ is small, can be approximated by the following *linear* system

$$\frac{d}{d\tau}\begin{pmatrix} \mu \\ \nu \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\cos(n\pi) & -\alpha \end{pmatrix}\begin{pmatrix} \mu \\ \nu \end{pmatrix}.$$

The matrix

$$J(n\pi, 0) = \begin{pmatrix} 0 & 1 \\ -\cos(n\pi) & -\alpha \end{pmatrix}$$

is called the *Jacobian* of system (3.18) at the fixed point $(n\pi, 0)$. The general solution of (3.19) can be written in terms of the eigenvalues and eigenvectors of $J(n\pi, 0)$, and this information may be used to assess the linear stability of the corresponding fixed point [2]. Since the characteristic polynomial of a 2×2 matrix A is given by $\lambda^2 - \lambda \mathrm{Tr}(A) + \det(A) = 0$, the eigenvalues of $J(n\pi, 0)$ satisfy the characteristic equation

$$\lambda^2 + \alpha\lambda + \cos(n\pi) = 0.$$

If $n = 2p$ is even, $\cos(n\pi) = 1$; the product of the two eigenvalues of $J(2p\pi, 0)$ or, equivalently, the determinant of $J(2p\pi, 0)$, is positive. The eigenvalues of $J(2p\pi, 0)$ are thus either of the same sign and real, or complex conjugate (recall that the eigenvalues of a matrix with real entries either are real or come in complex conjugate pairs). The sum of the eigenvalues of $J(2p\pi, 0)$ or, equivalently, the trace of $J(2p\pi, 0)$, is $-\alpha$, which is negative. The fixed points $(2p\pi, 0)$, $p \in \mathbb{Z}$, are thus always stable. We can decide on the nature of these fixed points by comparing the square of the trace of $J(2p\pi, 0)$ to four times its determinant or, more directly, by calculating the eigenvalues of $J(2p\pi, 0)$. They are given by

$$\lambda^{\pm}_{(2p\pi,0)} = -\frac{\alpha}{2} \pm \sqrt{\frac{\alpha^2}{4} - 1}.$$

The fixed points $(2p\pi, 0)$ are therefore stable nodes (if $\alpha^2 > 4$) or stable spirals (if $\alpha^2 < 4$). Similarly, if $n = 2p + 1$ is odd, the eigenvalues of $J((2p + 1)\pi, 0)$ have a product equal to $\cos((2p + 1)\pi) = -1$ and are therefore both real and of opposite signs. [3] As a consequence, the fixed points $((2p + 1)\pi, 0)$ are saddle points. For completeness, we give the eigenvalues of $J((2p + 1)\pi, 0)$, which read

$$\lambda^{\pm}_{((2p+1)\pi,0)} = -\frac{\alpha}{2} \pm \sqrt{\frac{\alpha^2}{4} + 1}.$$

These facts are summarized in the phase portrait of Figure 3.5, obtained with PPLANE. We see that trajectories tangent to the eigenvectors of $J((2p + 1)\pi, 0)$ at the points $((2p + 1)\pi, 0)$ are *separatrices* of the phase plane. They are called the *stable* and unstable *manifolds* of the saddle points $((2p + 1)\pi, 0)$. As expected, the unstable manifold of $((2p + 1)\pi, 0)$ contains the the stable fixed points $(2p\pi, 0)$ and $((2p + 2)\pi, 0)$. Trajectories which connect two different fixed points are called *heteroclinic* orbits.

[2] See the phase plane analysis section of the appendix on ordinary differential equations for a review.

[3] Since $J(n\pi, 0)$ has real entries, if its eigenvalues were complex, then they would be complex conjugate and their product would have to be positive.

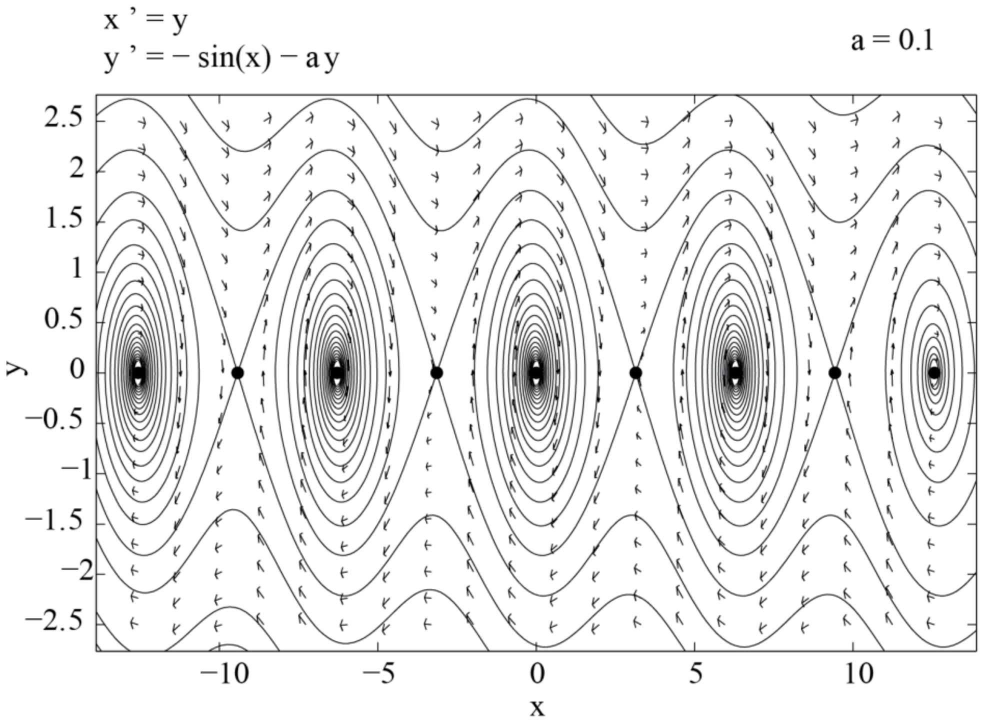


Figure 3.5. Phase plane of the nonlinear pendulum with friction, as obtained with the software PPLANE.

Summary

The nonlinear pendulum provides a good illustration of the modeling process. Newton’s law, together with a set of simplifying hypotheses, allowed us to derive a differential equation model for the angle of the nonlinear pendulum, with or without friction. We then reviewed how to use dimensional analysis to rescale dependent and independent variables in order to obtain a dimensionless model with a reduced number of parameters. The second order differential equations derived in this chapter were studied in terms of phase plane analysis. The phase plane of a two-dimensional conservative system may be obtained by plotting level-sets of the conserved energy. Curves of constant energy may either be found analytically or drawn from the graph of the potential energy. The phase plane of non-conservative two-dimensional dynamical systems may often be inferred from the analysis of the fixed points of the system and of their linear stability. More precisely, nonlinear terms do not change the nature of stable or unstable nodes and spirals, and of saddles points. The reader may want to consult an elementary text on dynamical systems for further details.

Food for Thought

Problem 1

Show that

$$\theta(\tau) = C\cos(\tau + \phi), \text{ with } A, B \in \mathbb{R}$$

can be written as

$$\theta(\tau) = A\cos(\tau) + B\sin(\tau),$$

with $C, \phi \in \mathbb{R}$ by expressing A and B in terms of C and ϕ.

Conversely, explain how you would find C and ϕ, knowing A and B. Are C and ϕ uniquely determined? Why or why not?

Problem 2

It is shown in the text that there exist solutions to Equation (3.8) which exhibit periodic oscillations.

1. Write down an equation for the period T of these oscillations, as an integral over the angle variable θ. Your answer should depend on the energy E.
2. Show that the period T is an increasing function of the energy E. In other words, show that the period of the nonlinear pendulum increases with its amplitude.

Problem 3

Consider a one-dimensional frictionless spring-mass system, where the forces acting on the mass m at position x are the force of gravity $F_g = -mg$ with $g > 0$ constant, and the restoring force of the spring given by $F_r = -k(x - x_0)$, where $k > 0$ and x_0 are constant.

1. Use Newton's law to write down an equation for the position x of the mass.

2. Re-scale the resulting equation. How many parameters can you get rid of?
3. Show that the frictionless spring-mass system is conservative.
4. Describe the dynamics of this system.

Problem 4

Consider the following system of differential equations,

$$\frac{dx}{dt} = F(x,y) \qquad \frac{dy}{dt} = G(x,y).$$

1. What conditions should be satisfied by the coordinates x and y of a fixed point of this system?
2. Assume that (x_0, y_0) is a fixed point of the above system. We are interested in describing the dynamics of the system in the vicinity of the fixed point. To do so, set $x = x_0 + \mu$, $y = y_0 + \nu$, substitute these expressions into the above system, and Taylor-expand the right-hand sides in powers of μ and ν, up to order two.
3. Using the results of part (2) above, show that the linearization of the system about the fixed point (x_0, y_0) reads

$$\frac{d}{dt}\begin{pmatrix}\mu\\ \nu\end{pmatrix} = J(x_0, y_0)\begin{pmatrix}\mu\\ \nu\end{pmatrix},$$

where the Jacobian $J(x_0, y_0)$ of the system at (x_0, y_0) is given by

$$J(x_0, y_0) = \begin{pmatrix} \left.\frac{\partial F(x,y)}{\partial x}\right|_{(x_0,y_0)} & \left.\frac{\partial F(x,y)}{\partial y}\right|_{(x_0,y_0)} \\ \left.\frac{\partial G(x,y)}{\partial x}\right|_{(x_0,y_0)} & \left.\frac{\partial G(x,y)}{\partial y}\right|_{(x_0,y_0)} \end{pmatrix}.$$

Problem 5

Consider a linear system of the form

$$\frac{d}{dt}\begin{pmatrix}x\\ y\end{pmatrix} = \begin{pmatrix}a & 0\\ 0 & b\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix},$$

where $a > 0,\ b < 0.$

1. Give the general solution to this system.
2. Sketch the phase plane of this system, paying particular attention to special trajectories. Can you tell in which direction the solution curves are traced out as time goes on? If so, add arrows to the trajectories you drew.

Problem 6

Sketch the phase plane in the vicinity of the following types of fixed points:

1. A stable node.
2. An unstable spiral.
3. A center.
4. A saddle point.

Problem 7

Describe the eigenvalues of the linearization of a system of dimension two in the vicinity of the following fixed points:

1. A stable node.
2. An unstable spiral.
3. A center.
4. A saddle point.

Problem 8

Find a 2 by 2 matrix A whose entries are all non-zero, and such that the linear system $\dot{X} = AX$ has a fixed point of the following type at the origin:

1. A stable node.
2. An unstable spiral.
3. A center.
4. A saddle point.

Check your answer with a phase plane analysis software, such as the Phase Plane App or equivalent software.

Problem 9

Explain how you would determine the nature (e.g. stable or unstable node or spiral, linear center or saddle point) of a fixed point (x_0, y_0) of the two-dimensional differential system

$$\frac{dx}{dt} = F(x,y), \qquad \frac{dy}{dt} = G(x,y).$$

Problem 10

Sketch the phase planes of the following dynamical systems. If the answer depends on the parameter(s) of the problem, all cases should be considered. Use the Phase Plane App or equivalent software to check your results.

1. $\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ -x - c\,y \end{pmatrix}, \qquad c > 0.$
2. $\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ x(x+1)(x-2) \end{pmatrix}.$
3. $\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ x(x+1)(x-2) - c\,y \end{pmatrix}, \qquad c > 0.$
4. $\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x(6-2y) \\ y(x-3) \end{pmatrix}.$
5. $\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x(1-x-2y) \\ y(1-3x-y)/2 \end{pmatrix}.$
6. $\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ x - x^3 \end{pmatrix}.$

Problem 11

Consider the van der Pol oscillator,

$$\frac{d^2x}{dt^2} + \omega^2 x = \epsilon \frac{dx}{dt}(1 - x^2), \qquad \epsilon \geq 0.$$

1. Describe the dynamics when $\epsilon = 0$.
2. What is the effect of the right-hand-side when $|x|$ is small?
3. What is the effect of the right-hand-side when $|x|$ is much larger than 1?
4. Based on your answers to the previous questions, explain what you expect to happen for intermediate values of $|x|$.
5. Sketch the phase plane of the van der Pol oscillator.
6. Check your answer with the Phase Plane App or equivalent software. Is the result surprising? Why or why not?

Problem 12

Consider the following differential equation, $\dfrac{dx}{dt} = \lambda x - \gamma x^3.$

1. What is the dimension of λ?
2. What is the dimension of γ? Your answer should be in terms of the dimension of x, denoted by $[x]$.
3. Let t_0 be a characteristic time scale for this problem. Define a dimensionless time variable $\tau = t/t_0$, and show that you can make a change of variable from t to τ, to “remove” the parameter λ.
4. Can you find a change of variable that would allow you to remove γ as well?

Problem 13

The force of gravity between two bodies of mass m_1 and m_2 has intensity $F = G\dfrac{m_1 m_2}{r^2}$, where r is the distance between the centers of mass of the two bodies, and G is the gravitational constant.

1. Use this formula to show that for an object of mass m at the surface of the Earth, one can approximate the force F by $F \simeq mg$, where the acceleration of gravity g is constant.
2. Express g in terms of G, the mass M of the Earth, and the radius R of the Earth.

Problem 14

Consider a smooth function $f(x)$, and its Taylor expansion near $x = x_0$,

$$f(x) = \sum_{i=0}^{n} f^{(i)}(x_0)\frac{(x-x_0)^i}{i!} + f^{(n+1)}(\bar{x})\frac{(x-x_0)^{n+1}}{(n+1)!}, \qquad (3.20)$$

where $\bar{x} \in [x_0, x]$. Let $E_n(x)$ be the error made by approximating f with its Taylor expansion truncated to order n,

$$E_n(x) = f(x) - \sum_{i=0}^{n} f^{(i)}(x_0)\frac{(x-x_0)^i}{i!}.$$

1. Use Equation (3.20) above to write the Taylor expansion of $f(x) = \cos(x)$, near $x_0 = 0$ to order $n = 5$.
2. Find a condition on $|x|$ which ensures that $|E_5(x)| < 0.05$.
3. Use a calculator or MATLAB to check your answer to the previous question.

Problem 15

Consider the following model

$$\frac{\partial u}{\partial t} = \mu u + \alpha\frac{\partial^2 u}{\partial x^2} - \beta u^3, \qquad u \in \mathbb{R}, \quad \alpha, \beta > 0.$$

1. What are the dimensions of the parameters α, β and μ? Write $[u]$ for the dimension of u.
2. How many relevant parameters does this model have? Explain.

Problem 16

Consider the second order differential equation $\dfrac{d^2x}{dt^2} + \dfrac{dV}{dx} = 0$, where $V(x)$ is a smooth potential.

1. Re-write this second order equation as a first-order system.
2. Find the fixed points of the first-order system and show that they correspond to critical points of V.
3. Find the linearized system about an arbitrary fixed point and show that maxima of V correspond to saddle points and minima of V to linear centers.

Problem 17

Consider the dynamical system $\dfrac{d^2x}{dt^2} + \dfrac{dV}{dx} = 0,$ where $V(x)$ is a smooth function. Assume that $(x_0, 0)$ is a saddle point of the corresponding first-order system of differential equations.

Find an equation describing the trajectories in the associated phase plane near the fixed point $(x_0, 0)$.

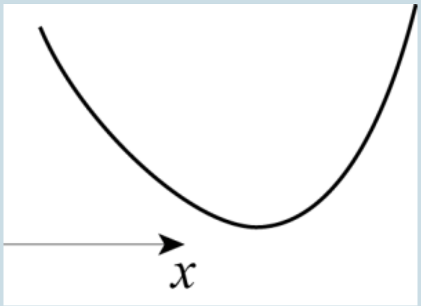
Problem 18

Consider the dynamical system $\dfrac{d^2x}{dt^2} + \dfrac{dV}{dx} = 0,$ where $V(x)$ is a smooth function.

Show that, in the phase plane of coordinates $\left(x, \dfrac{dx}{dt}\right)$, trajectories with energy E such that $\min(V) < E < \max(V)$ (where $\min(V)$ and/or $\max(V)$ may be infinite depending on the potential V) cross the x-axis. In addition, show that they have a vertical tangent at points of crossing where $\dfrac{dV}{dx} \neq 0$.

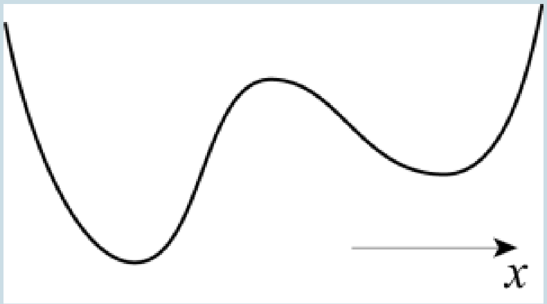
Problem 19

Sketch the phase plane associated with the differential equation $\dfrac{d^2x}{dt^2} + \dfrac{dV}{dx} = 0,$ where the potential $V(x)$ has the shape given below. Only the portion of the graph of $V(x)$ near its minimum is shown. You may assume that the trends suggested by the picture continue outside of the plot area, i.e. that $\lim_{x \to -\infty} V(x) = \lim_{x \to +\infty} V(x) = +\infty.$



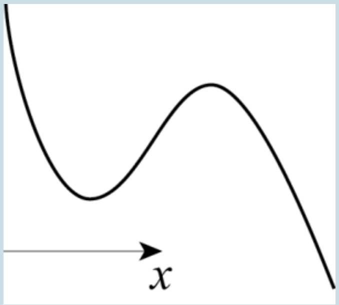
Problem 20

Sketch the phase plane associated with the differential equation $\frac{d^2x}{dt^2} + \frac{dV}{dx} = 0,$ where the potential $V(x)$ has the shape given below. Only the portion of the graph of $V(x)$ near its extrema is shown. You may assume that the trends suggested by the picture continue outside of the plot area, i.e. that $\lim_{x \to -\infty} V(x) = \lim_{x \to +\infty} V(x) = +\infty.$



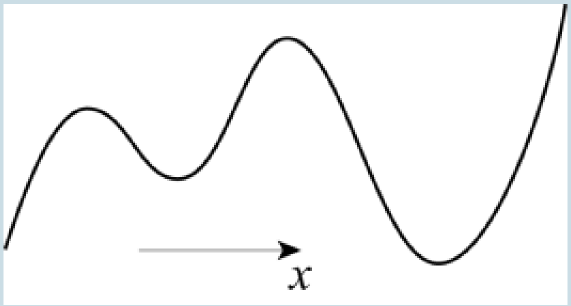
Problem 21

Sketch the phase plane associated with the differential equation $\frac{d^2x}{dt^2} + \frac{dV}{dx} = 0,$ where the potential $V(x)$ has the shape given below. Only the portion of the graph of $V(x)$ near its extrema is shown. You may assume that the trends suggested by the picture continue outside of the plot area, i.e. that $\lim_{x \to -\infty} V(x) = +\infty, \quad \lim_{x \to +\infty} V(x) = -\infty.$



Problem 22

Sketch the phase plane associated with the differential equation $\frac{d^2x}{dt^2}+\frac{dV}{dx}=0,$ where the potential $V(x)$ has the shape given below. Only the portion of the graph of $V(x)$ near its extrema is shown. You may assume that the trends suggested by the picture continue outside of the plot area, i.e. that $\lim_{x\to-\infty} V(x) = -\infty, \quad \lim_{x\to+\infty} V(x) = +\infty.$



Problem 23

Consider the Navier-Stokes equation

$$\frac{\partial\vec{v}}{\partial t}+(\vec{v}\cdot\nabla)\,\vec{v}=-\frac{1}{\rho}\nabla p+\nu\nabla^2\vec{v}+\frac{1}{\rho}\vec{F},$$

where the vector $\vec{v}$ represents the velocity of a fluid in 3 spatial dimensions, ρ is the density of the fluid, p is its pressure, ν is the kinematic viscosity of the fluid, and $\vec{F}$ stands for the density of bulk forces applied to the fluid. The operator ∇ is the gradient in 3 dimensions.

1. What is the dimension of the two terms on the left hand side of the equation? Why should these dimensions be the same?

2. Use this information to find the dimension of p and that of $\vec{F}$. Is the result what you expected, based on your knowledge of forces and pressure?

3. Find the dimension of ν.

4. Let U be a characteristic velocity, and l a characteristic length. Define dimensionless variables $\vec{V}$ and $\vec{X}$ such that $\vec{v} = U\vec{V}$ and $\vec{x} = l\vec{X}$, where $\vec{x}$ is the position vector. Re-write the Navier-Stokes equation in terms of these new variables.

5. Show that if you now re-scale time and introduce dimensionless versions of p and $\vec{F}$, then the dimensionless Navier-Stokes equation involves only one parameter, $Re = Ul/\nu$. This parameter is called the *Reynolds number*.

4.

STONE-SKIPPING

Learning Objectives

At the end of this chapter, you will be able to do the following.

- Understand stone skipping as an alternation of free flight and collision phases.
- Apply concepts of classical mechanics to describe the dynamics of the stone during each phase.
- Justify approximations to simplify the equations of motion.
- Formulate initial conditions for each phase, based on the outcome of the previous phase.
- Analyze the dynamics of each phase.
- Evaluate theoretical predictions against experimental observations.

We now consider the problem of modeling stone-skipping. The goal is to understand how the stone skims the surface of the water and how far it goes. With this information, it should be possible to suggest an optimal way of throwing the stone, in order to increase the number of rebounds. One of the major differences with the problem of the pendulum discussed previously, is that the stone cannot be considered as a point mass. As a consequence, we will use two sets of model equations: the conservation of momentum to describe the motion of the center of mass (a point), and the equations of motion for a rigid body, to describe the rotation of the stone about its center of mass.

Experimental data

We probably all have some sort of experience with stone-skipping, and we intuitively know that in order to get a large number of bounces, the stone should be rather flat, and thrown with a small incidence angle, and with a spin. We also know that if we do not throw the stone correctly, it may bounce once, but will then start tumbling while in the air and then sink. Remarkably enough, controlled experimental data is available for this problem. A 2004 article[1] by C. Clanet, F. Hersen and L. Bocquet reports laboratory measurements giving the minimal initial velocity of the stone, the angle between the velocity vector of the stone and the water surface, as well as the collision time, as functions of the angle between the stone and the water surface.

Assumptions and model equations

The discussion below is based on the 2003 article[2] by L. Bocquet, entitled *The physics of stone skipping*. We use the same notation, in order to make the reading easier.

We will assume that the motion of the stone alternates between free flight, when the stone is in the air, and collisions with the surface of the water. Each of these stages can be described separately, using equations of classical mechanics. The initial conditions for say a free flight phase will be the position and velocity of the stone as it takes off from the water at the end of the preceding collision phase, and vice-versa. We will start by describing the collision process, and then discuss the free flight part of the motion.

Collision phase

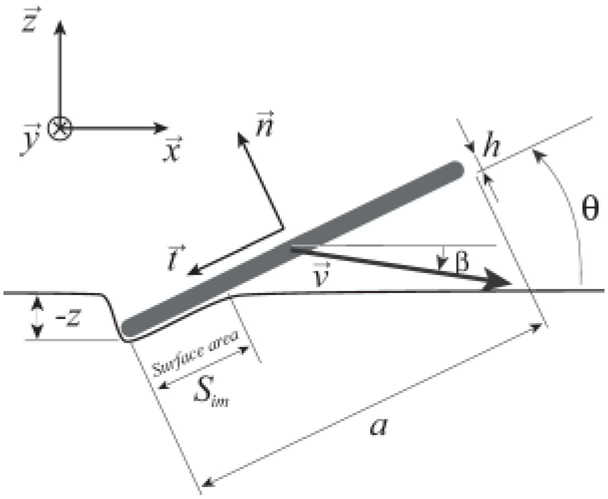


Figure 4.1. Cross-section of a flat stone during the collision process (adapted from L. Bocquet's 2003 article).

Consider a thin, flat (planar), homogeneous and symmetric (e.g. square or circular) stone of total mass M and assume that during the collision process, the stone does not spin or tumble, and thus remains parallel to itself.[3] As a consequence, the problem may be simplified by considering the cross-section of the stone through its center of mass and parallel to the plane defined by the vertical and the velocity vector of the stone (see Figure 4.1). All of the points in this cross-section (and in fact in the stone) thus move with the same velocity vector.

The parameters for this problem are then the size a of the stone (a is the side length of a square stone or the diameter of a circular stone), its thickness $h \ll a$, its velocity vector $\vec{v}(t)$, and the angle θ (assumed constant during

[1] C. Clanet, F. Hersen, and L. Bocquet, *Secrets of successful stone-skipping*, Nature **427**, 29 (2004).

[2] L. Bocquet, *The physics of stone skipping*, Am. J. Phys. **71**, 150-155 (2003).

[3] Is this assumption reasonable? Why or why not?

each collision phase) made by the stone with the surface of the water. We will call $\beta(t)$ the angle between $\vec{v}(t)$ and the horizontal, and define two orthonormal bases: $(\vec{x}, \vec{y}, \vec{z})$ and $(\vec{n}, \vec{y}, \vec{t})$, such that $\vec{y}$ is perpendicular to the plane of the cross-section, $\vec{z}$ points upward, $\vec{n}$ is normal and $\vec{t}$ is tangent to the surface of the stone. See Figure 4.1 for an illustration.

Since we made the assumption that the stone does not rotate during the collision process, its motion can be reduced to that of its center of mass, and is completely described by Newton's law. The forces applied to the stone are the force of gravity, $-Mg\,\vec{z}$, and the force $\vec{F} = F_x\,\vec{x} + F_z\,\vec{z}$ exerted by the water on the stone. If we write $\vec{v} = v_x\,\vec{x} + v_z\,\vec{z}$, we have

$$M\,\frac{dv_x}{dt} = F_x, \qquad M\,\frac{dv_z}{dt} = -Mg + F_z. \qquad (4.1)$$

Of course, we now need a model for the force $\vec{F}$. First, this force can be considered as being the sum of a lift force, parallel to $\vec{n}$, and of a friction force, parallel to $\vec{t}$. Second, let us consider the dynamic of the water during the collision process. The fluid moves according to the Navier-Stokes equation,

$$\rho_w\left(\frac{d\vec{V}}{dt} + \left(\vec{V}\cdot\vec{\nabla}\right)\vec{V}\right) = -\nabla p + \mu\nabla^2\vec{V} + \vec{f},$$

where ρ_w is the density of the water, $\vec{V}$ is the velocity vector of a fluid particle, p is the pressure, μ is the dynamic viscosity of the water, and $\vec{f}$ represents bulk forces (such as gravity) exerted on the water. At the interface between the water and the stone, the pressure p is given by

$$p \simeq \frac{||\vec{F}||}{S_{im}},$$

where S_{im} is the surface area of the immersed portion of the stone (see Figure 4.1). We will now perform some dimensional analysis to compare the inertial and diffusive terms in this equation. We have

$$\frac{\left[\rho_w\left(\vec{V}\cdot\vec{\nabla}\right)\vec{V}\right]}{\left[\mu\vec{\nabla}^2\vec{V}\right]} = \left[\frac{\rho_w||\vec{V}||L}{\mu}\right] = \left[\frac{||\vec{V}||L}{\nu_w}\right] \equiv Re,$$

where L is a characteristic length, $\nu_w = \mu/\rho_w$ is the kinematic viscosity of water, and Re is called the Reynolds number. If we assume that $||\vec{V}||$ is of the order of a few meters per second, that L is of the order of the size of the stone – say a few centimeters – and since, for water, $\nu_w = 10^{-6}\ \mathrm{m}^2\,\mathrm{s}^{-1}$, we find

$$Re = \frac{10\cdot 10^{-2}}{10^{-6}} = 10^5 \gg 1.$$

As a consequence, viscous terms are negligible, and we can balance inertial terms with pressure terms. This gives

$$\left[\frac{p}{L}\right] \simeq \left[\rho_w \left(\vec{V}\cdot\vec{\nabla}\right)\vec{V}\right] = \left[\frac{\rho_w ||\vec{V}||^2}{L}\right],$$

which together with $p \simeq ||\vec{F}||/S_{im}$ gives

$$||\vec{F}|| \propto \rho_w \, S_{im} \, ||\vec{V}||^2.$$

Since $||\vec{V}||$ and $||\vec{v}||$ are comparable, we will thus consider that the intensity of the force $\vec{F}$ is proportional to the density of water ρ_w, the immersed surface area S_{im}, and the square of the velocity vector of the stone, $||\vec{v}||^2$. We can then write

$$\vec{F} = \left(\frac{1}{2}C_l\,\vec{n} + \frac{1}{2}C_f\,\vec{t}\right)\rho_w\,S_{im}\left(v_x^2 + v_z^2\right),$$

where the dimensionless coefficients C_l and C_f describe how the normal (lift) and tangential (friction) components of $\vec{F}$ vary with $\rho_w\,S_{im}\,||\vec{v}||^2$. We will assume that these coefficients are constant. We can now substitute this expression into Equations (4.1). Since $\vec{n} = -\sin(\theta)\,\vec{x} + \cos(\theta)\,\vec{z}$ and $\vec{t} = -\cos(\theta)\,\vec{x} - \sin(\theta)\vec{z}$, we get

$$\begin{aligned} M\,\frac{dv_x}{dt} &= -\frac{1}{2}\rho_w\,S_{im}(z)\left(v_x^2 + v_z^2\right)\left(C_l\sin(\theta) + C_f\cos(\theta)\right), \\ M\,\frac{dv_z}{dt} &= -Mg + \frac{1}{2}\rho_w\,S_{im}(z)\left(v_x^2 + v_z^2\right)\left(C_l\cos(\theta) - C_f\sin(\theta)\right). \end{aligned} \tag{4.2}$$

These equations are nonlinear in v_x and v_z. Moreover, we made explicit the dependence of S_{im} on the vertical coordinate z of the part of the stone that is the deepest under water (see Figure 4.1). Equations (4.2) are valid as long as the stone is not completely immersed. If this happened, the lift exerted by the water on the stone would be given by Archimedes' law, and the stone would sink. At the beginning of the collision phase, $v_z < 0$. In order for the stone to skim the surface of the water, we need to have $dv_z/dt > 0$, so that v_z is positive at the end of the collision phase. We see that the angle θ needs to be small for this to happen, since $C_l\cos(\theta) - C_f\sin(\theta)$ must be positive if we want

$$\frac{1}{2}\rho_w\,S_{im}(z)\left(v_x^2 + v_z^2\right)\left(C_l\cos(\theta) - C_f\sin(\theta)\right)$$

to balance $-Mg$.

Free flight phase

During a free flight phase, the motion of the center of mass of the stone is described by Newton's equation in which the only force is gravity:

$$M\,\frac{dv_x}{dt} = 0, \qquad M\,\frac{dv_z}{dt} = -Mg. \qquad (4.3)$$

If we denote by X and Z the coordinates of the stone, we have

$$X(t) = X(0) + v_x(0)t \qquad Z(t) = v_z(0)t - \frac{1}{2}gt^2, \qquad (4.4)$$

where $t = 0$ is taken at the beginning of the free flight phase. The initial position $(X(0), Z(0) = 0)$ and velocity $(v_x(0), v_z(0))$ of the stone are given by the final position and velocities of the preceding collision phase, unless of course this is the first free flight phase.

The rigid rotation of the stone about its center of mass is given by Euler's equations,

$$\begin{aligned} I_{-y}\frac{d\omega_y}{dt} - \omega_n\omega_p(I_n - I_p) &= N_{-y} \\ I_n\frac{d\omega_n}{dt} - \omega_p\omega_{-y}(I_p - I_{-y}) &= N_n \qquad (4.5) \\ I_t\frac{d\omega_p}{dt} - \omega_{-y}\omega_n(I_{-y} - I_n) &= N_p, \end{aligned}$$

where I_α, ω_α and N_α are respectively the moment of inertia, angular frequency and torque about the body principal axis parallel to $\vec{\alpha}$. Here, $-\vec{y}$ and $\vec{p}$ are unit vectors tangent to the surface of the stone, and parallel to two of its principal axes. The orthonormal frame $(-\vec{y}, \vec{n}, \vec{p})$ is attached to the center of mass of the stone, and rotates with it. A similar set of equations could be written for the collision phase of the motion, if we wanted to take into account the fact that the stone may be spinning and tumbling as it collides with the water.

Since the stone is square or circular, one has $I_{-y} = I_p \equiv J_1$, and we will use the notation $I_n \equiv J_0$. Moreover, the only force acting on the stone, $-Mg\,\vec{z}$, is applied at its center of mass, so that all of the torques are zero. Equations (4.5) can then be simplified into

$$\begin{aligned}
&J_1 \frac{d\omega_{-y}}{dt} - \omega_n \omega_p (J_0 - J_1) = 0 \\
&J_0 \frac{d\omega_n}{dt} = 0 \\
&J_1 \frac{d\omega_p}{dt} - \omega_{-y} \omega_n (J_1 - J_0) = 0.
\end{aligned} \tag{4.6}$$

Analysis

We will start by analyzing the equations for the free flight phase of the motion, and then go back to the collision process.

Free flight phase

From Equations (4.6), we see that ω_n is conserved, and is thus equal to the initial spinning velocity Ω_0 of the stone about the axis normal to its surface and through its center of mass. We are thus left with two linear equations,

$$\begin{aligned}
J_1 \tfrac{d\omega_{-y}}{dt} - \Omega_0\, \omega_p (J_0 - J_1) \quad &= 0 \\
J_1 \frac{d\omega_p}{dt} - \omega_{-y}\, \Omega_0 (J_1 - J_0) \quad &= 0,
\end{aligned} \tag{4.7}$$

which read, in matrix form, $\dfrac{d}{dt}\begin{pmatrix}\omega_{-y}\\ \omega_p\end{pmatrix} = \begin{pmatrix}0 & \delta\\ -\delta & 0\end{pmatrix}\begin{pmatrix}\omega_{-y}\\ \omega_p\end{pmatrix}$ with $\delta = \dfrac{\Omega_0(J_0 - J_1)}{J_1}$. In the absence of initial spin, i.e. if $\Omega_0 = 0$, we obtain

$$\omega_{-y} = \omega_{-y}(0) \qquad \omega_p = \omega_p(0),$$

which implies, since $\omega_{-y} = d\theta/dt$,

$$\theta(t) = \omega_{-y}(0)\, t.$$

In other words, if during the collision phase, the stone is even slightly put into rotation about the $\vec{y}$ axis, i.e. if $\omega_{-y}(0)$ is non-zero, then the stone will start tumbling during the free flight phase. At the next collision, the stone is likely to hit the water with a large incidence angle θ, and then sink.

It is thus critical that Ω_0 be non-zero. In such a case, both ω_{-y} and ω_p oscillate in time with frequency $|\delta|$ and are given by

$$\omega_{-y}(t) = \omega_{-y}(0)\cos(\delta t) + \omega_p(0)\sin(\delta t)$$
$$\omega_p(t) = \omega_p(0)\cos(\delta t) - \omega_{-y}(0)\sin(\delta t).$$

As a consequence,

$$\theta(t) = \theta(0) + \frac{1}{\delta}\omega_{-y}(0)\sin(\delta t) - \frac{1}{\delta}\omega_p(0)\cos(\delta t)$$

and will remain close to $\theta(0)$ if the coefficients of the above equation are small. Thus, a spin about the $\vec{n}$ direction stabilizes the stone and prevents it from tumbling. This is called the *gyroscopic effect*.

We now turn to the motion of the center of mass of the stone during a free flight phase. Equations (4.4) tell us that the duration τ_f of each free flight phase is such that $Z(\tau_f) = 0$, i.e.

$$\tau_f = \frac{2v_z(0)}{g},$$

and the length of each skip, that is the distance covered by the stone along the $\vec{x}$ direction is

$$\lambda = v_x(0)\tau_f = v_x(0)\frac{2v_z(0)}{g}.$$

From Equations (4.2), it is clear that when θ is small, v_x is decreasing during each collision phase. Thus the value $v_x(0)$ of v_x at the beginning of each free flight phase will decrease from one free flight phase to the next. Since we do not expect v_z to increase between two free flight phases, we see that the length of each skip decreases from one skip to the next, as expected.

Collision phase

Our analysis of Equations (4.2) should first of all indicate under what conditions the stone is going to emerge and take off from the water at the end of a collision phase. Consider the equation for v_z,

$$M\frac{dv_z}{dt} = -Mg + \frac{1}{2}\rho_w\, S_{im}(z)\left(v_x^2 + v_z^2\right)\left(C_l\cos(\theta) - C_f\sin(\theta)\right). \quad (4.10)$$

We can appreciate the information contained in this equation if we make the approximation $v_x \gg v_z$, which is reasonable, and also assume that $v_x^2 + v_z^2 = v^2 \equiv v_0^2$ is constant during the collision phase. The latter statement is not quite correct since the stone is slowed down by friction with the water, but if the stone bounces about 20 times, then the relative loss $\left(v^2(0) - v^2(\tau_c)\right)/v^2(0)$ during each collision is small, at least for most collisions. Here, τ_c denotes the duration of the collision. Under these conditions, Equation (4.10) reads

$$M\frac{dv_z}{dt} = -Mg + \frac{1}{2}\rho_w S_{im}(z) v_0^2 \left(C_l \cos(\theta) - C_f \sin(\theta)\right) \equiv -\mathcal{F}(z). \quad (4.11)$$

The right-hand-side, $-\mathcal{F}(z)$, is a function of z only (recall that we assumed at the beginning that θ is constant during each collision phase), and since $v_z = dz/dt$, we can multiply both sides of this equation by v_z and integrate with respect to t, to get

$$\frac{M}{2}\left(\frac{dz}{dt}\right)^2 = \int -\mathcal{F}(z)\frac{dz}{dt}\,dt = \int -\mathcal{F}(z)\,dz \equiv -\mathcal{U}(z) + E, \quad (4.12)$$

where $\mathcal{U}(z) = \displaystyle\int_0^z F(s)\,ds$, and the constant E is arbitrary. When $z = 0$, $S_{im}(z) = 0$; moreover, $S_{im}(z)$ increases as z becomes more negative. Therefore, the potential $\mathcal{U}(z)$ has a shape like that drawn in Figure 4.2 for negative z‘s, provided θ is small enough.

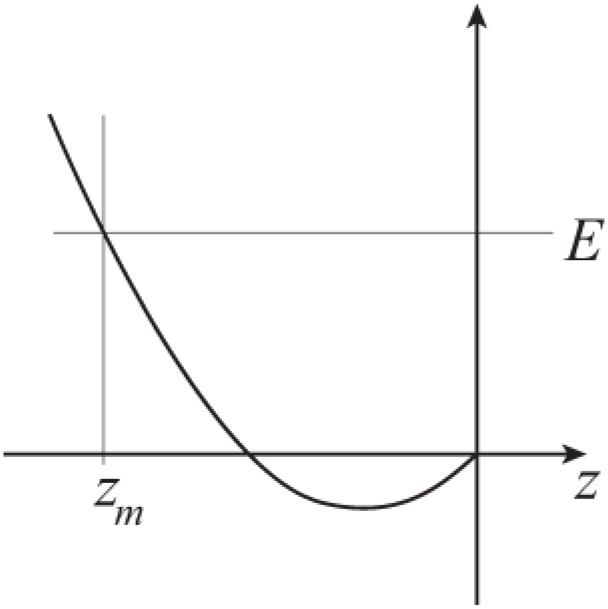


Figure 4.2. Shape of the potential $U(z)$ as a function of z for $z < 0$.

From the shape of $\mathcal{U}$, one can see that the trajectory of the stone will be such that z will have a turning point at $z = z_m$, and that at the end of the collision phase, the stone will emerge with a vertical velocity $v_z(\tau_c) = -v_z(0)$, opposite to the vertical velocity it had at the beginning of the collision phase. This occurs only if the energy E is such that the stone is not completely immersed when $z = z_m$. If we keep E fixed, this imposes a condition on v_0^2 since the larger v_0^2, the smaller $|z_m|$. Therefore, the stone will not sink during a collision phase, provided $v_0^2 > v_c^2$, where the critical velocity v_c depends on the parameters of the problem. Since $v_x \gg v_z$, the condition on v_0 can also be re-written as a condition on the initial velocity of the stone in the $\vec{x}$ direction, which thus reads $v_x(0) > v_c$. As shown in the 2003 article by L. Bocquet, and in the exercises at the end of this Chapter, one can calculate $S_{im}(z)$ for a stone of a particular shape, solve Equation (4.11) explicitly, and obtain an expression for v_c.

As mentioned above, we made the assumption that the dynamics was conservative (i.e. $v^2 = v_0^2$ was assumed constant). This is obviously not true since the stone does not keep bouncing forever. There is a loss of energy during the collision process, due to the friction of the stone on the surface of the water. Everything we said before will be qualitatively correct – in particular the stone will emerge with a vertical velocity only slightly smaller than $|v_z(0)|$ – as long as the relative loss of energy during a collision phase is small. This will certainly be true for the first bounces, but will cease to be valid for the last ones. Since this energy loss is the major reason for the slowing down of the stone, we will now try to estimate this quantity. The (signed) energy loss $\mathcal{W}$ is equal to the work of the friction force exerted by the water on the stone. This can be roughly estimated as

$$\begin{aligned}\mathcal{W} &\simeq [\text{force along } \vec{x}] \cdot [\text{distance covered by the stone along } \vec{x}] \\ &\simeq \vec{F} \cdot \vec{x} l,\end{aligned}$$

where $l = v_x(0)\ \tau_c$. From Equation (4.11), we see that

$$\left[\frac{\rho_w}{M} v_0^2\right] = L^{-1} T^{-2},$$

and we can thus expect

$$\tau_c \propto \sqrt{\frac{M}{a \rho_w v_0^2}} \simeq \sqrt{\frac{M}{a \rho_w}} \frac{1}{v_x(0)},$$

where a is the size (side length or diameter) of the stone. Therefore, $l = v_x(0)\tau_c$ does not depend on $v_x(0)$ at lowest order. During the collision, the right-hand-side of Equation (4.10) is small, so we can write that

$$Mg \simeq \vec{F} \cdot \vec{z}.$$

Moreover, since

$$\frac{\vec{F} \cdot \vec{x}}{\vec{F} \cdot \vec{z}} = -\frac{C_l \sin(\theta) + C_f \cos(\theta)}{C_l \cos(\theta) - C_f \sin(\theta)} \equiv -\mu,$$

(the ratio μ should not be confused with the dynamic viscosity of water), we have

$$\vec{F} \cdot \vec{x} \simeq -Mg\,\mu,$$

and

$$\mathcal{W} \simeq -Mg\,\mu\,l.$$

Thus, since the difference in kinetic energy between the beginning and the end of the collision phase is equal to $\mathcal{W}$, we have

$$\frac{1}{2} M v_x^2(\tau_c) - \frac{1}{2} M v_x^2(0) = \mathcal{W} = -Mg\,\mu\,l,$$

and the initial velocity $v_x(0)$ must satisfy

$$v_x(0) > \sqrt{2\,g\,\mu\,l} \equiv V_c.$$

Putting everything together, the velocity along $\vec{x}$ at the beginning of a collision phase must satisfy

$$v_x(0) > \max(v_c, V_c),$$

where v_c is obtained by specifying that the stone should not become completely immersed during a collision, and V_c is such that the stone is moving fast enough to compensate for the energy due to friction forces. Our discussion of the free flight phase shows that the length of each jump is proportional to the horizontal velocity v_x. We know that v_x is conserved during a free flight phase (see Equation (4.3)), and that v_x^2 decreases by $2\,g\,\mu\,l$ during each collision phase. Since $|v_z|$ is the same at the beginning and end of each free flight phase, and since v_z varies only slightly during a collision phase, the length of a jump after N collision phases will be given by

$$\begin{aligned}\lambda_N &= \frac{2|v_z(0)|}{g}\sqrt{v_{x,0}^2 - 2\,N\,g\,\mu\,l} = \frac{2|v_z(0)|v_{x,0}}{g}\sqrt{1-\frac{2\,N\,g\,\mu\,l}{v_{x,0}^2}} \\ &= \lambda_0\sqrt{1-\frac{N}{N_c}}, \qquad \lambda_0 = \frac{2|v_z(0)|v_{x,0}}{g},\ N_c = \frac{v_{x,0}^2}{2\,g\,\mu\,l},\end{aligned}$$

where $v_{x,0}$ is the $\vec{x}$-velocity of the stone as it is initially thrown. We thus see that the length of each jump decreases as N gets large, and scales like $\sqrt{1-N/N_c}$.

Summary

This model explains the existence of a minimum velocity at which the stone should be thrown; it explains why the stone should be given a spin; and it describes how the length of each jump decreases as the number of bounces increases. It also indicate that the angle θ between the stone and the surface of the water should be small. Experiments described in the article by C. Clanet *et al.*[4] show that $\theta = 20^o$ is optimal. This value of θ minimizes the collision time; it is also the incidence angle for which successful stone-skipping occurs with the smallest initial speed. In order to explain these observations, one would for instance have to take into account the dependence of C_l and C_n on θ, and find the value of θ that minimizes the energy loss $\mathcal{W}$.

[4] Figures 1.b and 1.c of the 2004 article by C. Clanet *et al.* show the regions of successful stone-skipping in the $(\theta, ||v||)$ and (θ, β) planes; Figure 1.d shows the collision time as a function of θ for different values of β (see Figure 4.1 for a definition of the parameters).

Food for thought

Problem 1

Consider Equations (4.5).

1. Use these equations to determine the dimension of ω_α, where α is each of $y,\ n$ and p.
2. Relate the dimension of I_α to that of N_α.
3. Given that a torque is of the form $\vec{N} = \vec{r} \times \vec{f}$ where $\vec{r}$ is a position vector and $\vec{f}$ a force, find the dimension of N_α.
4. Use the above to find the dimension of I_α. Does this agree with the definition $I_\alpha \equiv \displaystyle\int ||\vec{r}||^2 \rho\, dV$, where ρ is a density and dV is a volume element?

Problem 2

This problem is based on the 2003 article by L. Bocquet.

Consider Equation (4.11), and assume that the stone has a square shape.

1. Show that $S_{im} = a|z|/\sin(\theta)$.
2. Using the expression for S_{im}, show that Equation (4.11) becomes a linear differential equation for z.
3. Find the general solution this ordinary differential equation.
4. Apply the boundary conditions and show that $z(t)$ is given by

$$z(t) = \frac{-g}{\omega_0^2} + \frac{g}{\omega_0^2}\cos(\omega_0 t) + \frac{v_z(0)}{\omega_0}\sin(\omega_0 t),$$

$$\text{where} \qquad \omega_0^2 = \frac{\left(C_l \cos(\theta) - C_f \sin(\theta)\right)\rho_w v_x(0)^2 a}{2M\sin(\theta)}.$$

Problem 3

This problem is based on the 2003 article by L. Bocquet.

The dynamics of the immersed edge of a square stone during a collision phase is described by (see Problem 2)

$$z(t) = \frac{-g}{\omega_0^2} + \frac{g}{\omega_0^2}\cos(\omega_0 t) + \frac{v_z(0)}{\omega_0}\sin(\omega_0 t),$$

where

$$\omega_0^2 = \frac{\left(C_l\cos(\theta) - C_f\sin(\theta)\right)\rho_w v_x(0)^2 a}{2M\sin(\theta)}.$$

1. Use this expression to show that the edge of the stone reaches a depth $|z_m|$, given by
$$z_m = -\frac{g}{\omega_0^2}\left[1 + \sqrt{1 + \frac{\omega_0^2 v_z(0)^2}{g^2}}\right].$$
2. What is the condition on z for the stone to be completely immersed?
3. Show that the stone will not be completely immersed during a collision phase if
$$v_x(0)^2 > \frac{4Mg}{C\rho_w a^2}\left[1 - \frac{2\tan^2\beta\, M}{a^3 C\rho_w\sin(\theta)}\right]^{-1}, \quad C = C_l\cos(\theta) - C_f\sin(\theta).$$

Problem 4

This problem is based on the 2003 article by L. Bocquet.

Using Euler's equations of motion (4.5), together with the appropriate expressions for the lift force applied to stone during the collision phase, derive Equation (20) of the 2003 article by L. Bocquet,

$$\frac{d^2\theta}{dt^2} + \frac{J_0 - J_1}{J_1}(\theta - \theta(0)) = \frac{\mathcal{M}_\theta}{J_1},$$

where $\mathcal{M}_\theta$ is the projection on the $\vec{y}$ axis of the torque exerted by the water on the stone.

PART III
POPULATION DYNAMICS AND EPIDEMIOLOGY

We now turn to models of population dynamics and epidemics. These typically involve difference or differential equations. We will start with one-species discrete models, discuss the occurrence of *chaos*, and consider populations with different age groups. We will then move on to continuous models, involving one or two species. The latter are too low-dimensional to exhibit any chaotic behavior, and we will concern ourselves with the appearance of oscillatory dynamics.

Contrarily to classical mechanics, we are still in the process of understanding the laws of biological evolution. Mathematical biology has recently become a topic of broad interest, and it is generally accepted that progress in comprehending the dynamics of populations, diseases, epidemics, etc can only be achieved through interdisciplinary activities involving both mathematicians and life scientists. Modelers should be able to know the significance of any term in their model equations and assess, by looking at experimental results, whether such terms are relevant to the problem in question. Successful modeling thus involves going back and forth between the model, its simulation and/or analysis, and experimental results.

Reliable biological data are difficult to obtain for a variety of reasons. First, in the case of population dynamics, data must be gathered over time periods much longer than the lifetime of an individual. At the human scale, this takes centuries. Fortunately, other systems, such as for instance bacterial systems, have a much shorter intrinsic time scale. Second, different studies are typically performed under different conditions, and the results may be affected by unknown confounding variables. It is also very rare to find very large-scale studies involving primates, in particular human subjects. As a consequence, trends may only become statistically significant when a variety of studies are combined. Finally, many biomedical studies are performed on mice or monkeys, and it is not always clear how the results may be transported to humans. More than ever, it is thus essential for a modeler to understand the nature and the limitations of the available data, as well as those of the models.

The examples discussed in this section are fairly simple, but they should be sufficient to help the reader develop an intuition for the basic generic modeling of population dynamics and epidemics in terms of difference or ordinary differential equations.

5.

SINGLE-SPECIES MODELS

Learning Objectives

At the end of this chapter, you will be able to do the following.

- Construct discrete-time models for isolated or interacting groups of individuals, including models with age groups.
- Solve linear discrete-time models exactly.
- Analyze the dynamics of first order nonlinear difference equations by assessing the stability of fixed points and periodic cycles.
- Describe the route to chaos in the logistic map.
- Recognize continuous-time models as limits of discrete-time models.

The population of Red-tailed Hawks in the US

We start by analyzing data collected by the Audubon Society through its Christmas Bird Count. Counts are currently available for every year since 1900. This information is discrete (we are given the number of birds that were counted at the end of each year), imprecise (since not all of the birds were counted and different people participate in the counting effort every year), and necessarily reflects a combination of many factors (in particular, not just growth or decay due to births and deaths). We will focus on the data for the red-tailed hawk, which is a short-distance migrant.

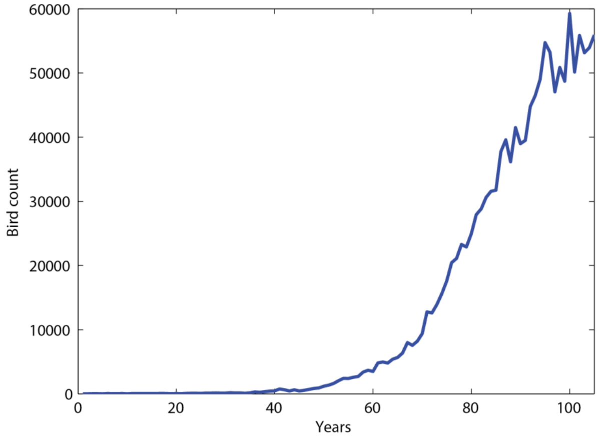


Figure 5.1. Red-tailed hawk bird count in the US, from 1900 (year 1) to 2004 (year 105), based on data collected by the Audubon Society.

Figure 5.1 shows the number of red-tailed hawks that were counted in the US, from 1900 (year 1) to 2004 (year 105). We can see that the population grew during most of last century, and seems to have begun to saturate in the late nineties. In order to understand this data, we may try to plot the rate of change of the population, defined as

$$\frac{\Delta N}{\Delta t} = \frac{N(t+\Delta t) - N(t)}{\Delta t},$$

where t is time in years, $N(t)$ is the bird count at time t, and $\Delta t = 1$ year in this case. Such a graph shows that the quantity $\Delta N / \Delta t$ fluctuates and the size of its fluctuations increases with time. Since the number of red-tailed hawks also increases as a function of time, we can look instead at the rate of change normalized by the number of individuals, which we denote by $R(t)$. We thus define the *per capita rate of change* of the red-tailed hawk population as

$$R(t) = \frac{N(t+\Delta t) - N(t)}{N(t)\Delta t}.$$

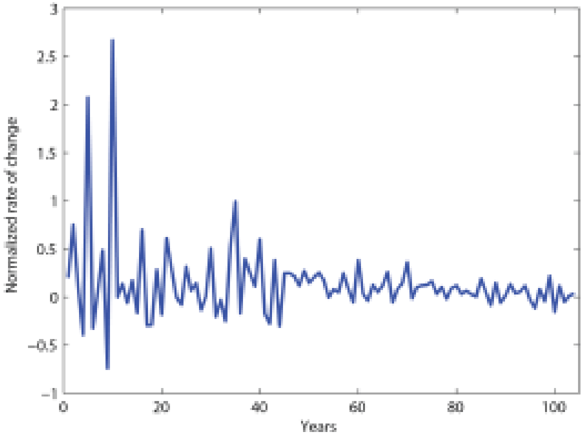


Figure 5.2. Plot of the quantity *R*(*t*) for the red-tailed hawk bird count, based on data collected by the Audubon Society.

Figure 5.2 shows $R(t)$ as a function of t. Fluctuations are of course still present, but they are now especially large only when $N(t)$ is relatively small. Moreover, the data appear to fluctuate about a mean value which is positive (although rather small). In the recent years, this trend breaks down, and from 1995 on, $R(t)$ seems to oscillate about a value closer to zero.

It is therefore reasonable to expect $N(t + \Delta t)$ to behave like a linear function of $N(t)$, up to some fluctuations, and to display saturation in the recent years. This is supported by the plot of Figure 5.3, which shows that $N(t + \Delta t) = N(t + 1)$ as a function of $N(t)$. We see that $N(t + 1)$ is very close to $1.025\ N(t)$. This slope was obtained by a least-square fit of the data with a straight line going through the origin (shown in red). As can be seen on Figure 5.3, the agreement is reasonably good. The value 1.025 is also quite close to the 95% confidence interval [1.026, 1.029] for the per-annum growth rate of red-tailed hawks in

North America, found in a 2016 article by C.S. Soykan *et al.*[1], which uses statistical analysis to estimate population trends from bird counts.

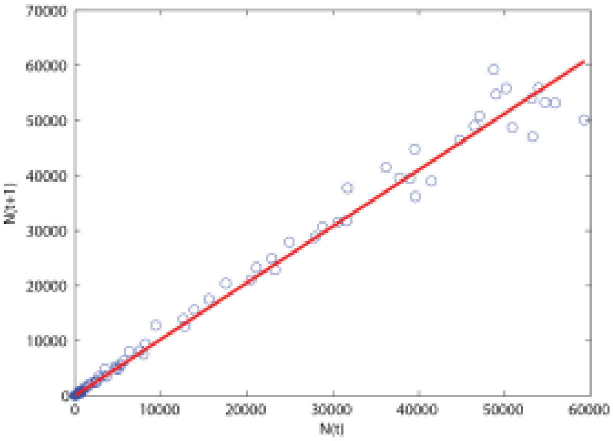


Figure 5.3. The first return map, showing $N(t + \Delta t)$ as a function of $N(t)$, for the Audubon Society red-tailed hawk bird count. The straight line has slope 1.025.

Note however that we do not expect – and in fact there is not – a good agreement at small values of N . However, the scale of Figure 5.3 is such that the scatter of the points near the origin is not apparent for $N(t) < 5,000$, which corresponds to times $t \leq 50$ (see Figure 5.1). In summary, even though the data is noisy and imprecise, a simple model for the red-tailed hawk population in the US from say 1950 to 1995 appears to be of the form

$$N(t+1) = \kappa\, N(t),$$

where κ is a constant close to 1.025. We can understand such a linear relationship as follows. If we neglect migration in and out of the United States – recall that the red-tailed hawk is a short distance migrant, we can write

$$N(t+\Delta t) = N(t) + \text{number of births} - \text{number of deaths}.$$

As a first approximation, it is reasonable to assume that both the number of deaths and the number of births may be written as the products of N with a function of N and t. Such an approximation is valid for most population dynamics models in a closed system. We thus define the *per capita birth rate* b, and the *per capita death rate* d as

$$b = \frac{\text{number of births per } \Delta t}{N(t)\Delta t}, \qquad d = \frac{\text{number of deaths per } \Delta t}{N(t)\Delta t}.$$

Then,

$$N(t+\Delta t) = N(t)\,(1 + b\Delta t - d\Delta t) = \kappa\, N(t),$$

$$R(t) = \frac{N(t+\Delta t) - N(t)}{N(t)\Delta t} = b - d.$$

[1] Candan U. Soykan, John Sauer, Justin G. Schuetz, Geoffrey S. LeBaron, Kathy Dale, Gary M. Langham, *Population trends for North American winter birds based on hierarchical models*, Ecosphere **7**, e01351 (2016); Table S3, line 206.

In a developmental phase, that is when the number of individuals is much lower than the maximum population that can be sustained by the environment, one can assume that b and d are constant, which gives

$$N(t+1) = \kappa\, N(t), \qquad \kappa = \text{constant}. \qquad (5.1)$$

For the red-tailed hawk, the above discussion suggests that $\kappa \simeq 1.025$, i.e. $b - d = 0.025$ individuals per year. Of course, fluctuations are always present in practice, but this simple model, known as *the Malthus equation in discrete time*, is sufficient to explain the exponential (or Malthusian) growth of the population. Indeed, Equation (5.1) implies that

$$N(t_0 + m\Delta t) = \kappa^m N(t_0).$$

If $\kappa > 1$, i.e. if births exceed deaths, $N(t)$ grows exponentially. Conversely, if $0 < \kappa < 1$, $N(t)$ decays exponentially and the species is driven towards extinction. We have assumed that $\kappa \geq 0$. If not, $N(t)$ would alternate between positive and negative values, and our population dynamics model would obviously be flawed.

When the number of individuals nears the carrying capacity of the environment, nonlinear effects can no longer be neglected, and the growth rate of the population changes with the number of individuals $N(t)$. For instance, logistic growth corresponds to

$$N(t+\Delta t) = \Big(1 + \kappa\big(N_\infty - N(t)\big)\Big) \cdot N(t) \qquad \kappa = \text{constant},$$

which has a N-dependent per-Δt growth rate equal to $\Big(1 + \kappa\big(N_\infty - N(t)\big)\Big)$. The parameter N_∞ represents the carrying capacity of the environment.

First-order difference equations

The example discussed above suggests that discrete population models are of the form $N(t+\Delta t) = f(N(t))$, or equivalently,

$$N_{k+1} = f(N_k), \qquad k \in \mathbb{N}, \qquad (5.2)$$

where $N_k = N(k\Delta t)$. Equation (5.2) is called a *first order difference equation*. It is *linear* if $f(x)$ is linear in x, and has *constant coefficients* if f does not depend on k. In the context of population dynamics, a model like Equation (5.2) is sometimes called a *metered model*. Equation (5.1), which can be written as

$$N_{k+1} = \kappa\, N_k,$$

is therefore a linear first order difference equation with constant coefficients. As mentioned above, its solution is

$$N_k = \kappa^k \; N_0,$$

where N_0 is the initial condition.

More generally, one can write

$$N_k = f^k(N_0),$$

where f^k is the k^{th} iterate of f. Even if the graph of f is complicated, it is possible to visualize the dynamics of N_k by iterating f graphically, as shown in Figure 5.4. Starting from the initial value N_0 of N, we first find $f(N_0)$ as the y-coordinate of the intersection of the graph of f with the line $x = N_0$. We then mark the value of $N_1 = f(N_0)$ on the x-axis by drawing a vertical line through the intersection of the line $x = y$ with the line $y = f(N_0)$. Finally, we repeat this process to find the successive iterates of f. In the example of Figure 5.4, iterates starting at $N = N_0$ converge towards a *fixed point* $N = N_s$.

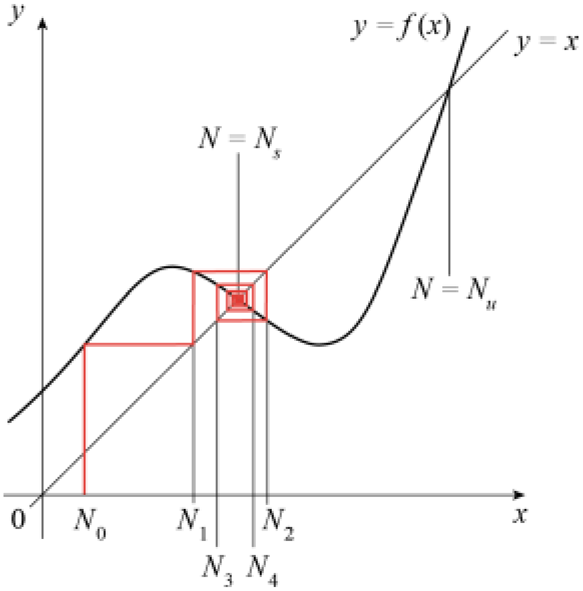


Figure 5.4. Graphical representation of successive iterations of the map $N_{k+1} = f(N_k)$. The first four iterates N_1 to N_4 of the function *f*, starting at $N = N_0$, are also indicated. The symbols N_s and N_u refer to stable and unstable fixed points, respectively.

Fixed points and their linear stability

Fixed points N_c of the map (5.2) are such that $N_c = f(N_c)$, and are found by intersecting the line of equation $y = x$ with the curve $y = f(x)$. In the example of Figure 5.4, the function f has two fixed points, $N = N_s$ and $N = N_u$. We can study their *linear stability* as follows. Let $N = N_c + u$, where u is a small perturbation. Then,

$$f(N) = f(N_c + u) = f(N_c) + uf'(N_c) + O(u^2),$$

and

$$f(N) - N_c = f(N) - f(N_c) = uf'(N_c) + O(u^2) \simeq uf'(N_c),$$

the last approximation being appropriate if u is small enough. The linear map $u_{k+1} = f'(N_c)u_k$ is such that its iterates converge towards zero if $|f'(N_c)| < 1$, and diverge if $|f'(N_c)| > 1$. The fixed point

$N = N_c$ is thus said to be *linearly stable* if $|f'(N_c)| < 1$, and *unstable* if $|f'(N_c)| > 1$. If $|f'(N_c)| = 1, N = N_c$ is *marginally stable*.

The logistic map

As an example, consider the logistic map defined by Equation (5.2), with $f(x) = a\,x\,(1-x)$, where a is a parameter such that $0 < a < 4$. If $a \leq 1$, then $x = 0$ is the only non-negative fixed point. It is stable for $a < 1$ and marginally stable if $a = 1$. For $a > 1$, there are two fixed points, $x = 0$ and $x = 1 - 1/a$. Linear stability analysis shows that $x = 0$ is now unstable since $a > 1$, and that $x = 1 - 1/a$ is linearly stable if $a < 3$ and unstable if $a > 3$ (see exercises). A numerical exploration of the dynamics of the difference equation $x_{n+1} = a\,x_n\,(1 - x_n)$ reveals the presence of attractors of increasing complexity as a is increased, such as period-two cycles (for instance at $a = 3.2$), period-four cycles (e.g. at $a = 3.5$), and period-eight cycles (for instance at $a = 3.55$). In fact, a *period-doubling* cascade occurs as a is increased, and the dynamics becomes chaotic near $a = 3.57$.

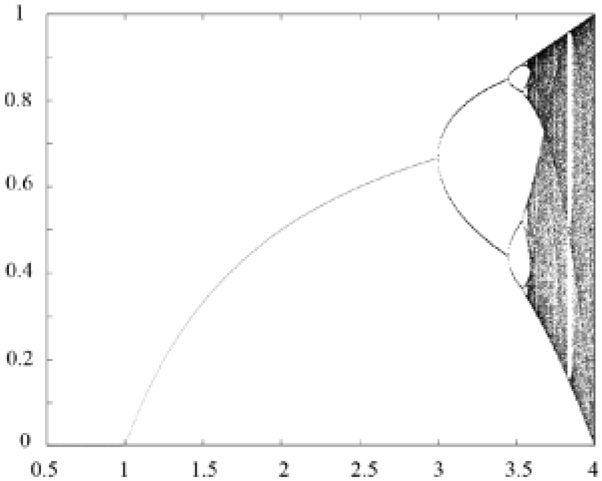


Figure 5.5. Bifurcation diagram of the logistic map.

Figure 5.5 shows the *bifurcation diagram* of the logistic map. As mentioned above, for a given value of a, successive iterations of f converge towards an attractor, for instance a fixed point, a period-two cycle, or a much more complicated structure. The bifurcation diagram of f gives a representation of this attractor as a function of a. It is numerically obtained by plotting the set of points

$$\{x_k,\ k_{min} < k < k_{max}\},$$

for different values of a. Here, the initial condition is a point $x_0 \in (0,1)$, say $x_0 = 1/2$, k_{min} is large enough so that the dynamics is sufficiently close to its attractor after k_{min} iterations, and k_{max} depends on the available computing power. The bifurcation diagram shown in Figure 5.5 was calculated with $k_{min} = 1500$ and $k_{max} = 1700$.

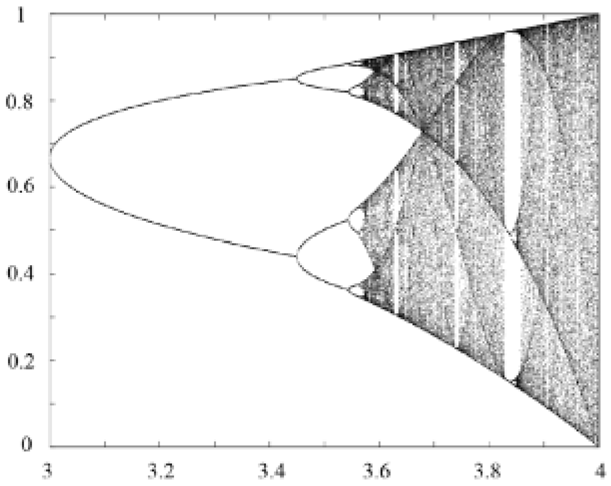


Figure 5.6. Bifurcation diagram of the logistic map, for $3 \leq a \leq 4$.

An enlargement for $3 \leq a \leq 4$ is shown in Figure 5.6. The cycles of period 2, 4 and 8 are clearly visible. Period 3 and 6 windows for larger values of a are also easily detected.

A period-n cycle may be analyzed by looking at fixed points of f^n. Of course, as n increases it quickly becomes difficult to solve for such fixed points. However, it can be shown that all of the period-doubling bifurcations occur in a similar, universal, fashion (see articles by P. Coullet & C. Tresser, 1978[2] and by M. Feigenbaum, 1980[3]). For a simple description of the various types of instabilities occurring in the logistic and other iterated maps, the reader is referred to the 1976 article by Robert May entitled *Simple mathematical models with very complicated dynamics* and to the exercises at the end of this chapter.

Continuous approximation of a one-species discrete model

The discrete model discussed in the first section is such that

$$\frac{N(t+\Delta t)-N(t)}{\Delta t} = N(t)R(t).$$

As $\Delta t \to 0$, the difference quotient on the left-hand-side can be approximated by dN/dt, and we thus obtain the following continuous approximation

$$\frac{dN}{dt} = R(t)N(t). \qquad (5.3)$$

Such an approximation is meaningful only if one can think of N as a differentiable function of t. In particular, N must be continuous. Equation (5.3) will thus be a reasonable model for a population only if N is large. In the case of a phenomenological model, if stochastic effects are negligible and N approximates an actual number of individuals, the latter should be close to the integer $\mathrm{round}(N(t))$ nearest to $N(t)$. Another possibility is to define $N(t)$ as a density, i.e. an average number of individuals per surface area for instance.

[2] P. Coullet and C. Tresser, *Itérations d'endomorphismes et groupe de renormalisation* J. Phys. Colloques **39**, C5: 25-28 (1978)

[3] Mitchell J. Feigenbaurn, *Universal Behavior in Nonlinear Systems*, Los Alamos Science **Summer 1980**, 4-27 (1980)

Alternatively, one may view $N(t)$ as the expected value of a large population. In that context, it is useful to think of $R(t)\delta t + O(\delta t^2)$ as a per-capita probability of growth, so that the expected change in N during the small amount of time between t and $t + \delta t$ is $R(t)N(t)\delta t + O(\delta t^2)$. In the rest of these notes, we mainly use continuous models and thus implicitly assume that the relevant quantities are adequately described by continuous variables.

Linear model

If $R(t) = b - d \equiv r$ with b and d constant, Equation (5.3) reads

$$\frac{dN}{dt} = r\,N(t) \qquad (5.4)$$

and its solution is $N(t) = \exp(rt)N_0$, where N_0 is the population at $t = 0$. If $r > 0$, the population grows exponentially; if $r < 0$, $N(t)$ decays to zero and the species goes extinct. There is thus a direct analogy between the two linear models given by Equations (5.1) and (5.4), which is further explored in the exercises. Equation (5.4) is known as *the Malthus equation in continuous time*.

The Logistic model

Consider now the logistic model, given by

$$\frac{dN}{dt} = aN(N_c - N),$$

where N_c and a are parameters. This equation, known as *the Verhulst equation*, can be simplified into

$$\frac{dM}{dt} = \lambda M(1 - M) \quad (5.5)$$

by making the change of variables $M = \dfrac{N}{N_c}$ and $\lambda = aN_c$. Note that M is dimensionless and $[\lambda] = 1/T$. We could thus rescale the time variable and obtain a parameter-free model, but we will not do this now. The solution of Equation (5.5) with initial condition $M(0) = M_0$ is $M(t) = \left[1 + \dfrac{1 - M_0}{M_0}\exp(-\lambda t)\right]^{-1}$, for all values of the parameter λ. As $t \to \infty$, $M \to 1$ if $\lambda > 0$, for all values of M_0. In terms of the original variables, we have $N \to N_c$. The quantity N_c is called the *carrying capacity* of the environment. It corresponds to an equilibrium state in which the number of deaths balances the number of births in order to sustain a population in the presence of limited resources. If $\lambda < 0$

, then $M \to 0$ if $M_0 < 1$, and $M \to \infty$ for $t \to t^*$, where $t^* = \frac{1}{-\lambda} \ln\left(\frac{M_0}{M_0 - 1}\right)$, if $M_0 > 1$. In this case, the solution therefore diverges in finite time. The dynamics of M is monotonic for all values of λ , and is therefore completely different from that of the discrete logistic map, for which chaos may be observed.

General autonomous model

More generally, consider the differential equation

$$\frac{dN}{dt} = f(N), \qquad (5.6)$$

where f is a nonlinear function of N. In the context of population dynamics, $f(N)$ is called the *net growth rate of the population*. It is often written as $f(N) = NR(N)$, where $R(N)$ is the *net per capita growth rate* of the population. Since the right-hand-side of Equation (5.6) does not depend on t, this differential equation is said to be *autonomous*.

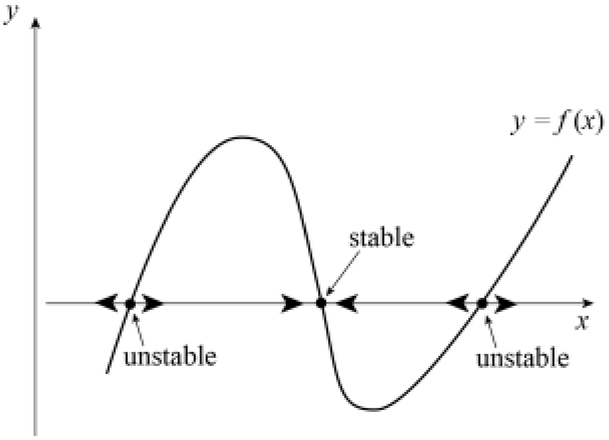


Figure 5.7. The fixed points of Equation (5.39) are the values of N at which the graph of $f(N)$ intersects the x-axis. The stability of a fixed point N_c can be inferred from the behavior of $f(N)$ near $N = N_c$.

The fixed points N_c of (5.6) are given by $f(N_c) = 0$, and can be obtained graphically by looking at where the graph of f intersects the horizontal axis. The *nonlinear* stability of these fixed points can also be assessed graphically. Indeed, $f(N)$ changes sign at $N = N_c$. We know that if $f(N) > 0$ for $N < N_c$, then N will grow towards N_c, whereas if $f(N) > 0$ for $N > N_c$, then N will move away from N_c. The stability of any fixed point N_c of f can thus be inferred by adding arrows to the horizontal axis that point to the left where f is negative and to the right where f is positive. A stable fixed point N_c will have arrows pointing towards it on both sides, whereas an unstable fixed point will be associated with arrows pointing away from it. This is illustrated in Figure 5.7.

Discrete models with age distribution

Models with age distribution describe the number of individuals in each age group of a population and allow to take into account different birth and death rates for different subgroups. Consider for instance a three-group model, in which $N_1(t)$ is the number of children at time t; $N_2(t)$ is the number of adults of child-bearing age, and $N_3\,(t)$ is the number of older adults. A simple set of difference equations describing the dynamics of such a population is as follows,

$$N_1(t+\Delta t) = N_1(t) + b_2 N_2(t)\Delta t - d_1 N_1(t)\Delta t - g_1 N_1(t)\Delta t$$

$$N_2(t+\Delta t) = N_2(t) + g_1 N_1(t)\Delta t - d_2 N_2(t)\Delta t - g_2 N_2(t)\Delta t$$

$$N_3(t+\Delta t) = N_3(t) + g_2 N_2(t)\Delta t - d_3 N_3(t)\Delta t,$$

where d_i is the per capita death rate of group i, g_i is the rate at which surviving individuals move out of group i (alternatively, $g_i\,\Delta t$ represents the probability that an individual in group i will move to group $i+1$ during Δt), and b_2 is the per capita rate at which individuals in group 2 give birth. The period of time Δt could be one year or five years for human populations. It should be shorter than the typical timescales of the system being modeled, such as the time it takes for one child to become an adult, or the time it takes for a young adult to move to group 3.

The above system is linear and can thus be conveniently rewritten as a difference equation for the vector $\vec{C} = (N_1, N_2, N_3)^T$, which reads

$$\begin{pmatrix} N_1 \\ N_2 \\ N_3 \end{pmatrix}(t+\Delta t) = A \begin{pmatrix} N_1 \\ N_2 \\ N_3 \end{pmatrix}(t), \qquad (5.7)$$

with $A = \begin{pmatrix} 1-(d_1+g_1)\Delta t & b_2\Delta t & 0 \\ g_1\Delta t & 1-(d_2+g_2)\Delta t & 0 \\ 0 & g_2\Delta t & 1-d_3\Delta t \end{pmatrix}$. The matrix A typically has constant, non-negative entries. It is called a *Leslie matrix* in the context of population dynamics models. We can look for a solution to Equation (5.7) in the form $\vec{C}(t_0 + m\Delta t) = r^m \vec{C}(t_0)$, where t_0 is some initial time. Substitution of this expression into Equation (5.7) implies that $A\vec{C}(t_0) = r\,\vec{C}(t_0)$, i.e. $\vec{C}(t_0)$ is an eigenvector of A with eigenvalue r. If the matrix A is diagonalizable, the general solution of Equation (5.7) may thus be written as

$$\vec{C}(m\,\Delta t) = a_1 r_1^m \vec{C}_1 + a_2 r_2^m \vec{C}_2 + a_3 r_3^m \vec{C}_3,$$

where the $\vec{C}_i$ are three linearly independent eigenvectors of A with associated eigenvalues r_i, and the a_i are coefficients imposed by the initial condition $\vec{C}(0) = \vec{C}_0$, where $\vec{C}_0$ is known. Note that since A has real entries, its eigenvalues are either all real or consist of a real eigenvalue and a complex conjugate pair of eigenvalues. In this latter case, two of the $\vec{C}_i$, say $\vec{C}_2$ and $\vec{C}_3$, are also complex conjugates and a real initial condition leads to $a_3 = a_2^*$, where the star denotes complex conjugation. The dynamics of Equation (5.7) depends on the position of the eigenvalues of A in the complex plane, relative to the unit circle. Indeed, if $|r_i| > 1$, then exponential growth will be observed in the direction of $\vec{C}_i$, with fastest and therefore dominant growth along the eigendirection associated with the principal eigenvalue of A. Conversely, if all of the eigenvalues of A have modulus less than 1, then $\vec{C}(m\,\Delta t)$ will converge towards zero. More complex behaviors, including chaos, may be observed if nonlinear effects are taken into account.

The LPA Model

Depending on the type of species whose populations is being modeled, effects other than births and deaths may have to be considered. As an illustration, we now discuss a model developed by R.F. Costantino and colleagues[456] for a population of flour beetles, for which adults are known to eat their offspring. The different stages in the development of a beetle are the larval, pupal and adult stages. If one chooses Δt (about 2 weeks for flour beetles) such that it corresponds to the time it takes on average for a larva to pupate and for a pupa to become a reproductive adult, one has, counting time in units of Δt, and in the absence of cannibalism,

$$\begin{cases} L(t+1) = bA(t), \\ P(t+1) = L(t) - d_l L(t), \\ A(t+1) = A(t) - d_a A(t) + P(t), \end{cases}$$

where the positive constants d_l and d_a are the probabilities of death for larvae and adults during the period $\Delta t\ (= 1)$, the parameter $b > 0$ is the average number of eggs per adult that have hatched during Δt, and one has neglected deaths in the pupal stage.

Cannibalism of eggs by adults and larvae is modeled through multiplicative terms of the form $\exp(-c_{ea}A(t))$ and $\exp(-c_{el}L(t))$ respectively, and cannibalism of pupae by adults is described by $\exp(-c_{pa}A(t))$ (see the articles by Costantino et al. and Cushing et al. mentioned above). Here, the constants c_{ea}, c_{el} and c_{pa} are all positive. The exponential terms represent the fraction of eggs or pupae which survive to the next stage in their development, and can be understood as follows. Beetles in a crawling stage (larvae and adults) move through the flour and may encounter individuals in a non-moving stage, such as eggs and pupae. When this happens, the moving larva or adult beetle will bite the egg or pupa, thereby killing it.

[4] R.F. Costantino, J.M. Cushing, B. Dennis, and R.A. Desharnais, *Experimentally induced transitions in the dynamics behaviour of insect populations*, Nature **375**, 227-230 (1995).

[5] R.F. Costantino, R.A. Desharnais, J.M. Cushing, and B. Dennis, *Chaotic dynamics in an insect population*, Science **275**, 389-391 (1997).

[6] J.M. Cushing, R.F. Costantino, B. Dennis, R.A. Desharnais, and S.M. Henson, *Nonlinear population dynamics: models, experiments and data*, J. Theor. Biol. **194**, 1-9 (1998).

We assume that larvae only kill eggs, since pupae are bigger than larvae. Suppose the probability that an adult encounters one egg during the period of time Δt is given by $p\,\Delta t$, where p is constant. Then, the probability for this egg not to have been eaten by this adult at the end of Δt is $1 - p\Delta t$, and the probability for this (or any other) egg to become a larva is then $(1 - p\Delta t)^{A(t)}$, since there are $A(t)$ adults. Thus, if eggs were only eaten by adults, one would write

$$L(t+\Delta t) = bA(t)(1-p\Delta t)^{A(t)} = bA(t)\exp(A(t)\ln(1-p\Delta t)) = bA(t)\exp(-c_{ea}A(t)),$$

where $c_{ea} = -\ln(1 - p\Delta t)$. Following similar arguments, the fact that eggs are also eaten by larvae is taken into account by multiplying the right hand side of the above equation by $\exp(-c_{el}L(t))$. Therefore, the *deterministic LPA model* (where LPA stands for Larva, Pupa, Adult), which includes cannibalism of pupae and eggs by adults and of eggs by larvae, reads

$$\begin{cases} L(t+1) = bA(t)\exp(-c_{ea}A(t) - c_{el}L(t)), \\ P(t+1) = L(t) - d_l L(t), \\ A(t+1) = A(t) - d_a A(t) + P(t)\exp(-c_{pa}A(t)). \end{cases}$$

A review of the dynamical properties of the LPA model can be found in the 2004 article by Jim Cushing *et al.* entitled *Nonlinear population dynamics: models, experiments and data*. In particular, it can be shown that the trajectories of the LPA model which start in the non-negative octant remain non-negative. Moreover, when stochastic terms are added to the above equations, the predictions given by the resulting model, including behavior consistent with chaotic dynamics, are in particular good agreement with experiments.

Summary

This chapter discusses discrete and continuous models for the dynamics of a single species. We first introduced exponential growth in the context of the Audubon Society's count of red-tailed hawks in the United States, and added nonlinear saturation to the corresponding model. We then discussed fixed points and periodic cycles of one-dimensional nonlinear maps and saw that simple nonlinear difference equations, such as the logistic map, could have very complex dynamics and exhibit chaos. We also described linear and nonlinear population models with age groups. An example of the latter is the LPA model, whose stochastic version gives an accurate representation of experimental observations. Finally, we introduced continuous models as approximations of discrete models, discussed the global stability of their fixed points, and pointed out that one-dimensional (as well as two-dimensional) continuous models have very simple dynamics. In particular, they cannot exhibit chaos.

The methods and techniques introduced in this chapter generalize to more complex situations. The exercises below discuss models with continuous age distributions, delay and harvesting, as well as the evolution of the population of the United States.

Food for Thought

Problem 1

Assume that a population grows according to $N(t+1) = \kappa N(t), \kappa > 1.$

1. How long does it take to double the number of individuals?
2. Estimate the value of κ from Figure 5.1, which shows the number of Red-tailed Hawks in the United States as a function of time.
3. Is your estimate of κ in reasonable agreement with the value $\kappa = 1.025$ inferred from Figure 5.3?

Problem 2

Consider the continuous model $\dfrac{dN}{dt} = rN, r > 0.$

1. How long does it take to double N?
2. If this model is an approximation of a difference equation of the form $N(t + \Delta t) = \kappa N(t),$ what is the relationship between κ and r?
3. How long does it take for a system described by the discrete model to double its population?
4. Is the discrete model faster or slower than its continuous approximation?

Problem 3

Consider the continuous model $\dfrac{dN}{dt} = r(N)N,$ where $r(N)$ is a linear function of N: $r(N) = aN + b$.

1. What should the signs of a and b be if one wants the population to grow for small values of N and saturate at large values of N?
2. What changes of variables should you make to turn this model into Equation (5.5)?

Problem 4

Find the non-trivial fixed point of the logistic map

$$x_{n+1} = ax_n(1 - x_n) \equiv f(x_n), 0 < a < 4.$$

1. Show that this fixed point becomes unstable when $a > 3$.
2. Show that as soon as this fixed point becomes unstable, a period-two cycle of the map starts to exist. Hint: look for fixed points of f^2.
3. Find the value of a at which this period-two cycle becomes unstable.
4. Check your answer against the bifurcation diagram of Figure 5.6.

Problem 5

Solve the difference equation $X(t+1) = \begin{pmatrix} 2 & 2 \\ 1 & 3 \end{pmatrix} X(t)$, where $X(t)$ has two components: $X(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$. Assume the initial conditions are $x(0) = 1$ and $y(0) = 0.$

Problem 6

Consider the differential equation $\dfrac{dx}{dt} = x(x-1)(x-2).$

1. What are the fixed point of the dynamics?
2. Are they stable or unstable?

Problem 7

Consider the differential equation $\dfrac{dx}{dt} = x^2(x-1)$.

1. What are the fixed point of the dynamics?
2. Are they stable or unstable?

Problem 8

Consider the differential equation $\dfrac{dx}{dt} = a + x^2(x-1) \equiv f(x), a \in \mathbb{R}$.

1. Plot $f(x)$ as a function of x. What happens as a changes?
2. Without performing any calculation, sketch the behavior of the fixed points as a function of a.
3. Indicate the stability of each fixed point on the bifurcation diagram of question (2).
4. Assign a value of +1 to each stable fixed point and a value of -1 to each unstable fixed point. For each value of a, define a function $g(a)$ as the sum of indices associated with the fixed points existing for this particular value of a. Sketch the graph of g as a function of a. What do you observe?

Problem 9

Consider a population model where the age a is a continuous independent variable. Define the age distribution $M(a,t)$ such that the number of individuals between age a_1 and age a_2 at time t is given by

$$\int_{a_1}^{a_2} M(a,t)da.$$

Show that the dynamics of M is described by the following partial differential equation, known as the *McKendrick* or *von Foerster* partial differential equation,

$$\frac{\partial M}{\partial t} + \frac{\partial M}{\partial a} = -d(a)M,$$

where the mortality function $d(a)$ represents the per capita death rate of individuals of age a.

Problem 10

Use the method of characteristics to solve the McKendrick partial differential equation derived in Problem 9.

Problem 11

Consider the LPA model described in this chapter. What are the dimensions of the parameters (b, c_{ea}, c_{el}, c_{pa}, d_a, d_l) appearing in this model?

Problem 12

Write a model describing a situation analogous to that of the LPA model (with cannibalism), but such that the time for a pupa to become an adult is twice as long as the time it takes for a larva to pupate.

Problem 13

Write a model describing a population with two subgroups, juvenile and adults, in which adults eat some of their own eggs. You can use an exponential term similar to that appearing in the LPA model to describe cannibalism.

Problem 14

Consider the following model,

$$\begin{cases} J(t+\Delta t) = bA(t)\exp(-cA(t)), \\ A(t+\Delta t) = (1-d_J)J(t) + (1-d_A)A(t), \end{cases}$$

where b, d_J and d_A are positive parameters.

1. Describe in words a situation modeled by the above equations.
2. Under what condition is there a non-trivial fixed point for this model? What is the biological significance of this condition?
3. Discuss the stability of the fixed points of this model.

Problem 15

Read the article by R. May, entitled *Simple mathematical models with complicated dynamics*, and address (i.e. explain, justifies, work out the details of, or prove) the following statements found in this paper.

1. Page 460, left column, above Equation (3): "*By writing $X = bN/a$, the equation may be brought into the canonical form $X_{t+1} = aX_t(1-X_t)$.*"
2. Page 460, left column, bottom: "*If X ever exceeds unity, subsequent iterations diverge towards $-\infty$.*"
3. Page 460, right column, below Equation (6): "*So long as this slope lies between 45^o and -45^o … the equilibrium point X^* will be at least locally stable.*"
4. Page 461, left column, below Equation (9): "*Clearly, the equilibrium point X^* of Equation (5) is a solution of Equation (9).*"
5. Equation (10): $\lambda^{(2)}(X^*) = [\lambda^{(1)}(X^*)]^2$.
6. Page 461, left column, bottom: "*As this happens, the curve $F^{(2)}(X)$ must develop a "loop", and two new fixed points of period 2 appear.*"
7. Page 461, left column, bottom: "*This slope is easily shown to be the same at both points, and more generally to be the same at all k points on a period k cycle.*"
8. Bottom of page 461 and beginning of page 462: "*… until at last the three-point cycle appears (at $a = 3.8284\ldots$ for equation (3)).*"
9. Page 462, right column, bottom: "*This period 3 cycle is never stable.*"

10. Page 462, right column, bottom: "*As* $F(X)$ *continues to steepen, the slope* $\lambda^{(3)}$ *for this initially stable three-point cycle decreases beyond -1; the cycle becomes unstable and gives rise by bifurcation process … to stable cycles of period 6, 12, 24, …,* 3×2^n."
11. Draw pictures illustrating the concepts of tangent and pitchfork bifurcations (see page 463, top of left column).
12. Page 464, right column, top: "*… the slope of the* k*-time iterated map* $F^{(k)}$ *at any point on a period* k *cycle is simply equal to the product of the slopes of* $F(X)$ *at each of the points* X^*_k *on this cycle.*"
13. Page 465, left column, bottom: "*… as each new pair of cycles is born by tangent bifurcation (see Fig. 5), one of them is at first stable, by virtue of the way smoothly rounded hills and valleys intercept the 45° line.*"
14. Page 466, right column, bottom: "*… in continuous two-dimensional systems … dynamic trajectories cannot cross each other.*"

Problem 16

The goal of this problem is to explore global changes in the population of the United States. The U.S. Census Bureau maintains a file (popclockest.txt) containing national population estimates between 1900 and 1999. Import this data set into MATLAB or EXCEL. Then answer the following questions.

1. Plot the U.S. population as a function of time. What do you conclude?
2. Can the growth of the U.S. population be modeled by a simple evolution equation of the form $N(t+1) = (1+R)N(t)$, where t is in years? Why or why not? If so, estimate R .
3. Post-census population estimates are obtained as described on the U.S. Census Bureau methodology page (see for instance the methodology file for the 2021 vintage). Explain the main formula given in the Overview section of this article.
4. Given the following estimates[7], find the population of the U.S. in 2004:
    1. Population in 2001: 285,102,075.
    2. Births, deaths, and net international immigration:

[7] Downloaded in 2005 from a now decommissioned US Census Bureau web page

- 2001-2002: 4,006,985; 2,429,999; 1,262,159.
- 2002-2003: 4,055,469; 2,432,874; 1,225,161.
- 2003-2004: 4,099,399; 2,453,984; 1,221,013.

Problem 17

The goal of this section is to explore the evolution of the population of the United States using different age groups. Population estimates by five-year age groups from 2010 to 2019 can be obtained from the U.S. Census Bureau web site (each year has its own table).

1. Use this information to plot the age distribution of the U.S. population for different years.
   1. Has there been major changes in the last 4 years of this data set?
   2. The data set has 18 age groups. Use the age distributions that you just plotted to define larger age groups that can be used in a simplified model.
2. Using recent birth and death rates estimates, as published by the Center for Disease Control and Prevention, create a model to predict the population in the age groups you defined, taking the 2010 data as initial condition. You may want to start with Figures 2 and 3 of the birth and deaths documents, respectively. There are also tables at the end that may be of use. (Note: do not attempt to print these files; they are more than 50 pages long!)
   1. How does your model compare to the Census Bureau estimates for 2019?
   2. Use your model to predict the population in each age group in 2050. What do you conclude?
   3. Discuss the limitations of your model.

6.

TWO-SPECIES MODELS

Learning Objectives

At the end of this chapter, you will be able to do the following.

- Create coupled differential equation models for predator-prey and competing species systems.
- Formulate dimensionless versions of two-dimensional models and appraise their dynamics by means of phase plane analysis.
- Translate the results of model analysis into biological terms and discuss their significance in that context.

Predator-prey models: the Lotka-Volterra equations

Consider a closed system involving a population of predators (e.g. sharks) and prey (e.g. little fish). Let F represent the density of prey and S that of the predator. It is clear that if there is abundant prey, predators will proliferate. As a consequence, more prey will be eaten and F will decrease, but so will S since there will be less food for the predators. However, if S decreases, F will have a chance to grow again since the prey will not be eaten as much. This reasoning suggests that parameter values may exist for which F and S oscillate in time[1][2].

[1] A.J. Lotka, *Elements of Physical Biology*, Williams & Wilkins, Baltimore, 1925; Dover, New York, 1956.

[2] V. Volterra, *Leçons sur la Théorie Mathématique de la Lutte pour la Vie*, Gauthier-Villars, Paris, 1931.

If we neglect migration, and assume that F and S are the only dependent variables – in particular we assume that there is an ample (but not infinite) supply of nutrients for the prey, we may write

$$\begin{cases} \dfrac{dF}{dt} = F\left(\alpha - \beta F - \gamma S\right) \equiv g(F,S), \\ \dfrac{dS}{dt} = S\left(-\kappa + \delta F\right) \equiv h(F,S), \end{cases} \tag{6.1}$$

where the constant parameters α, β, γ, κ and δ are all non-negative. This model indicates that when $S = 0$, the prey population evolves according to a logistic model, and that when $S \neq 0$, prey are eaten proportionally to both their own abundance (F) and that of the predators (S). The predators have a growth (birth minus death) rate equal to $-\kappa$ if $F = 0$. Since κ is positive, this means that in the absence of prey, the predator population would be driven to extinction. The growth rate of S is then corrected by a linear term $\delta\, F\, S$, proportional to F.

If $\beta = 0$, this model is called the Lotka-Volterra system (see footnotes 1 and 2 for references) and is the simplest continuous model that incorporates the basic factors describing the interaction of a predator and its prey. It should be considered as a starting point for more complicated and more realistic models. The rest of this section is devoted to an analysis of Equations (6.1).

Scalings

Equations (6.1) involve three variables, F, S and t, and five parameters, which we assume are positive (except possibly for β). Scaling all of the variables should allow us to reduce this system to a problem with two (i.e. five minus three) parameters. A typical value for F is given by $F_0 = \kappa/\delta$ (see the second equation of (6.1)), and a typical value for S can be chosen as $S_0 = \alpha/\gamma$ in a similar fashion. For a characteristic time, we may take $\tau = 1/\kappa$. Then, if we define $f = \dfrac{F}{F_0}$, $s = \dfrac{S}{S_0}$, and $\tau = \kappa t$, we obtain a dimensionless, canonical model, which reads

$$\begin{cases} \dfrac{df}{d\tau} = af(1 - bf - s), \\ \dfrac{ds}{d\tau} = s(-1 + f), \end{cases} \tag{6.2}$$

Here, the parameters a and b are related to the original parameters by $a = \dfrac{\alpha}{\kappa}$, $b = \dfrac{\beta\kappa}{\alpha\delta}$, and are therefore positive.

Phase plane analysis

A description of the phase plane associated with system (6.2) gives us a qualitative understanding of the predator-prey equations. With this information, we are able to address questions such as those listed below, even if we do not have explicit forms for all of the solutions of (6.2).

- Is it possible for the populations of sharks or little fish to go extinct?
- Does the model make biological sense?
- Can oscillations be observed?

Before going any further, it is important to check that the model is "biologically well-posed," i.e. that if s and f are initially non-negative, they will remain so. Assume for instance that $s = 0$. Then, the second equation of (6.2) indicates that s will remain zero. Trajectories in the (f, s) plane are therefore along the f axis or away from it, but do not cross this axis. Similarly, the first equation of (6.2) shows that trajectories do not cross the s axis.

We now turn to a description of the phase portrait of system (6.2). The fixed points of this system are given below in terms of their coordinates in the (f, s) plane. They are

$$P_0 = (0, 0), \qquad P_1 = \left(\frac{1}{b}, 0\right), \qquad P_2 = (1, 1 - b),$$

and are all located in the first quadrant provided $0 < b \leq 1$, which we will assume is true. Linear stability analysis can be used to describe the local dynamics of (6.2) near its three fixed points. The Jacobian of (6.2) is given by

$$J(f, s) = \begin{pmatrix} a(1 - 2bf - s) & -af \\ s & -1 + f \end{pmatrix}.$$

Analysis near the origin, P_0

For P_0, $s = f = 0$ and the Jacobian $J(0, 0)$ is diagonal. Its eigenvalues are a and -1, with associated eigenvectors $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ respectively. The origin is therefore a saddle point.

Analysis near P_1

For P_1, $f = 1/b$ and $s = 0$. The Jacobian $J(1/b, 0)$ is an upper-triangular matrix and its eigenvalues are therefore its diagonal entries. They are $-a$ and $-1 + 1/b$. Since $b \leq 1$, this fixed point is also a saddle (with possibly a neutral direction if $b = 1$). The eigenspace associated with $-a$ is spanned by $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, and the eigenspace associated with $-1 + 1/b$ is spanned by $\begin{pmatrix} 1 \\ (-ab + b - 1)/a \end{pmatrix}$.

Analysis near P_2

For P_2, $f = 1$ and $s = 1 - b$. The Jacobian $J(1, 1-b)$ reads

$$J(1-b, 1) = \begin{pmatrix} -ab & -a \\ 1-b & 0 \end{pmatrix},$$

and its eigenvalues λ_1 and λ_2 are such that

$$\det(J) = \lambda_1 \lambda_2 = \det\left[J(1-b, 1)\right] = a(1-b) > 0,$$

and

$$\mathrm{Tr}(J) = \lambda_1 + \lambda_2 = \mathrm{Tr}\left[J(1-b, 1)\right] = -ab < 0.$$

Therefore P_2 is either a stable node or a stable spiral.

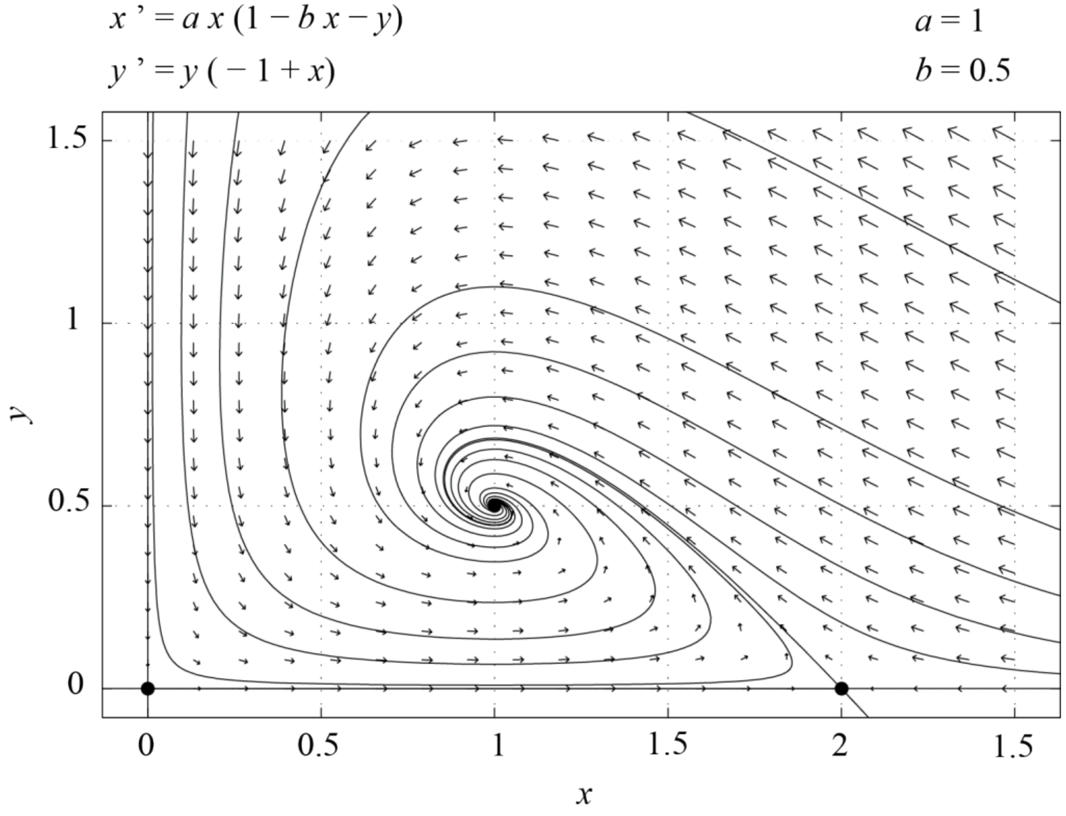


Figure 6.1. Phase plane of model (6.2) with $a = 1$ and $b = 0.5$, plotted with the software PPLANE.

This analysis thus indicates that if the initial values of f and s are positive, then the dynamics will converge to the fixed point P_2, for which both shark and fish populations are of finite size. In other words, no periodic oscillations are observed. Figure 6.1 shows the phase plane of system (6.2) for $a = 1$ and $b = 0.5$, obtained with PPLANE. In this case, the fixed point P_2 is a stable spiral.

As an exercise, the reader should work out the case where $b > 1$, for which only two of the fixed points are in the first quadrant. Then, P_1 is a stable node and all initial conditions with $f > 0$ and $s > 0$ converge to P_1. In this situation, the sharks are extinct and the fish have reached the carrying capacity of the environment they live in.

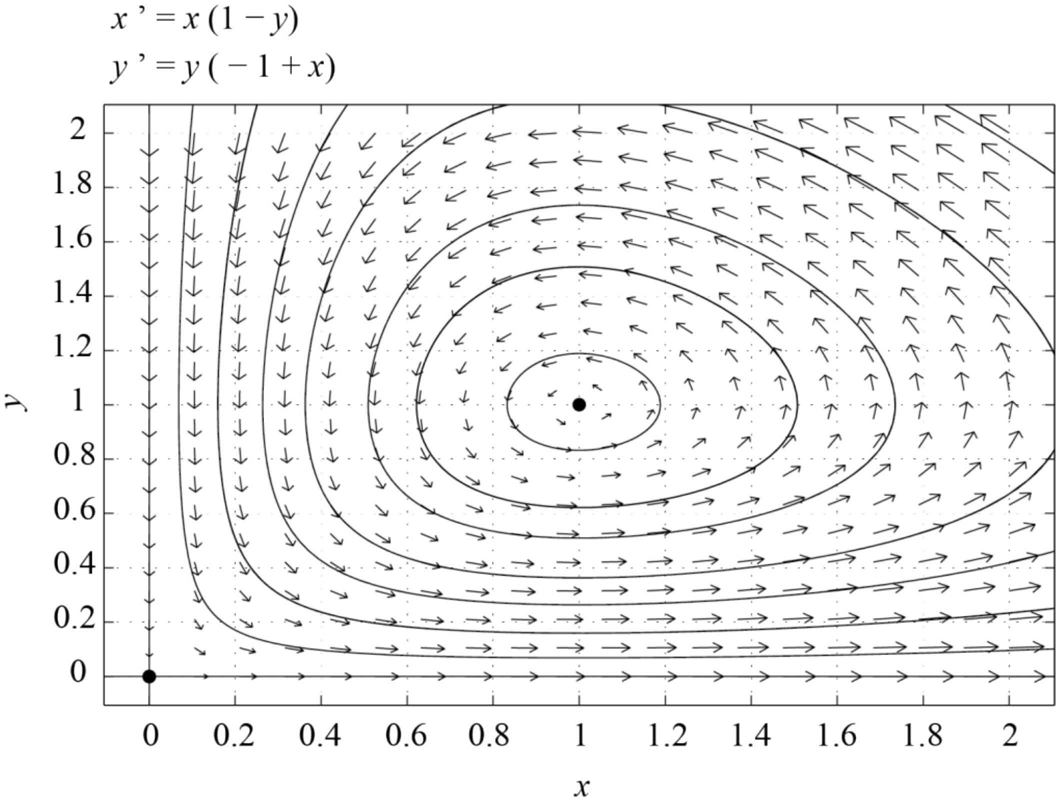


Figure 6.2. Phase plane of the Lotka-Volterra model (6.2) with *a* = 1 and *b* = 0, plotted with the software PPLANE.

It is possible to observe oscillations, provided we set $b = 0$. This is indeed a necessary condition if we want P_2 to be a center, since $\text{Tr}(J(P_2)) = -ab$ and $\det(J(P_2)) = a(1 - b)$. Indeed, setting $b = 0$ makes the trace of $J(P_2)$ vanish, but keeps its determinant positive. In this case, P_1 disappears (it is in fact sent to infinity as b goes to zero), and we are left with two fixed points, $P_0 = (0, 0)$ and $P_2 = (1, 1)$. The Jacobian $J(1, 1) = \begin{pmatrix} 0 & -a \\ 1 & 0 \end{pmatrix}$ has eigenvalues $\pm i\sqrt{a}$, so P_2 is now a linear center, as expected. Figure 6.2 shows the corresponding phase plane, with closed trajectories around P_2. We can prove the existence of close trajectories as follows. Any trajectory in the phase plane is such that

$$\frac{df}{ds} = \frac{af(1 - bf - s)}{s(-1 + f)}.$$

When $b = 0$, an implicit solution to this equation is

$$(f \exp(-f)) \, (s \exp(-s))^a = \zeta,$$

where ζ is a non-negative constant. It is easy to see that there are values of ζ for which this equation defines a closed curve around P_2. Indeed, the function $x \exp(-x)$ is zero at the origin, has a maximum equal to $1/e$ at $x = 1$, and goes to zero as $x \to \infty$. As a consequence, for given values of s and a, one can find a range of values of ζ such that the equation $f \exp(-f) = \dfrac{\zeta}{(s \exp(-s))^a}$ has two solutions, one on each side of $f = 1$. A similar reasoning applies to cross-sections parallel to the s-axis. These closed trajectories are drawn counterclockwise in the first quadrant of the (f, s) plane, since

$$\frac{df}{d\tau} = af(1-s) > 0 \Leftrightarrow s < 1 \text{ and } \frac{ds}{d\tau} = s(-1+f) > 0 \Leftrightarrow f > 1.$$

The Lotka-Volterra model therefore predicts periodic oscillations of the predator and prey populations.

The ideas developed in this section can be generalized to more complex systems. For instance, it is shown in a 2000 article by G.F. Fussmann *et al*. entitled *Crossing the Hopf Bifurcation in a Live Predator-Prey System*, that the predictions of a predator-prey model involving environmental factors as well as reproductive and non-reproductive fractions of a population compare very well with experimental results.

Two competing species

We now turn to the problem of two species competing for the same resources. We will assume that in the absence of the other species, each species grows according to a logistic law. The competition for food makes the growth rate of each species limited by the presence of the other. As a first approximation, this effect is linear. If we denote by X and Y the average density of each species, we thus have

$$\begin{cases} \dfrac{dX}{dt} = X(\alpha - \beta X - \gamma Y), \\ \dfrac{dY}{dt} = Y(\delta - \kappa Y - \zeta X), \end{cases} \qquad (6.3)$$

where α, β, γ, δ, κ and ζ are positive parameters. As for the Lotka-Volterra model, we first rescale these equations in order to reduce the number of independent parameters. We then perform a phase plane analysis to describe the dynamics of the competing species.

Scalings

Characteristic values of X and Y are $X_0 = \alpha/\beta$ and $Y_0 = \delta/\kappa$. A characteristic time is for instance $t_0 = 1/\alpha$. So we can define dimensionless quantities as

$$x = \frac{X}{X_0}, \qquad y = \frac{Y}{Y_0}, \qquad \tau = \frac{t}{t_0},$$

substitute these expressions into Equations (6.3), and look for a simplified system for the variables x and y. We obtain

$$\begin{cases} \dfrac{dx}{d\tau} = x\,(1 - x - a\,y), \\ \dfrac{dy}{d\tau} = c\,y\,(1 - y - b\,x), \end{cases} \tag{6.4}$$

where $a = \dfrac{\gamma\delta}{\alpha\kappa}$, $b = \dfrac{\zeta\alpha}{\beta\delta}$, and $c = \dfrac{\delta}{\alpha}$. We thus have a three-parameter model. This was to be expected since we initially had a six-parameter nonlinear system, and had the possibility of rescaling three variables, t, X and Y. We leave the question of biological well-posedness as an exercise and directly move to an analysis of the fixed points of Equations (6.4).

Phase plane analysis

As usual, fixed points are obtained by solving $x(1 - x - ay) = 0$ and $y(1 - y - bx) = 0$. This system of equations has four solutions, given by

$$P_0 = (0, 0), \quad P_1 = (1, 0), \quad P_2 = (1, 0), \quad P_3 = \left(\frac{1-a}{1-ab}, \frac{1-b}{1-ab} \right).$$

The last fixed point P_3 is in the first quadrant only if $1 - a$, $1 - b$ and $1 - ab$ are all of the same sign, and $ab \neq 1$. In what follows, we assume that these conditions are satisfied, and denote by $\epsilon = \pm 1$, the sign of any of these three quantities, i.e.

$$\epsilon = \operatorname{sign}(1 - a) = \operatorname{sign}(1 - b) = \operatorname{sign}(1 - ab).$$

The Jacobian of (6.4) is

$$J(x, y) = \begin{pmatrix} 1 - 2x - ay & -ax \\ -b\,c\,y & c(1 - 2y - bx) \end{pmatrix}.$$

We can now calculate the eigenvalues of $J(x,y)$ evaluated at each of the fixed points.

Analysis near the origin, P_0

For P_0, $J(0,0)$ is diagonal and its eigenvalues are 1 and c. Since they are both positive, the origin is an unstable node.

Analysis near P_1

For P_1, $J(1,0)$ is upper-triangular, and its eigenvalues are -1 and $c(1-b)$. If $\epsilon = 1$, then P_1 is a saddle. If $\epsilon = -1$, it is a stable node.

Analysis near P_2

For P_2, $J(0,1)$ is lower-triangular, and its eigenvalues are $1-a$ and $-c$. Thus P_2 is a saddle if $\epsilon = 1$, and a stable node if $\epsilon = -1$.

Analysis near P_3

For P_3, the Jacobian is

$$J\left(\frac{1-a}{1-ab}, \frac{1-b}{1-ab}\right) = \frac{1}{1-ab}\begin{pmatrix} a-1 & a\,(a-1) \\ b\,c\,(b-1) & c\,(b-1) \end{pmatrix}.$$

Since its trace T and determinant D are given by

$$T = \frac{1}{1-ab}\left(a-1+c(b-1)\right), \quad D = \frac{(a-1)c(b-1)}{1-ab},$$

we see that $\operatorname{sign}(T) = -1$ and $\operatorname{sign}(D) = \epsilon$. If $\epsilon = -1$, P_3 is a saddle. If $\epsilon = 1$, P_3 is a stable node. Indeed, the discriminant of the characteristic polynomial is

$$T^2 - 4D = \frac{1}{(1-ab)^2}\left[(a-1) - c(b-1)\right]^2,$$

and is therefore positive. Thus, both eigenvalues are real.

Figures 6.3 and 6.4 show phase portraits of system (6.4) for $\epsilon = 1$ and $\epsilon = -1$ respectively.

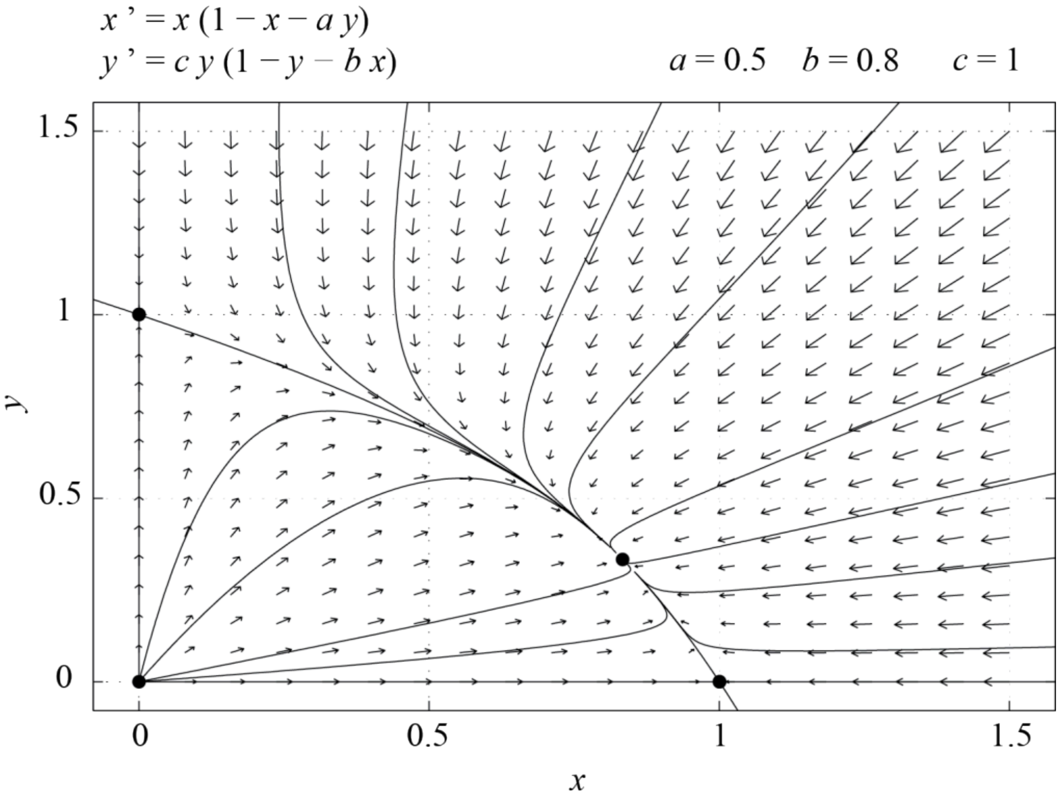


Figure 6.3. Phase plane of system (6.4), with a = 0.5, b = 0.8, and c = 1 (ε = 1), plotted with the software PPLANE.

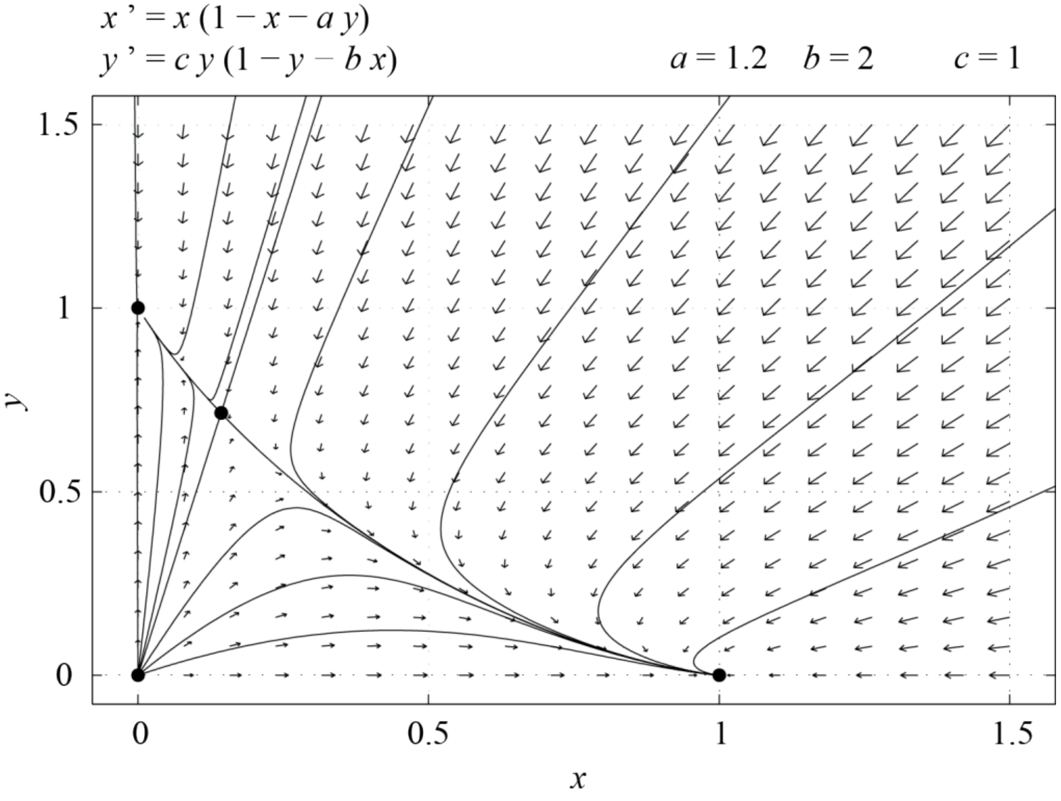


Figure 6.4. Phase plane of system (6.4), with a = 1.2, b = 2, and c = 1 (ε = -1), plotted with the software PPLANE.

From an ecological point of view, the two species coexist if P_3 is stable, i.e. if $\epsilon = 1$. If not, then one of the species is driven to extinction by the other, since the only fixed points which are stable when $\epsilon = -1$ are P_1

and P_2, for which one of the populations is zero. From this simple example, we see how relationships between model parameters may be inferred from biological facts. For instance, the above analysis indicates that if the two species coexist, then $a \leq 1$ and $b \leq 1$, with $ab \neq 1$.

Summary

In this chapter, we introduced two classical population dynamics models, the predator-prey system and a system describing two species competing for the same resources. The main tool we used for the analysis of these two-dimensional models was phase plane analysis. We only discussed elementary continuous models, but discrete analogues as well as more complex models may of course be considered.

From a biological point of view, it is important to remember that predator-prey systems may exhibit temporal oscillations. The latter typically arise from a Hopf bifurcation, as discussed for instance in the paper by Fussmann *et al.*[3] (see exercises), and are not related to the presence of an infinite number of closed orbits, as was the case in the Lotka-Volterra system with $a = 1$ and $b = 0$.

In the case of competing species, the simple model presented in this section shows that the two species can only coexist if a and b in system (6.4) are both less than unity. Otherwise, one of the species will drive the other to extinction.

Food for Thought

Problem 1

Write a time-continuous model which describes the following situation. Species x survives by eat-

[3] G.F. Fussmann, S.P. Ellner, K.W. Shertzer, N.G. Hairston Jr., *Crossing the Hopf Bifurcation in a Live Predator-Prey System*, Science **290**, 1358-1360 (2000).

ing nutrients n and has a per-capita death rate equal to d. The nutrients n are supplied to the system at a constant rate.

Problem 2

Is model (6.4) biologically well-posed? Why or why not?

Problem 3

Consider model (6.4) with $a < 1$ and $b > 1$.

1. Use phase plane analysis to describe the long-term dynamics of this system. Check your results with the Phase Plane App or equivalent software.
2. What is the biological significance of what you found out in part (1)?

Problem 4

Consider the following model

$$\begin{cases} \dfrac{dx}{dt} = -x + 4 - xy, \\ \dfrac{dy}{dt} = -xy + y + y^2, \end{cases}$$

where x and y are non-negative.

1. Find the fixed points of this system.
2. Linearize the system about its fixed points and find the stability of each fixed point.
3. Use the above information to sketch the phase plane of this system.
4. Check your answer with the Phase Plane App or equivalent software.
5. Could this model describe a biological system? Why or why not?

Problem 5

Consider the system

$$\begin{cases} \dfrac{dX}{dt} = -\alpha X + \beta, \\ \dfrac{dY}{dt} = \gamma XY - \delta Y - \zeta Y^2, \end{cases} \qquad \alpha, \beta, \gamma, \delta, \zeta > 0.$$

1. What is the dimension of each parameter?
2. Write this system in dimensionless form. Explain how you choose to scale the variables.

Problem 6

Consider the model described in the paper by G.F. Fussmann *et al.* entitled *Crossing the Hopf Bifurcation in a Live Predator-Prey System*.

1. Which species is the predator and which is the prey?
2. What is the role of N in the model?
3. How would you modify the model if you did not want to distinguish between reproducing and non-reproducing *Brachionus*?

7.

EPIDEMIOLOGY

Learning Objectives

At the end of this chapter, you will be able to do the following.

- Construct compartmental models describing infections and disease outbreaks.
- Formulate dimensionless versions of two-dimensional models and appraise their dynamics by means of phase plane analysis.
- Translate the results of model analysis into biological terms and discuss their significance in that context.

Viral infections

We first model the dynamics of a viral infection, such as hepatitis B or C, and are interested in describing how the corresponding virus can spread and multiply in a person's body. Denote by X the average number of uninfected cells, which virions will try to infect; let Y be the average number of infected cells, and V be the average viral load (or the number of free virions in the body). Consider that uninfected cells are produced at a constant rate λ by the body, die at rate δX, and become infected at rate $f(X,V)X$, where f is some function of X and V. As a consequence, infected cells Y are created at rate $f(X,V)X$, and we assume they die at rate aY. Finally, free virions are produced at a rate proportional to the number of infected cells kY, and are removed or destroyed at rate κV. If we consider that, as a first approximation, f is a linear function of V, i.e. $f(X, V) = bV$, we have the following model.[1]

$$\begin{cases} \dfrac{dX}{dt} = \lambda - \delta X - bVX, \\ \dfrac{dY}{dt} = bVX - aY, \\ \dfrac{dV}{dt} = kY - \kappa V, \end{cases} \tag{7.1}$$

This model has six parameters and four variables. We can rescale time and V, but it is not a good idea to rescale X and Y independently since they both count cells and the term in bVX transfers cells from the X compartment to the Y compartment. We can therefore reduce Equations (7.1) to a model with three parameters. More precisely, let

$$\tau = \delta t, \qquad x = \frac{kb}{\delta^2}X, \qquad y = \frac{kb}{\delta^2}Y, \qquad v = \frac{b}{\delta}V.$$

Then, the scaled version of Equations (7.1) is

$$\begin{cases} \dfrac{dx}{d\tau} = \zeta - x - v\,x, \\ \dfrac{dy}{d\tau} = v\,x - \eta\,y, \\ \dfrac{dv}{d\tau} = y - \mu\,v, \end{cases} \tag{7.2}$$

where $\zeta = \dfrac{\lambda k b}{\delta^3}$, $\eta = \dfrac{a}{\delta}$, and $\mu = \dfrac{\kappa}{\delta}$ are dimensionless parameters.

We scaled time according to the death rate of normal cells. Alternatively, we could have scaled time according to the rate at which normal cells are produced by the body, i.e. we could have defined $\tau = \lambda t / X_0$, with $X_0 = \delta^2/(kb)$. We could also have used a combination of these two time scales. In general, there is more than one possible way of defining dimensionless variables. The most convenient choice is often that which gives dimensionless parameters of order one, if at all possible.

We refer the reader to the 1998 article by A.U. Neumann *et al.*, entitled *Hepatitis C viral dynamics in vivo and*

[1] M.A. Nowak, S. Bonhoeffer, A.M. Hill, R. Boehme, H.C. Thomas, and H. McDade, *Viral dynamics in hepatitis B virus infection*, Proc. Natl. Acad. Sci. USA 93, 4398-4402 (1996).

the antiviral efficacy of interferon-alpha therapy, to see how estimating parameters in model (7.1) may be used to understand the role of interferons in hepatitis C therapy.

Modeling infectious diseases

There is ample literature on the modeling of infectious diseases (for a review, see for instance the 2000 article by H. Hethcote, entitled *The Mathematics of Infectious Diseases*). The most general basic model is called the MSEIRS model. Each letter in the acronym refers to a particular class or compartment of the population, and the ordering of the letters is such that individuals belonging to one class move to the class on the right under the effect of the disease. The various classes are defined as follows.

- Class M corresponds to *infants with passive immunity*. Typically, these are newborns whose mothers were infected, and who received antibodies through the placenta before birth.
- Class S is the class of *susceptible* individuals, who may become infected by the disease. Infants in class M have temporary immunity and eventually move to class S.
- Class E corresponds to *exposed* individuals, who have been in contact with an infected person.
- Class I is the class of *infectious* individuals, who can transmit the disease.
- Class R corresponds to individuals who have *recovered* (or died) from the disease, or who have been *removed* from the group of people affected by the disease.

Depending on the type of infectious disease, individuals in class R may have acquired permanent immunity. In this case, the appropriate model is of the MSEIR type. However, if individuals in class R eventually become susceptible again, then an MSEIRS model should be used. Most realistic models couple the classes described above with age groups. Some models consider the age a of a person as an independent variable. Other classes, such as symptomatic and asymptomatic groups of individuals, may be considered, and the appropriate number of compartments, as well as their nature, is selected by the modeler. Applications of compartmental models include disease forecasting, as well as predicting the effectiveness of vaccination or of a disease eradication campaign. Below, we only discuss two simple models, namely the classic and endemic SIR models.

The SIR model

The classic SIR model does not include classes M and E, and reads

$$\begin{cases} \dfrac{dS}{dt} = -\alpha S\dfrac{I}{N}, \\ \dfrac{dI}{dt} = \alpha S\dfrac{I}{N} - \beta I, \\ \dfrac{dR}{dt} = \beta I, \end{cases} \qquad (7.3)$$

with initial conditions $S(0) = S_0 \geq 0$, $I(0) = I_0 > 0$, and $R(0) \geq 0$. Here, S, I and R are the expected numbers of individuals in each compartment, and $N = S + I + R$ is the total population. Note that births and deaths are not included in the model. This typically works for diseases which evolve over a short period of time, so that changes in the total population are negligible.

In this model, I/N represents the fraction of infectious individuals. The product of this quantity with the *contact rate* $\alpha > 0$ measures the average number of positive (i.e. giving rise to transmission of the disease) contacts per susceptible individual per unit of time. Since there are on average S susceptible individuals, the rate of change of S is $-\alpha SI/N$. The number of infected individuals increases by contact with susceptibles, and decreases due to recovery, at rate $-\beta I$, with $\beta \geq 0$. By adding up the three equations, one easily checks that $dN/dt = 0$, as expected since we neglected births and deaths. System (7.3) can thus be reduced to a two-dimensional system of ordinary differential equations, by omitting the last equation for R. The remaining two equations may be written in dimensionless form by letting

$$s = \frac{S}{N}, i = \frac{I}{N}, \text{ and } \tau = \alpha t.$$

Then, the first two equations of (7.3) become

$$\begin{cases} \dfrac{ds}{d\tau} = -si, \\ \dfrac{di}{d\tau} = si - \delta i, \end{cases} \qquad (7.4)$$

where $\delta = \beta/\alpha > 0$. The quantity $\sigma = 1/\delta$ is the contact rate α multiplied by the characteristic time $1/\beta$ during which a person remains infectious. It is called the *contact number* of the disease, and in this case is equal to the *basic reproduction number* R_0 of the infection described by the SIR model.

Since $(1 - s - i)N = R \geq 0$, Equations (7.4) only make biological sense if s and i remain positive and such that $s + i \leq 1$, provided initial conditions satisfy these requirements. In other words, trajectories of the dynamical system (7.4) that start in the triangle

$$\mathcal{T} = \{(s, i) | s \geq 0,\ i \geq 0,\ s + i \leq 1\}$$

should remain in $\mathcal{T}$. To check this, consider the dynamics on the boundary of $\mathcal{T}$. First assume that $s = 0$

and $i \leq 1$. Then $ds/d\tau = 0$, i.e. s remains equal to zero, and i decreases towards zero, but does not become negative. Similarly, if $i = 0$, then both s and i remain constant (what is the biological significance of this fact?). Finally, if $s + i = 1$, then $\dfrac{d}{d\tau}(s+i) = -\delta i \leq 0$, so that $s + i$ will not increase past the value 1. System (7.4) has an infinite number of fixed points in $\mathcal{T}$, which are such that $i = 0$ and s is arbitrary.

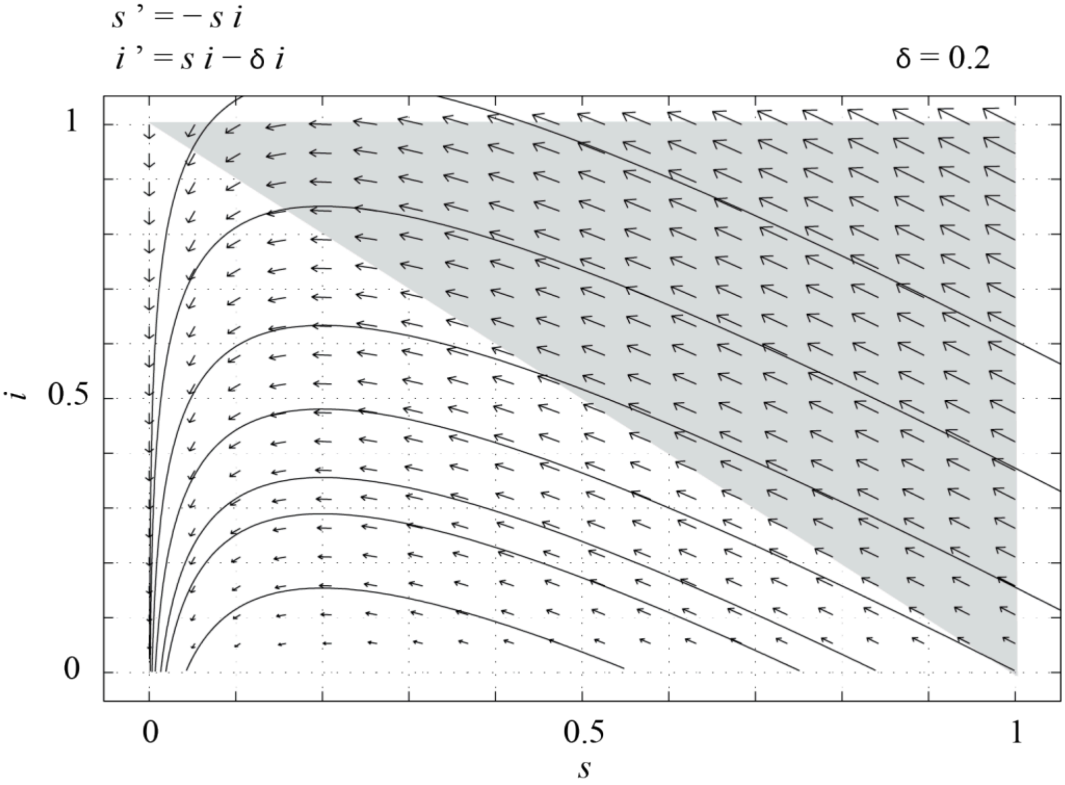


Figure 7.1. Phase plane of system (7.4), with δ = 0.2, plotted with the software PPLANE. Only the dynamics inside *T* (not shaded) is relevant.

Figure 7.1 shows the phase portrait of (7.4), obtained with PPLANE, with $\delta = 0.2$. In this case, we see that all trajectories in $\mathcal{T}$ converge towards one of the fixed points, i.e. $\lim_{t\to\infty} i = 0$. This means that the epidemic eventually dies out, and we are only left with susceptible individuals and/or those who have recovered from the disease.

The classic endemic model

For an endemic disease, births and deaths need to be taken into account, and the SIR model becomes

$$\begin{cases} \dfrac{dS}{dt} = \nu N - \mu S - \alpha S \dfrac{I}{N}, \\ \dfrac{dI}{dt} = -\mu I + \alpha S \dfrac{I}{N} - \beta I, \\ \dfrac{dR}{dt} = -\mu R + \beta I, \end{cases} \qquad (7.5)$$

with initial conditions $S(0) = S_0 \geq 0$, $I(0) = I_0 > 0$, and $R(0) \geq 0$. Here the new parameters are the per capita death rate μ and per capita birth rate ν of the population. By choosing $\mu = \nu$, the total population $N = S + I + R$ is constant. In this case, using the same dimensionless variables as for the classic SIR model, we are left with a two-dimensional dynamical system, which reads, in dimensionless form,

$$\begin{cases} \dfrac{ds}{d\tau} = \eta - \eta s - si, \\ \dfrac{di}{d\tau} = -(\eta + \delta)i + si, \end{cases} \qquad (7.6)$$

where $\eta = \nu/\alpha = \mu/\alpha$.

As before, it is easy to check that trajectories starting in $\mathcal{T}$ remain in $\mathcal{T}$ (see exercises). The fixed points of (7.6) in the (s, i) plane are

$$P_1 = (1, 0) \qquad \text{and} \qquad P_2 = \left(\eta + \delta, \frac{\eta(1 - \eta - \delta)}{\eta + \delta} \right).$$

The Jacobian of (7.6) is

$$J(s, i) = \begin{pmatrix} -\eta - i & -s \\ i & s - (\eta + \delta) \end{pmatrix}$$

and

$$J(P_1) = \begin{pmatrix} -\eta & -1 \\ 0 & 1 - (\eta + \delta) \end{pmatrix}.$$

Whether P_2 is in $\mathcal{T}$ depends on the parameters δ and η. More precisely, $P_2 \in \mathcal{T} \Leftrightarrow 0 < \eta + \delta \leq 1$. Therefore, if $\eta + \delta > 1$, P_1 is the only fixed point in $\mathcal{T}$ and since the eigenvalues of $J(P_1)$ are $-\eta$ and $1 - (\eta + \delta)$, P_1 is a stable node. All of the trajectories starting in $\mathcal{T}$ must converge to this fixed point, which means that in the long run there are only susceptible individuals in the population. This is because those who have recovered from the disease eventually die and are replaced by newborns, who are susceptible. This is illustrated in Figure 7.2, which shows the phase portrait of (7.6), obtained with PPLANE, with $\delta = 0.2$ and $\eta = 1$.

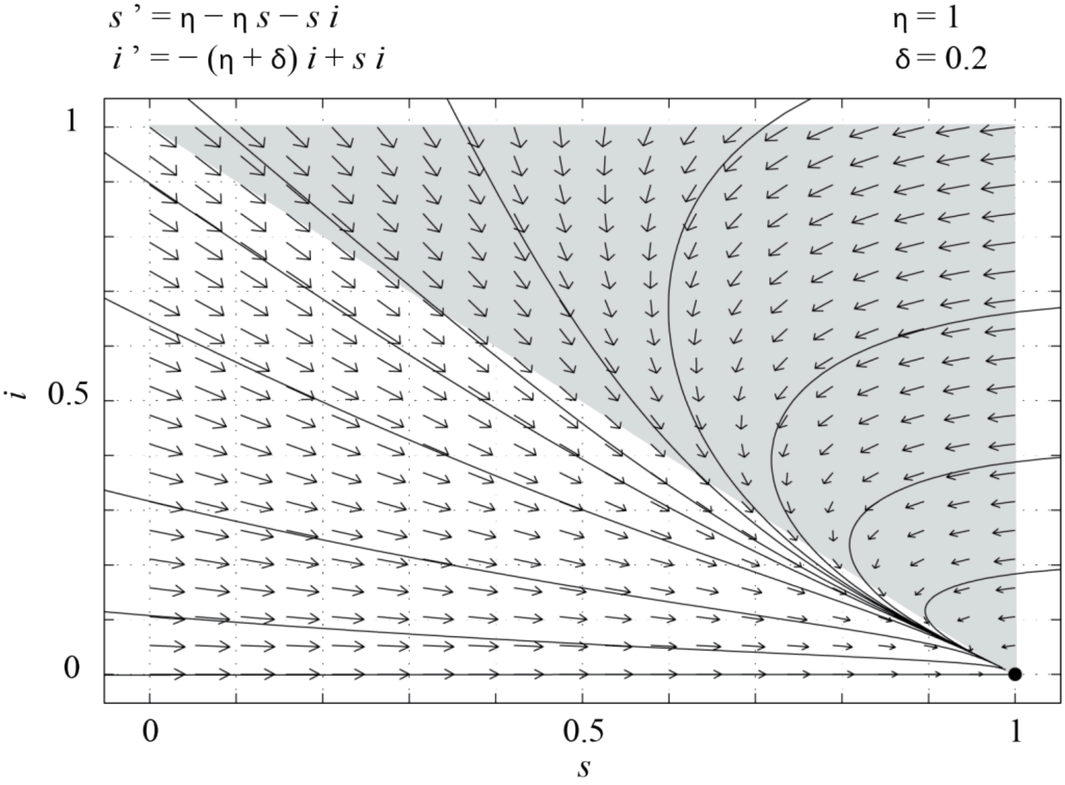


Figure 7.2. Phase plane of system (7.6), with **δ** = 0.2 and **η** = 1, plotted with the software PPLANE. Only the dynamics inside *T* (not shaded) is relevant.

If on the other hand $\eta + \delta < 1$, then P_1 is a saddle, and

$$J(P_2) = \begin{pmatrix} -\dfrac{\eta}{\eta+\delta} & -(\eta+\delta) \\ \dfrac{\eta(1-\eta-\delta)}{\eta+\delta} & 0 \end{pmatrix}.$$

The determinant of $J(P_2)$ is equal to $\eta(1-\eta-\delta)$ and is positive. The trace of $J(P_2)$ is negative, so P_2 is either a stable spiral or a stable node. Trajectories starting in $\mathcal{T}$ converge to P_2, which is called the *endemic equilibrium*. In this case, the disease is always present in the population and there is always a non-zero number of infected individuals. Figure 7.3 shows the phase portrait of (7.6) with $\eta = 0.2$ and $\delta = 0.1$.

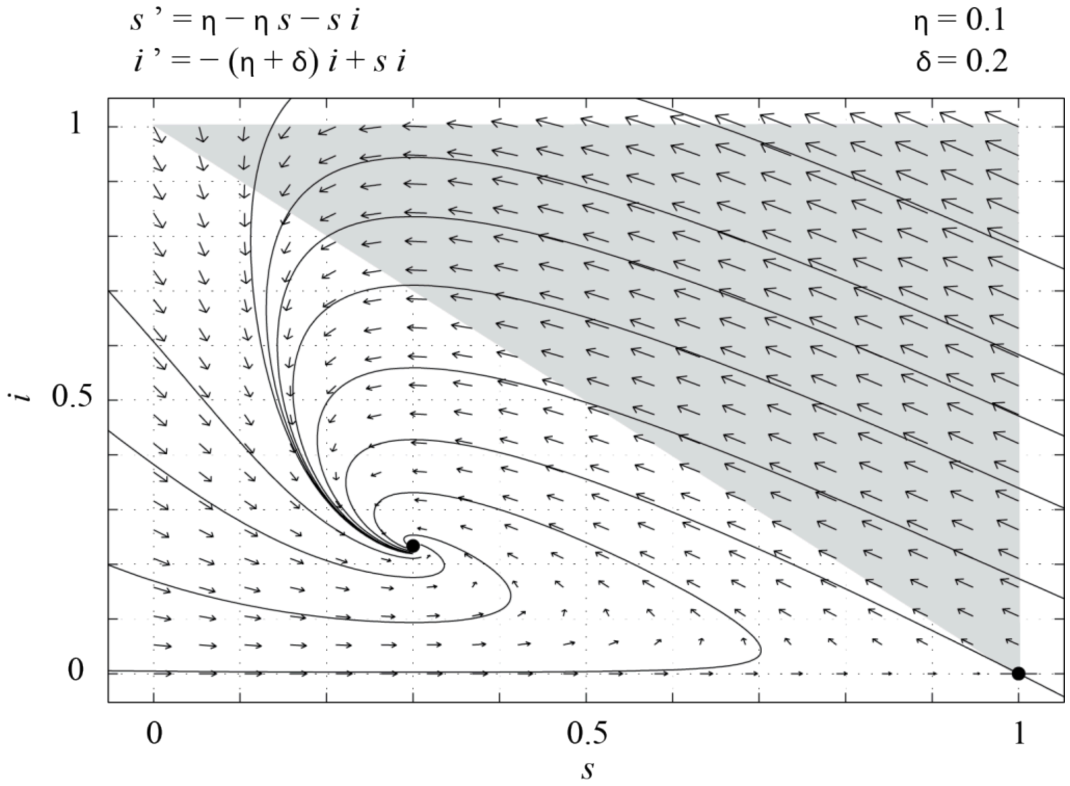


Figure 7.3. Phase plane of system (7.6), with **δ** = 0.2 and **η** = 0.1, plotted with the software PPLANE. Only the dynamics inside *T* (not shaded) is relevant.

Summary

This chapter illustrates how the modeling concepts discussed in the previous chapters may be applied to the description, in terms of ordinary differential equations, of the dynamics of infectious diseases and the spread of epidemics. Moreover, it provides the reader with a basic introduction to the terminology and applications of compartmental models.

It should be clear by now that models of arbitrary complexity may be built from the simple tools discussed in this text. The modeling process is always the same, no matter how involved the model is. The methods of analysis, in terms of maps or differential equations, are also similar, but become more complicated as the dimension of the model is increased. In particular, three-dimensional continuously differentiable dynamical systems may exhibit chaos, the understanding of which requires more advanced techniques than those discussed here. The section on further reading includes texts on dynamical systems and chaos that the reader may want to consult.

Food for Thought

Problem 1

Write model (7.1) in dimensionless form by defining $\tau = \lambda kbt/\delta^2$ and appropriate variables x, y and v. Explain what you are doing and check that τ is dimensionless.

Problem 2

Consider model (7.2).

1. Find the fixed points of this system.
2. What are the conditions on the parameters for these fixed points to be in the first octant? What is the biological significance of these conditions?
3. Discuss the linear stability of the fixed point such that $y = v = 0$.

Problem 3

Consider model (7.6). Show that trajectories starting in the triangle $\mathcal{T}$ remain in $\mathcal{T}$.

Problem 4

Consider model (7.6) with $\eta + \delta < 1$. Are there values of η and δ for which P_2 is a stable node and not a stable spiral?

Problem 5

The MSEIR model is suitable for diseases such as rubella or measles, since once individuals are infected, they become immune to the disease. Write a system of equations for the MSEIR model, in a population with per capita birth rate ν and per capita death rate μ, with $\nu \neq \mu$.

PART IV
CHEMICAL REACTIONS AND SPATIAL EFFECTS

Chapter 8 focuses on chemical reactions. We first introduce the *law of mass action* to obtain rate equations, and consider two classical models for oscillatory chemical reactions, which only involve ordinary differential equations. These models may be simplified into two-dimensional dynamical systems and are amenable to the kind of analysis done in the previous chapters.

So far, we did not discuss or even consider the fact that most models describe quantities which vary not only in time, but also in space. For instance, if N measures the density of a population, it is normal to assume that N depends on the landscape: for humans, N is larger in cities than in mountain or desert areas; for animals, N varies according to the presence of forests, rivers, prey, etc.

In order to include spatial effects in the models we have discussed, we proceed as follows. First, we assume that the nature of the model itself does not change. For instance, if species u eats species v, then at any point with coordinates (x, y) in the plane, $u(x, y)$ grows proportionally to the local concentration of $v(x, y)$. One could imagine more complicated situations where the growth rate of u depends on say the spatially averaged density of v, but such cases are beyond the scope of these notes. In other words, we only consider *local* models. Second, we need to describe how each species behaves when it is not uniformly distributed over a region of space. Our intuition tells us that motion should take place away from regions of high concentration, in order to reach a uniform distribution. This is the phenomenon of *diffusion*, which is discussed in Chapter 9. In the presence of an external flow, other transport terms have to be included, but we will not consider such situations.

Adding diffusion to nonlinear differential equations typically leads to *reaction-diffusion* equations. These models are systems of partial differential equations, and there is a vast literature on this and related topics. Although a complete discussion is beyond the scope of these notes, we nevertheless give a flavor of the kind of modeling done with such systems. Chapter 9 briefly describes an example of surface chemical reaction, in which diffusion is coupled to the corresponding chemical rate equations. The resulting reaction-diffusion system is able to sustain chemical waves and we encourage the reader to study the articles referenced in the text for additional information. Chapter 10 discusses the general phenomenon of *pattern formation* in systems driven far from equilibrium. There, we explore the dynamics of a generic pattern-forming model, namely the *Swift-Hohenberg* equation, and also consider the specific example of a two-dimensional model leading to vegetation patterns.

8.

CHEMICAL REACTIONS

Learning Objectives

At the end of this chapter, you will be able to do the following.

- Assemble systems of coupled nonlinear ordinary differential equations describing how concentrations of chemical reactants evolve in time.
- Predict the dynamics of chemical reactions by applying the methods of analysis (dimensional analysis, scalings, fixed points and their linear stability) discussed in previous chapters.
- Explain why oscillatory chemical reactions are possible.
- Explain how chemical waves may be observed in spatially extended systems.

The law of mass action

Consider a chemical reaction

$$A + B \quad \overset{k}{\rightarrow} \quad C,$$

where A, B and C represent chemicals, A and B are the reactants, C is the product, and $k > 0$ is the *rate constant* of the reaction. The *law of mass action* describes how the concentrations of A, B and C change as a consequence of this reaction. The idea is that the reaction will take place if chemicals A and B collide, and if their energy is higher than the energy of activation of the reaction. The number of successful (i.e. leading to a reaction) collisions between A and B is proportional to the product of the concentrations of A and B. The constant of proportionality is the constant k. We thus write the following differential equations for the expected values of $[A]$, $[B]$, $[C]$, (recall that we are in a situation where large numbers of molecules of reactants and products are present)

$$\frac{d[A]}{dt} = \frac{d[B]}{dt} = -k[A]\,[B] = -\frac{d[C]}{dt},$$

where $[I]$ denotes the concentration of chemical I. If $A = B$, so that the chemical reaction is

$$\begin{array}{ccc} & k & \\ 2A & \rightarrow & C, \end{array}$$

we need two molecules of A to create one product molecule C, leading to

$$\frac{1}{2}\frac{d[A]}{dt} = -k[A]^2 = -\frac{d[C]}{dt}.$$

If a reaction is reversible, for instance if

$$\begin{array}{ccc} & k_1 & \\ A + B & \rightleftharpoons & C, \\ & k_2 & \end{array}$$

then A is consumed at relative rate k_1 and produced at relative rate k_2, so that

$$\frac{d[A]}{dt} = -k_1[A]\,[B] + k_2[C],$$

and similarly for $[B]$ and $[C]$.

Finally, if a reaction process involves more than one chemical reaction, the rates of change of a chemical due to each of the reactions are added up in order to obtain the global rate of change of that chemical. For instance, assume that we have a system described by

$$\begin{array}{ccc} & k_1 & \\ A + B & \rightleftharpoons & C, \\ & k_2 & \\ & & \\ & k_3 & \\ C + A & \rightarrow & D. \end{array}$$

Then the concentration of A evolves according to

$$\frac{d[A]}{dt} = -k_1[A][B] + k_2[C] - k_3[C][A],$$

and the rate equations for B, C and D are

$$\frac{d[B]}{dt} = -k_1[A][B] + k_2[C],$$

$$\frac{d[C]}{dt} = k_1[A][B] - k_2[C] - k_3[C][A],$$

$$\frac{d[D]}{dt} = k_3[C][A].$$

Note that the net reaction associated with this process is $2A + B \rightarrow D$. The rate equations must be written for each of the steps involved in the reaction, and not on the basis of the net reaction.

Chemical reactions can thus be described in terms of coupled nonlinear ordinary differential equations, and the theory of dynamical systems therefore applies to their analysis.

The Brusselator and Oregonator models

Because rate equations describing chemical oscillations are nonlinear, it is possible to observe reactions which are oscillatory in time, in the same way as the Lotka-Volterra equations may describe oscillations in a predator-prey system. The Belousov-Zhabotinsky reaction is the classical example of an oscillatory chemical reaction. When Russian biochemist Boris P. Belousov[1] reported his findings in 1951, his results were initially received with disbelief. It is only after A.M. Zhabotinsky[2] reproduced and improved Belousov's experiments, that the existence of oscillatory reactions was finally accepted. There are two classical models for oscillatory reactions, called the "Brusselator" (proposed by a group in Brussels[3]) and the "Oregonator" (proposed by chemists at the University of Oregon[4]). We briefly discuss each of them below.

[1] B.P. Belousov, Sb. Ref. Radiats. Med., 1958, Megiz, Moscow, 145 (1950).

[2] A.M. Zhabotinsky, *Oscillatory Processes in Biological and Chemical systems*, Science Publ., Moscow, p. 149, 1967. A.N. Zaikin and A.M. Zhabotinsky, Concentration *Wave Propagation in Two-dimensional Liquid-phase Self-oscillating System*, Nature **225**, 535-537 (1970).

[3] P. Glansdor and I. Prigogine, *Thermodynamic Theory of Structure, Stability and Fluctuations*, Wiley Interscience, London, 1971. Page 233.

[4] R.J. Field and R.M. Noyes, *Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction*, J. Chem. Phys. **60**, 1877-1884 (1974).

The Brusselator

Consider the following hypothetical chemical process (Glansdorff & Prigogine, 1971).

$$
\begin{aligned}
A &\overset{k_1}{\rightarrow} X, \\
2X + Y &\overset{k_2}{\rightarrow} 3X, \\
B + X &\overset{k_3}{\rightarrow} Y + D, \\
X &\overset{k_4}{\rightarrow} E,
\end{aligned}
$$

where the rate constants k_i are all equal to 1. The corresponding rate equations for $[X]$ and $[Y]$ form the *Brusselator* model, which reads

$$
\begin{cases}
\dfrac{d[X]}{dt} = [A] + [X]^2\,[Y] - [B]\,[X] - [X], \\
\dfrac{d[Y]}{dt} = -[Y]\,[X]^2 + [B]\,[X],
\end{cases}
\qquad (8.1)
$$

where $[A]$ and $[B]$ are parameters. It has a unique fixed point P, given by $[X] = [A]$ and $[Y] = [B]/[A]$. From a dimensional analysis point of view, these expressions may look strange, but we lost track of the dimensions when we set all of the reaction constants (which had different dimensions) to unity (see exercises).

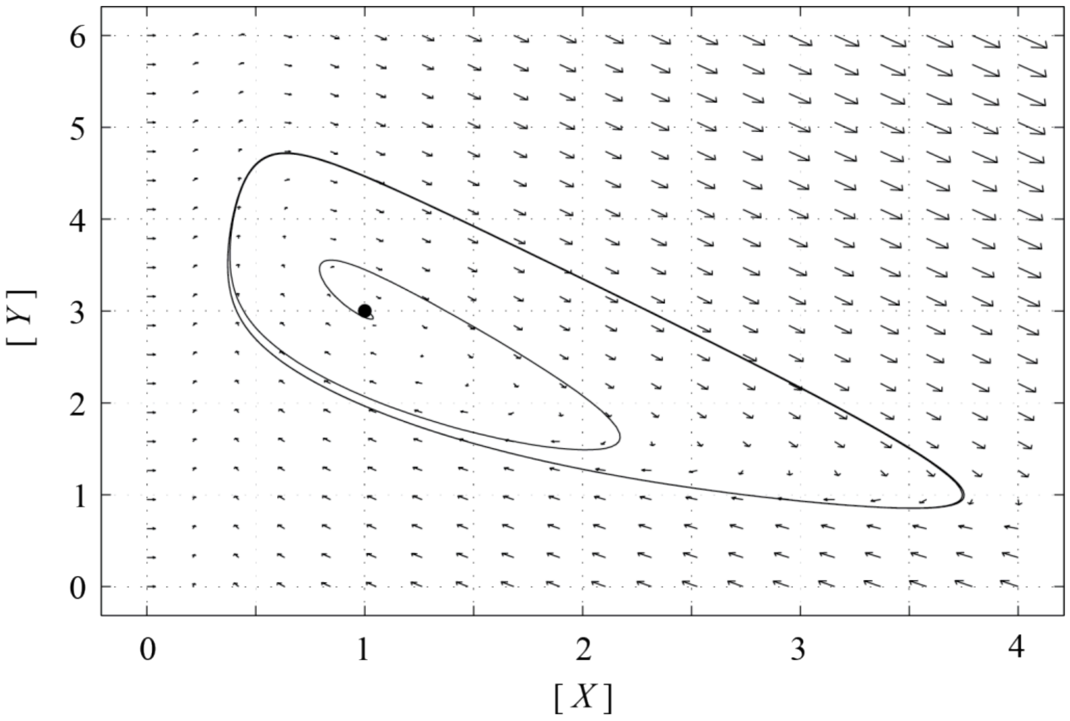


Figure 8.1. Phase plane of system (8.1), with [A]=1 and [B]=3, plotted with the software PPLANE.

The Jacobian of (8.1) about the fixed point P is

$$J(P) = \begin{pmatrix} [B]-1 & [A]^2 \\ -[B] & -[A]^2 \end{pmatrix},$$

and its determinant is equal to $[A]^2 > 0$. The stability of the fixed point therefore depends on the sign of the trace of $J(P)$, which is equal to $T = [B] - 1 - [A]^2$. It is easy to see that if $[X]$ or $[Y]$ are sufficiently large, then $d[Y]/d[X] \simeq -1$, and trajectories are almost straight lines with slope -1. However, as $[Y]$ gets close to 0, these trajectories cannot leave the first quadrant, since $d[Y]/dt > 0$ if $[Y] = 0$ and $[X] > 0$. Since at that time we also have $d[X]/dt < 0$ if $[X]$ is large enough, we expect the trajectories to be brought back towards regions where both $[X]$ and $[Y]$ are of order one.

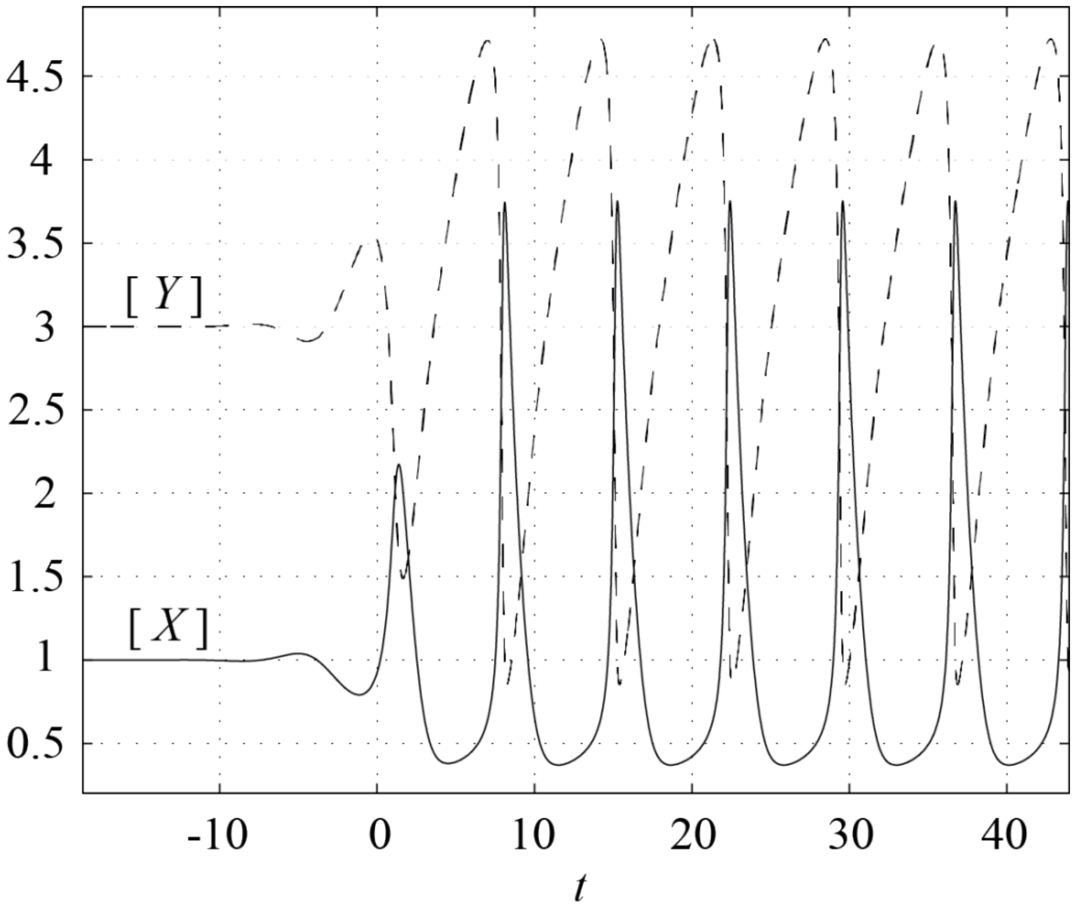


Figure 8.2. Time dependence of [X] and [Y] for the trajectory of Figure 8.1, plotted with the software PPLANE.

If P is stable, i.e. if $T < 0$, then all trajectories should end up at P. However, if P is an unstable node or an unstable spiral, trajectories cannot converge towards P, and there must be another attractor of the dynamics. Figure 8.1 shows the phase portrait of (8.1) with $[A] = 1$ and $[B] = 3$ (so that $T = 1 > 0$). The second attractor is a *limit cycle*, towards which all trajectories converge. Only one trajectory is plotted in this figure. The corresponding plots of $[X]$ and $[Y]$ as functions of time are shown in Figure 8.2. One can see that on the limit cycle, $[X]$ remains small most of the time, then quickly increases to a maximum value and comes back towards zero, in a periodic fashion. As $[X]$ increases, $[Y]$ drops abruptly, and then slowly grows back to its maximum value while $[X]$ remains small. Such *relaxation oscillations* are typical of many oscillatory chemical reactions.

The Oregonator

The *Oregonator* was proposed by R.J. Fields and R.M. Noyes in 1974. It corresponds to the following chemical reactions

$$A + Y \overset{k_1}{\rightarrow} X,$$

$$X + Y \xrightarrow{k_2} P,$$

$$B + X \xrightarrow{k_3} 2X + Z,$$

$$2X \xrightarrow{k_4} Q,$$

$$Z \xrightarrow{k_5} fY,$$

where $[A]$ and $[B]$ are constant, and f is a stoichiometric factor. The corresponding rate equations are

$$\begin{cases} \dfrac{d[X]}{dt} = k_1[A]\,[Y] - k_2[X]\,[Y] + k_3[B]\,[X] - 2k_4[X]^2, \\ \dfrac{d[Y]}{dt} = -k_1[A]\,[Y] - k_2[X]\,[Y] + k_5 f[Z], \\ \dfrac{d[Z]}{dt} = k_3[B]\,[X] - k_5[Z]. \end{cases}$$

Field and Noyes (1974) defined the following dimensionless quantities

$$x = \frac{[X]}{X_0}, \qquad y = \frac{[Y]}{Y_0}, \qquad z = \frac{[Z]}{Z_0}, \qquad \tau = \frac{t}{T_0},$$

where

$$X_0 = \frac{k_1}{k_2}[A], \quad Y_0 = \frac{k_3}{k_2}[B], \quad Z_0 = \frac{k_1 k_3}{k_2 k_5}[A][B], \quad T_0 = \frac{1}{\sqrt{k_1 k_3 [A][B]}}.$$

Then, the dimensionless form of the Oregonator becomes

$$\begin{cases} \dfrac{dx}{d\tau} = s(y - xy + x - qx^2), \\ \dfrac{dy}{d\tau} = \dfrac{1}{s}(-y - xy + fz), \\ \dfrac{dz}{d\tau} = \omega(x - z), \end{cases} \tag{8.2}$$

where

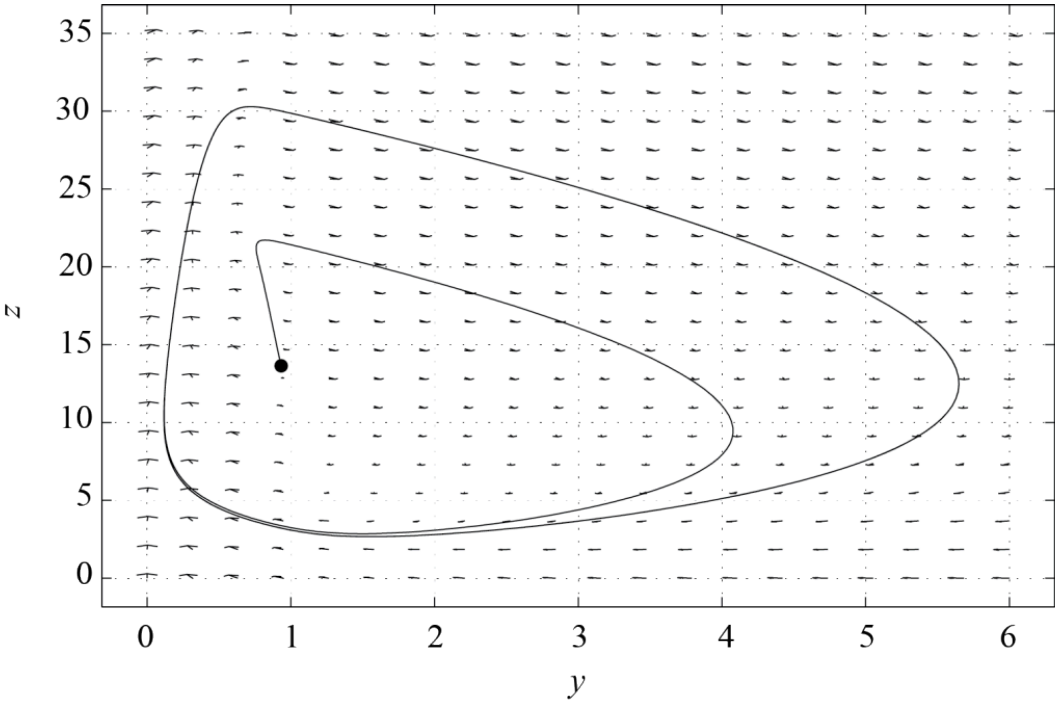
$$s = \sqrt{\frac{k_3[B]}{k_1[A]}}, \qquad q = \frac{2k_1k_4[A]}{k_2k_3[B]}, \qquad \omega = \frac{k_5}{\sqrt{k_1k_3[A][B]}}.$$

By assuming that x quickly reaches a steady-state value, it is possible to reduce Equations (8.2) to a two-dimensional dynamical system. Indeed, setting $dx/d\tau = 0$ gives

$$x = \frac{1}{2q}\left(1 - y + \sqrt{(1-y)^2 + 4qy}\right),$$

and by substituting this expression into (8.2), one obtains

$$\begin{cases} \dfrac{dy}{d\tau} = \dfrac{1}{s}\left(-y - \dfrac{y}{2q}\left(1 - y + \sqrt{(1-y)^2 + 4qy}\right) + fz\right), \\ \dfrac{dz}{d\tau} = \omega\left(\dfrac{1}{2q}\left(1 - y + \sqrt{(1-y)^2 + 4qy}\right) - z\right). \end{cases} \tag{8.3}$$



Phase plane of system (8.3), with f=1, s=1, q=0.01 and ω=2, plotted with the software PPLANE.

In what follows, we set $f = 1$. Field and Noyes (1974) estimated the values of s, q, ω as well as of all dimensionless variables for the Belousov-Zhabotinsky reaction. Details can be found in their article entitled *Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction*. They also concluded that the Oregonator model is capable of predicting oscillations for the concentrations of the chemicals involved in the reaction process. Realistic parameters make the problem very stiff (i.e. there are both very short and very long characteristic time scales, which is typical for excitable systems), but model (8.3) also exhibits oscillations in reasonably stiff cases. Figure 8.3 shows the phase plane of model (8.3) with parameters $s = 1, q = 0.01$, and $\omega = 2$. Trajectories starting from the fixed point with both y and z non-zero converge towards a limit cycle. As a consequence, the variables y and z oscillate in time, and so do $[X]$, $[Y]$ and $[Z]$ in the Oregonator model.

Chemical waves

When a chemical reaction takes place in a spatially extended system, diffusion should be taken into account. As a consequence, the rate equations discussed above are turned into partial differential equations. The reaction terms remain the same, but the concentration of each chemical now diffuses with a diffusion coefficient which depends on the size and weight of the molecule in question. If the reaction is oscillatory, wave fronts, corresponding to say large concentrations of some chemical, propagate in the system. Moreover, if the reaction is constrained to a two-dimensional surface and the wave is initiated at some point in space, then the wave fronts are circular. As an example, a 2001 article by C. Sachs *et al.*[5] describes how such wave fronts are observed in an experiment, and proposes a reaction-diffusion model which reproduces this behavior. The authors also discuss the limitation of reaction-diffusion models when macroscopic parameters are affected by the details of microscopic interactions.

What would happen if the wave front was broken, for instance if the wave went over a region where the reaction could not take place? If the wave front is anchored at one point, then it will curve and eventually form a spiral wave. Such waves are often observed in experiments where the Belousov-Zhabotinsky reaction is constrained to a two-dimensional surface, such as a thin film, a porous glass disk, or even a sheet of filter paper. The article by S.C. Müller *et al.*[6] discusses experiments revealing the structure of the core of such spiral waves.

Summary

We started with the law of mass action, which allowed us to describe the dynamics of the (average) concentrations of reactants and products involved in chemical reactions. In particular, we considered two hypothetical sets of chemical reactions called the Brusselator and the Oregonator. By applying the methods of analysis discussed in the previous chapters, we concluded that it was possible for these chemical systems to exhibit oscillatory dynamics. This provides a proof of concept for oscillatory reactions such as the Belousov-Zhabotinsky

[5] C. Sachs, M. Hildebrand, S. Völkening, J. Wintterlin, G. Ertl, *Spatiotemporal self-organization in a surface reaction: from the atomic to the mesoscopic scale*, Science **293**, 1635-1638 (2001).

[6] S.C. Müller, T. Plesser and B. Hess, *The structure of the core of the spiral wave in the Belousov-Zhabotinskii reaction*, Science **230**, 661-663 (1985).

reaction. In spatially extended systems, the latter leads to chemical waves and spiral defects, whose structure is well documented in the research literature.

Food for Thought

Problem 1

Consider the following chemical reactions

$$A + B \underset{k_2}{\overset{k_1}{\rightleftharpoons}} C \overset{k_3}{\rightarrow} B + D,$$

$$B + E \underset{k_5}{\overset{k_4}{\rightleftharpoons}} F.$$

1. Write differential equations describing the dynamics of the concentrations of A, B, C, E and F.
2. Show that $[B] + [C] + [F]$ is constant. What is the chemical significance of this fact?

Problem 2

Consider the chemical reaction

$$A + B \underset{k_2}{\overset{k_1}{\rightleftharpoons}} C.$$

1. Explain why the rate equation for A is $\dfrac{d[A]}{dt} = -k_1 [A] [B] + k_2 [C]$.
2. What are the dimensions of k_1 and k_2?
3. Write this equation in dimensionless form.

Problem 3

Consider the chemical reaction

$$n\,A \quad \xrightarrow{k} \quad B,$$

where n is a positive integer.

1. What is the rate equation for A?
2. Given that $[A](0) = [A]_0$, find the solution of this equation.
3. Does the solution make sense for all times? If not, what happens?

Problem 4

What are the dimensions of k_1, k_2, k_3 and k_4 in the Brusselator reactions?

Problem 5

Assume that the rate constants k_1, k_2, k_3 and k_4 in the Brusselator reactions are not equal to 1. Can you rescale the corresponding rate equations, in order to remove these parameters from the dimensionless model? Explain.

Problem 6

Using PPLANE, simulate the Brusselator model (8.1) with $[A] = 1$. Describe what happens as the parameter $[B]$ varies from 1.5 to 2.5.

Problem 7

Using PPLANE, simulate the Brusselator model (8.1) with $[A] = 1$.

1. Describe what happens as the parameter $[B]$ varies from 3 to 5. What is special with the value $[B] = 4$?
2. Plot $[X]$ and $[Y]$ as functions of t for $[B] = 6$. Describe what happens in your own words.

Problem 8

What are the dimensions of k_1, k_2, k_3, k_4 and k_5 in the Oregonator reactions?

Problem 9

Check that x, y, z and τ defined in the discussion of the Oregonator reactions are dimensionless.

Problem 10

Using PPLANE, simulate the reduced Oregonator model (8.3) with $f = 1$, $s = 1$ and $\omega = 2$. Describe what happens as the parameter q varies from 0.1 to 0.01.

Problem 11

Find the fixed points of the reduced Oregonator model (8.3) with $f = 1$, and analyze their stability. You may want to use a symbolic calculation package, such as MAPLE or MATHEMATICA.

Problem 12

Consider the chemical reactions

$$\begin{aligned} A + X &\xrightarrow{k_1} 2X, \\ X + Y &\xrightarrow{k_2} 2Y, \\ Y &\xrightarrow{k_3} P. \end{aligned}$$

1. Write the rate equations for $[X]$ and $[Y]$.
2. Show that the two-dimensional dynamical system hence obtained is a special case of the Lotka-Volterra model.
3. Based on this information, what kind of dynamics do you expect?

Problem 13

The Brusselator and Oregonator models exhibit periodic oscillations because of the existence of a limit cycle. On the other hand, the Lotka-Volterra equations (6.2) with $b = 0$ possess an infinite number of periodic solutions. If you had an experimental system which exhibited oscillations, how would you distinguish between the existence of a limit cycle and that of many periodic orbits?

9.

DIFFUSION

Learning Objectives

At the end of this chapter, you will be able to do the following.

- Construct continuity equations using ideas from vector calculus.
- Interpret reaction-diffusion equations as special cases of continuity equations.
- Solve the heat equation.
- Describe diffusion at the microscopic level.
- Compare diffusive processes at the microscopic and macroscopic levels.
- Justify the existence of a traveling wave solution to the Fisher-KPP equation.

Diffusion at the macroscopic level

Reaction-diffusion equations

Consider a quantity $F(x, y, z, t)$, which depends on the three space variables x, y, and z, as well as time t. Assume that F measures the density of some species, in number of individuals per unit volume. A similar treatment would apply to the number of individuals per unit area in a two-dimensional model, or per unit length in a one-dimensional model. We are interested in describing the evolution of F.

Consider a closed region of space Ω, and call N_i the number of individuals inside Ω. From the definition of F, we know that

$$N_i = \iiint_\Omega F(x,y,z,t)\,dV,$$

where dV denotes the volume element in Ω. Let us evaluate the time derivative of N_i. We will assume that F is regular enough to allow us to swap the derivative and integral signs, so that

$$\frac{dN_i}{dt} = \frac{d}{dt}\iiint_\Omega F(x,y,z,t)\,dV = \iiint_\Omega \frac{\partial F}{\partial t}\,dV. \qquad (9.1)$$

On the other hand, the change in N_i should reflect the local variations of N_i, as well as transport through the boundary $\partial\Omega$ of Ω. We thus have

$$\frac{dN_i}{dt} = \iiint_\Omega R(x,y,z,t,F)\,dV - \iint_{\partial\Omega} \vec{\jmath}(\vec{r})\cdot\vec{n}\,dS, \qquad (9.2)$$

where R is a *reaction term* describing the local change in F, $\vec{r}$ is the position vector, $\vec{n}$ is the normal to Ω pointing outwards, dS is the surface element on $\partial\Omega$, and $\vec{\jmath}$ is a vector describing the flow of N_i. The last term in (9.2) is the negative of the flux of $\vec{\jmath}$ through the boundary of Ω. The negative sign comes from the fact that $\vec{n}$ point outwards, and that individuals leaving Ω contribute to a decrease in N_i. With the divergence theorem (again, we implicitly assume all quantities have enough regularity) this term can be re-written as

$$-\iint_{\partial\Omega} \vec{\jmath}(\vec{r})\cdot\vec{n}\,dS = -\iiint_\Omega \vec{\nabla}\cdot\vec{\jmath}\,dV,$$

so that

$$\frac{dN_i}{dt} = \iiint_\Omega \left[R(x,y,z,t,F) - \vec{\nabla}\cdot\vec{\jmath}\right]\,dV. \qquad (9.3)$$

By combining Equations (9.1) and (9.3), we get

$$0 = \iiint_\Omega \left[-\frac{\partial F}{\partial t} + R(x,y,z,t) - \vec{\nabla}\cdot\vec{\jmath}\right]\,dV.$$

Since this equality is true for an arbitrary closed domain Ω, we obtain the following *continuity equation*

$$\frac{\partial F}{\partial t} = R(x,y,z,t,F) - \vec{\nabla}\cdot\vec{\jmath}, \qquad (9.4)$$

which describes the conservation of F. We now have to relate $\vec{\jmath}$ to F. As a first approximation, we will assume that *Fick's law* is valid, i.e. that $\vec{\jmath}$ only depends on the gradient of F,

$$\vec{\jmath} = -D\,\vec{\nabla}F,$$

where the *diffusion tensor* D is a matrix of diffusion coefficients. The entries of D may depend on F, in which case we have *nonlinear diffusion*. Depending on the properties of the medium, D may not be diagonal or even isotropic. In what follows, me make the simplifying assumption that $D = D_0 I_3$, where I_3 is the 3×3 identity matrix, and D_0 is a constant. We thus have that $\vec{\jmath}$ is proportional to the gradient of F,

$$\vec{\jmath} = -D_0 \vec{\nabla} F. \qquad (9.5)$$

Here D_0 is positive, which describes the fact that F "moves away" from regions regions of high concentration, towards regions of low concentration. By combining (9.4) with (9.5), we obtain the following *reaction-diffusion* equation,

$$\frac{\partial F}{\partial t} = R(x, y, z, t, F) + D_0 \Delta F, \qquad (9.6)$$

where $\Delta = \vec{\nabla}^2$ denotes the Laplacian. If F and R are uniform, i.e. $F(x, y, z, t) = G(t)$ and $R(x, y, z, t, F) = H(t, G)$, then Equation (9.6) reduces to an ordinary differential equation,

$$\frac{dG}{dt} = H(t, G),$$

where H can for instance be the logistic model, $H(t, G) = G(1 - G)$. We can thus extend all of the models discussed before by converting systems of ordinary differential equations into systems of partial differential equations, with appropriate diffusion terms.

The heat equation

In the absence of reaction terms, Equation (9.6) becomes the *heat equation*,

$$\frac{\partial F}{\partial t} = D_0 \Delta F. \qquad (9.7)$$

This is a linear equation in F. If appropriate initial and boundary conditions are supplied, a unique solution can be found in terms of Green's functions. On an infinite domain for instance, the solution to (9.7) with initial condition $F(x, y, z, 0) = H(x, y, z)$ is

$$F(x, y, z, t) = \left(\frac{1}{4\pi D_0 t}\right)^{3/2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[-\frac{|\vec{r} - \vec{\mu}|^2}{4 D_0 t}\right] H(\xi, \zeta, \eta)\, d\xi\, d\zeta\, d\eta,$$

where

$$|\vec{r} - \vec{\mu}|^2 = (x - \xi)^2 + (y - \zeta)^2 + (z - \eta)^2.$$

From a dimensional analysis point of view, we know that

$$[D_0] = \frac{L^2}{T}.$$

As a consequence, one expects the square of characteristic length scales of the problem to vary like characteristic time scales.

Diffusion at the microscopic level

A microscopic description of diffusion is as follows. Consider for instance a collection of molecules in a fluid, such as molecules of dye in water. If the temperature is non-zero, these molecules move randomly, and undergo what is called *Brownian motion*. What we call diffusion at the macroscopic level is the consequence of random motion at the microscopic level. To make this more intuitive, consider a particle undergoing a random walk in the plane: each step has a given length l, but can be taken in a random, uniformly distributed, direction. After N steps, of equivalently a time $T = N\delta t$, where δt is the time elapsed between any two consecutive steps, the particle will be at a distance L from its original position, such that

$$\langle L^2 \rangle = N l^2.$$

To see this, call $\vec{r}_i$ the position vector of the particle after i steps. Assume for simplicity that the particle started its walk from the origin. Since each step has length l,

$$|\vec{r}_1|^2 = l^2, \qquad \text{and} \qquad \langle |\vec{r}_1|^2 \rangle = l^2,$$

where the average $\langle \cdot \rangle$ is taken over all possible realizations of the particle taking one step. We will show by induction that after k steps, $k \in \mathbb{N}$, we have

$$\langle |\vec{r}_k|^2 \rangle = k\, l^2.$$

We know this statement is true for $k = 1$. We now show that if it is true for $k = n$, then it is also true for $k = n + 1$. We thus assume that

$$\langle |\vec{r}_n|^2 \rangle = n\, l^2,$$

and calculate $\langle |\vec{r}_{n+1}|^2 \rangle$. We have

$$\begin{aligned} |\vec{r}_{n+1}|^2 &= |[\vec{r}_n + (\vec{r}_{n+1} - \vec{r}_n)]|^2 \\ &= |\vec{r}_n|^2 + 2\vec{r}_n \cdot (\vec{r}_{n+1} - \vec{r}_n) + |(\vec{r}_{n+1} - \vec{r}_n)|^2. \end{aligned}$$

But

$$\vec{r}_{n+1} - \vec{r}_n = l\,[\cos(\theta_{n+1})\vec{\imath} + \sin(\theta_{n+1})\vec{\jmath}],$$

where $(\vec{\imath}, \vec{\jmath})$ is an orthonormal basis of the plane, and the angle θ_{n+1} is uniformly distributed. As a consequence,

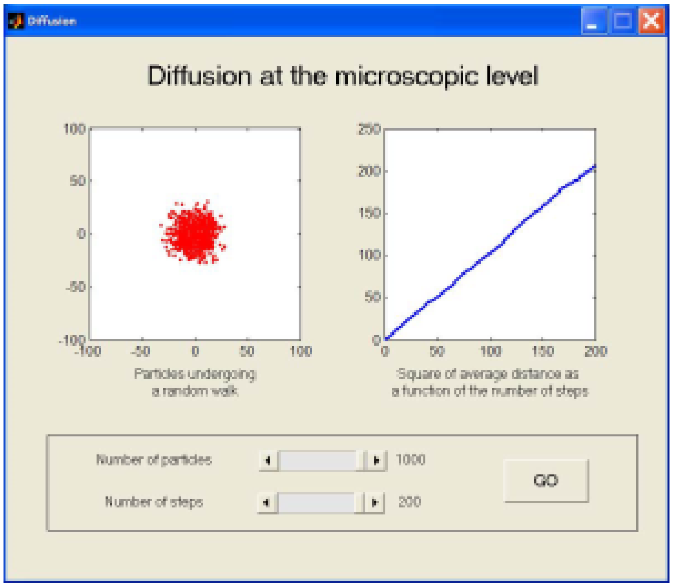
$$\begin{aligned} \langle \vec{r}_n \cdot (\vec{r}_{n+1} - \vec{r}_n) \rangle &= l\,(\vec{r}_n \cdot \vec{\imath}) \langle \cos(\theta_{n+1}) \rangle + l\,(\vec{r}_n \cdot \vec{\jmath}) \langle \sin(\theta_{n+1}) \rangle \\ &= 0, \end{aligned}$$

and $|(\vec{r}_{n+1} - \vec{r}_n)|^2 = l^2$. Thus,

$$\langle |\vec{r}_{n+1}|^2 \rangle = \langle |\vec{r}_n|^2 \rangle + 0 + l^2 = (n+1)\,l^2.$$

Therefore, $\langle |\vec{r}_k|^2 \rangle = k\,l^2$ for any positive integer k. Thus $\langle L^2 \rangle = Nl^2$ and since $T = N\delta t$, we have $\langle L^2 \rangle \propto T$.

Consider the following experiment. Many non-interacting particle are simultaneously released at the origin, and each particle performs an isotropic random walk in the plane. The above calculation tells us that after a period of time T, we expect to see a cloud of particles, and that if we measure the distance L between each particle and the origin, we should find that $\langle L^2 \rangle \propto T$.



Number of particles 1000

Number of steps 200

Figure 9.1. Graphic User Interface showing (left) the cloud of particles after two hundred steps, and (right) the linear relationship between $\langle L^2 \rangle$ and *N*.

To provide a visual understanding of this result, the *Diffusion* MATLAB GUI (see Figure 9.1) simulates the random motion of M non-interacting particles on a grid – so that each particle can only go up, down, left or right, with equal probability. All of the particles star their random walk from the origin, at the center of the box. For each particle, the distance L between its position after N steps (or equivalently after a time $T = N\delta t$, where δt is fixed) and the origin is measured as a function of N. The result is averaged over all of the particles, and plotted. The user can choose the maximum number of steps, N_{max}, as well as the number of particles M. The interface shows the position of all of the particles as time evolves. One of the particles is marked in blue, so that the user can follow its random walk. At the end of the simulation, a plot of the average of L^2 is shown as a function of N.

This simulation suggests that diffusion as modeled by the heat equation at the macroscopic level can be understood as resulting from the random walk of independent particles at the microscopic level. These two phenomena indeed have the same scaling properties. This correspondence can in fact be made rigorous, but we will not discuss this here. It is however useful to keep in mind the microscopic description of diffusion. Indeed, we often tend to develop macroscopic models based on our understanding of behaviors at the microscopic level. It is thus important to know under which conditions a particular phenomenon may be adequately modeled at the macroscopic level by diffusive terms.

The Fisher-Kolmogorov-Petrovsky-Piscounov equation

We now turn to a simple example of a one-dimensional reaction-diffusion equation, known as the Fisher-Kolmogorov-Petrovsky-Piscounov (Fisher-KPP) equation.[1][2] Consider the one-dimensional version of the logistic equation with diffusion,

$$\frac{\partial N}{\partial t} = rN\left(1 - \frac{N}{K}\right) + D\frac{\partial^2 N}{\partial X^2},$$

where r, K and D are constant. This equation can be made dimensionless by scaling space, time and the dependent variable N. To this end, we define

$$n = \frac{N}{K}, \qquad x = X\sqrt{\frac{r}{D}}, \qquad y = Y\sqrt{\frac{r}{D}}, \qquad \tau = r\,t,$$

and obtain the dimensionless Fisher-KPP equation,

$$\frac{\partial n}{\partial \tau} = n\,(1 - n) + \frac{\partial^2 n}{\partial x^2}. \qquad (9.8)$$

1. R.A. Fisher, *The wave of advance of advantageous genes*, Annu. Eugenics **7**, 255-369 (1937). Statement from the publisher, with which this author agrees: "The work of eugenicists was often pervaded by prejudice against racial, ethnic and disabled groups. Publication of this material online is for scholarly research purposes is not an endorsement or promotion of the views expressed in any of these articles or eugenics in general."
2. A. Kolmogorov, I. Petrovsky, N. Piscounoff, *Study of the diffusion equation with growth of the quantity of matter and its application to a biology problem*, Bulletin de l'Université d'état à Moscou, Ser. int., Section A, Vol. 1 (1937); translated in P. Pelcé, *Dynamics of curved fronts*, Academic Press, San Diego, 1988.

This equation has been widely studied in the literature[3], and we only mention below one of its properties, namely that it admits a family of traveling wave solutions defined on the real line.

Equation (9.8) has two uniform and constant solutions, given by $n = 0$ and $n = 1$, and we can look for a traveling wave solution connecting these two solutions. To do so, we set $n(x,t) = v(\xi)$, where $\xi = x - ct$ and the speed c is arbitrary, and substitute into (9.8). Using the chain rule, we obtain

$$\frac{\partial n}{\partial t} = -c\frac{dv}{d\xi}, \qquad \frac{\partial n}{\partial x} = \frac{dv}{d\xi},$$

so that v satisfies the ordinary differential equation

$$\frac{d^2v}{d\xi^2} + c\frac{dv}{d\xi} + v(1-v) = 0. \qquad (9.9)$$

The dynamics of this equation may be qualitatively described by looking at the corresponding phase plane. Let $\dfrac{dv}{d\xi} = w$. Then,

$$\frac{dw}{d\xi} = \frac{d^2v}{d\xi^2} = -cw - v(1-v),$$

and (9.9) is equivalent to the following dynamical system

$$\frac{d}{d\xi}\begin{pmatrix} v \\ w \end{pmatrix} = \begin{pmatrix} w \\ -cw - v(1-v) \end{pmatrix}. \qquad (9.10)$$

Its fixed points in the (v, w) plane are are $P_0 = (0,0)$ and $P_1 = (1,0)$. The Jacobian is

$$J(v,w) = \begin{pmatrix} 0 & 1 \\ -1+2v & -c \end{pmatrix},$$

and

$$J(0,0) = \begin{pmatrix} 0 & 1 \\ -1 & -c \end{pmatrix}, \qquad J(1,0) = \begin{pmatrix} 0 & 1 \\ 1 & -c \end{pmatrix}.$$

For P_1, $\det(J(P_1)) = -1$, so that P_1 is a saddle. The origin has eigenvaluesλ_1 and λ_2 with $\lambda_1\lambda_2 = 1 > 0$ and $\lambda_1 + \lambda_2 = -c$. We can assume without loss of generality that c is non-negative, since

[3] For a review, see for instance W. van Saarloos, *Front propagation into unstable states*, Physics Reports **386**, 29-222 (2003).

changing c into $-c$ is the same as changing ξ into $-\xi$ in Equation (9.9). If $c = 0$, the origin is a center. If $c > 0$, the origin is either a stable node or a stable spiral. Since P_1 is a saddle and the origin a node or a spiral, there is, for each value of $c > 0$, a trajectory which connects P_1 to P_0. This corresponds to a *front*, moving at speed c, and describing the growth of $n = 1$ into a region with $n = 0$.

The discriminant of the characteristic polynomial of $J(P_0)$, $c^2 - 4$, is positive for $c > 2$, in which case the origin is a stable node. The front solution is therefore monotonic if $c > 2$, and oscillatory if $0 < c < 2$. Figures 9.2 and 9.3 show the phase portraits of (9.10), for $c = 1$ and $c = 3$ respectively. In both cases, the front corresponds to the heteroclinic connection between P_1 and P_0, which is plotted as a thick solid line. If n describes a population density, it cannot become negative, and only speeds larger than 2 are thus possible. There is nevertheless a whole family of fronts, only one of which is selected by the partial differential equation (9.8).

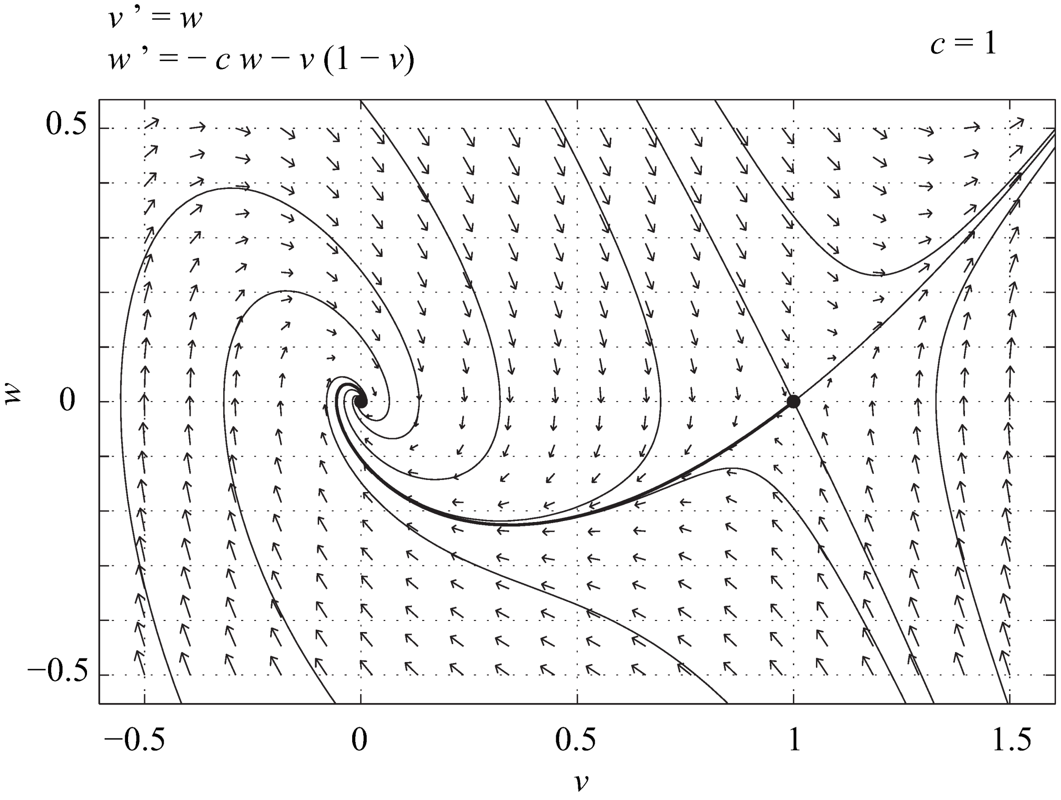


Figure 9.2. Phase plane of system (9.10), with $c = 1$, plotted with the software PPLANE.

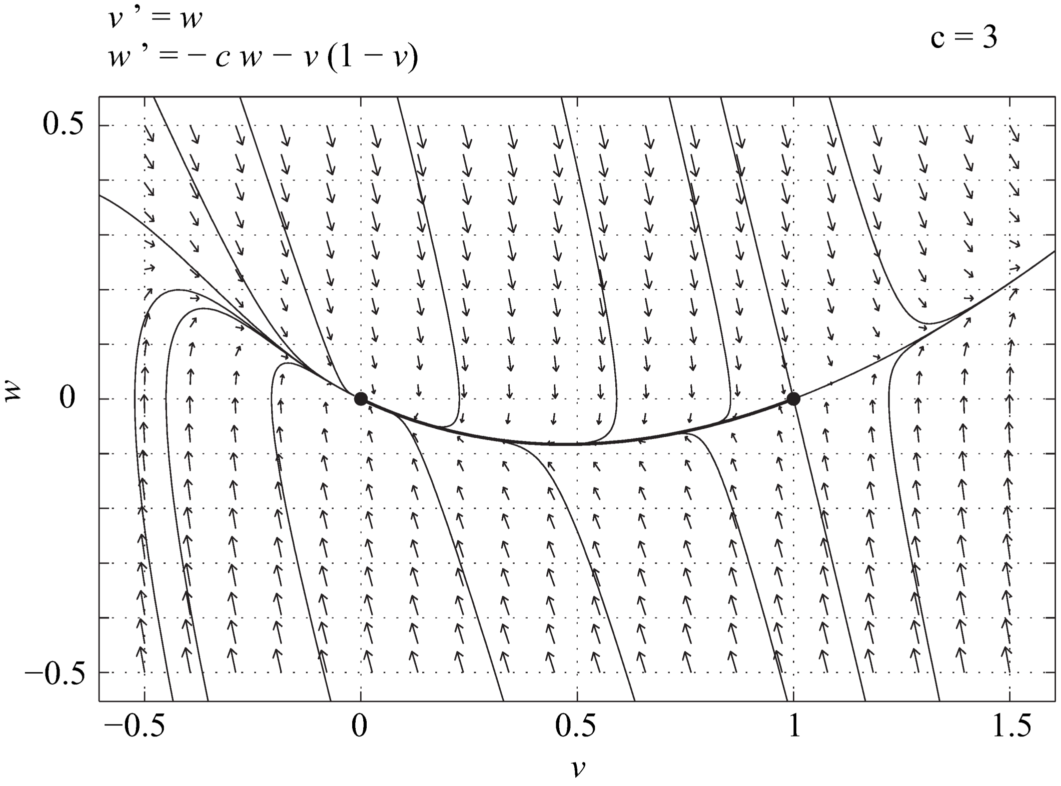


Figure 9.3. Phase plane of system (9.10), with $c = 3$, plotted with the software PPLANE.

Chemical waves

When a chemical reaction takes place in a spatially extended system, diffusion should be taken into account. As a consequence, the rate equations discussed in Chapter 8 are turned into partial differential equations. The reaction terms remain the same, but the concentration of each chemical now diffuses with a diffusion coefficient which depends on the size and weight of the molecule in question. If the reaction is oscillatory, wave fronts, corresponding to say large concentrations of some chemical, propagate in the system. Moreover, if the reaction is constrained to a two-dimensional surface and the wave is initiated at some point in space, then the wave fronts are circular. A 2001 article by C. Sachs *et al.*[4] describes how such wave fronts are observed in an experiment, and proposes a reaction-diffusion model which reproduces this behavior. The authors also discuss the limitation of reaction-diffusion models when macroscopic parameters are affected by the details of microscopic interactions.

What would happen if the wave front was broken, for instance if the wave went over a region where the reaction could not take place? If the wave front is anchored at one point, then it will curve and eventually form a spiral wave. Such waves are often observed in experiments where the Belousov-Zhabotinsky reaction is con-

[4] C. Sachs, M. Hildebrand, S. Völkening, J. Wintterlin, G. Ertl, *Spatiotemporal self-organization in a surface reaction: from the atomic to the mesoscopic scale*, Science **293**, 1635-1638 (2001).

strained to a two-dimensional surface, such as a thin film, a porous glass disk, or even a sheet of filter paper. For more information, the reader is referred to the original article by S.C. Müller *et al.*[5], which discusses experiments revealing the structure of the core of such spiral waves.

Summary

Reaction-diffusion equations extend ordinary differential equation models to entire spatial areas in one, two, or three dimensions. At the macroscopic level, diffusion describe the tendency of a quantity to spread out, by moving in a direction opposite to its local gradient. At the microscopic level, diffusion is associated with an isotropic random walk. Some reaction-diffusion equations admit traveling wave solutions, which may be found by means of dynamical systems methods.

Food for thought

Problem 1

Describe the behavior of system (9.10) near the origin if $c = 2$.

Problem 2

Consider the function

$$f(x,t) = \frac{1}{2\sqrt{\pi D_0 t}} \int_{-\infty}^{\infty} \exp\left[-\frac{(x-x_0)^2}{4D_0 t}\right] H(x_0)\, dx_0.$$

[5] S.C. Müller, T. Plesser and B. Hess, *The structure of the core of the spiral wave in the Belousov-Zhabotinskii reaction*, Science **230**, 661-663 (1985).

1. Calculate $\partial f/\partial t$.
2. Calculate $\partial^2 f/\partial x^2$.
3. Show that f solves the differential equation

$$\frac{\partial f}{\partial t} = D_0 \frac{\partial^2 f}{\partial x^2}.$$

Problem 3

Consider a collection of bacteria, which are *chemotactic* to food. This means that at the macroscopic level, the flow $\vec{j}$ of bacteria is given by

$$\vec{j} = \chi \vec{\nabla} n - D \vec{\nabla} b,$$

where n is the concentration of nutrients, b the density of bacteria, D a diffusion coefficient, and χ is a *chemotactic coefficient*.

Write a simple reaction-diffusion model for n and b, which takes into account bacterial motion, the diffusion of nutrients, and the fact that bacteria multiply by eating nutrients.

Problem 4

Consider the chemotactic bacteria described in Problem 3. Describe how you would modify the random motion of each bacterium at the microscopic level in order to include chemotaxis.

Problem 5

Consider a particle performing a one-dimensional random walk, such that its probability of taking a step to the right (resp. to the left) is p (resp. $1-p$), with p not necessarily equal to 1/2. Describe how far you expect the particle to have moved after N steps.

Problem 6

Consider the heteroclinic trajectory of Figure 9.2. Sketch the graph of v as a function of ξ. Explain why such a function cannot represent a population density.

Problem 7

Consider the heteroclinic trajectory of Figure 9.3. Sketch the graph of v as a function of ξ.

10.

PATTERN FORMATION

Learning Objectives

At the end of this chapter, you will be able to do the following.

- Explain why different sets of parameters in a single generic model may lead to different types of patterns.
- Analyze the linear stability of a homogeneous solution to a pattern-forming system.
- Predict the wavelength of the pattern that emerges above threshold.
- Recognize that different nonlinear terms lead to different types of patterns.
- Reconstruct a simple model for the formation of vegetation patterns.

Pattern Formation

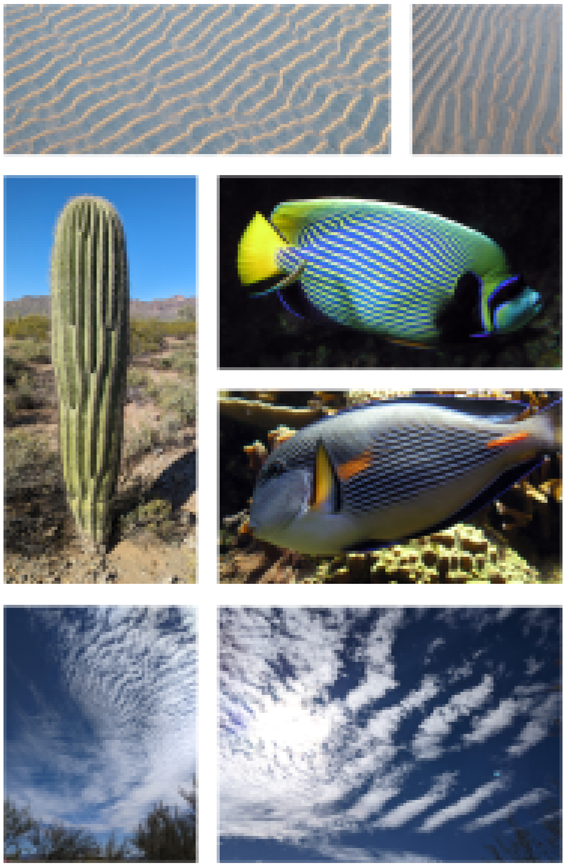


Figure 10.1. Stripe patterns in nature: sand ripples, saguaro ribs, colorful bands on fish coats, roll structures in clouds.

Patterns (see Figure 10.1), such as stripes and spots on animal coats or sand ripples on a beach, are very common in nature and in carefully controlled laboratory experiments. They typically occur in systems which are *driven far from equilibrium* by external forces or sources of energy. When the forcing is larger than what is necessary to balance inertial forces, the system responds by reorganizing itself into a periodic structure, called a pattern. There are many types of patterns, such as rolls, squares or hexagons, and they can be stationary or dynamic. The study of pattern-forming systems has been the subject of active research in physics, chemistry, nonlinear optics and applied mathematics for more than half a century[1]. In what follows, we will briefly mention Turing patterns, explore a canonical pattern-forming model, known as the complex Swift-Hohenberg equation, and discuss a reaction-diffusion model for vegetation patterns.

Turing patterns

In 1952, A.M. Turing[2] proposed the idea that differentiation patterns such as those selecting regions where a hydra grows its tentacles, were in fact chemical patterns, in particular those corresponding to stationary periodic structures[3]. Turing explained how such patterns could result from the instability of a homogeneous solution to a set of coupled reaction-diffusion equations. This instability was described as a *symmetry-breaking instability*, since the periodic structure that grew as a result of the instability broke the translational invariance of the initial homogeneous solution. It is only in the 1990s that Turing patterns were seen in chemical experiments[45],

[1] See for instance the book by P. Ball, entitled *The Self-made Tapestry: Pattern formation in nature* (Oxford, New York, 1999).

[2] A.M. Turing, *The chemical basis of morphogenesis*, Phil. Trans. R. Soc. London B **237**, 37-72 (1952).

[3] Turing also mentioned the possibility of wave patterns, but these were not the main focus of his paper.

[4] V. Castets, E. Dulos, J. Boissonade, and P. De Kepper, *Experimental evidence of a sustained standing Turing-type nonequilibrium chemical pattern*, Phys. Rev. Lett. **64**, 2953-2956 (1990).

[5] Q. Ouyang and H.L. Swinney, *Transition from a uniform state to hexagonal and striped Turing patterns*, Nature **352**, 610-612 (1991).

in the form of hexagons and stripes, but also as irregular structures corresponding to a state of chemical turbulence[6].

The complex Swift-Hohenberg equation

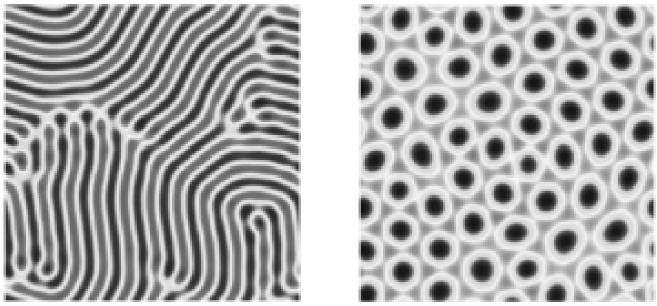


Figure 10.2. Numerical simulations (with periodic boundary conditions) of the complex Swift-Hohenberg equation showing stripe and hexagon patterns.

We will explore pattern formation by means of a single model, called the complex Swift-Hohenberg equation. Depending on whether the parameters of this model are real or complex, stationary or traveling patterns will be observed. Figure 10.2 shows stripe and hexagon patterns produced by this model. The general theory of pattern formation[7] explains why different models may produce patterns that are similar, and as a consequence why different chemical, physical or biological systems may display patterns that look alike. It also justifies the use of a generic pattern forming model to understand the mechanisms involved in pattern formation, as we do in the next section.

Consider the following partial differential equation

$$\frac{\partial\psi}{\partial t} = (\mu + i\nu)\psi - (\alpha + i\beta)\left(\Omega + \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^2 \psi + i\eta\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\psi - (\gamma + i\delta)|\psi|^2\psi - \zeta|\psi|^2, \qquad (10.1)$$

where all of the parameters are real, $\alpha > 0, \gamma > 0$, and $\psi(x, y, t)$ is *a priori* complex. This equation is a version of the Swift-Hohenberg equation[8] with complex coefficients[9] and a quadratic term. It is easy to see that $\psi = 0$ is a solution, and it is natural to ask under what conditions such a solution is also stable. We proceed in the same way as we did for systems of ordinary differential equations, that is we linearize Equation (10.1) about the solution $\psi = 0$. The linearized equation reads

$$\frac{\partial\psi}{\partial t} = (\mu + i\nu)\psi - (\alpha + i\beta)\left(\Omega + \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^2 \psi + i\eta\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\psi. \qquad (10.2)$$

[6] Q. Ouyang and H.L. Swinney, *Transition to chemical turbulence*, Chaos **1**, 411-420 (1991).

[7] For a review see for instance M.C. Cross, and P.C. Hohenberg, *Pattern formation outside of equilibrium*, Rev. Mod. Phys. **65**, 851-1112 (1993) and A. C. Newell, T. Passot and J. Lega, *Order parameter equations for patterns*, Ann. Rev. Fluid Mech. **25**, 399-453 (1993).

[8] J. Swift and P.C. Hohenberg, *Hydrodynamic fluctuations at the convective instability*, Phys. Rev. **A 15**, 319-328 (1977).

[9] J. Lega, J.V. Moloney, and A.C. Newell, *Swift-Hohenberg equation for lasers*, Phys. Rev. Lett. **73**, 2978-2981 (1994).

Typically, we are interested in spatially extended systems, which are large enough for patterns to develop (how large is "large" will be made more precise soon). We will thus only worry about instabilities that develop in the bulk, and consider that ψ can be decomposed into Fourier modes. From the linearized equation, we see that the rate of change (i.e. the derivative) of Fourier mode $u_{\vec{k}} \equiv a_{\vec{k}}\ \exp(i\vec{k}\cdot\vec{r})$ with wave vector $\vec{k} = (k_x, k_y)$ is $\lambda_{\vec{k}}\ u_{\vec{k}}$, where

$$\lambda_{\vec{k}} = \mu + i\nu - (\alpha + i\beta)(\Omega - k_x^2 - k_y^2)^2 - i\eta(k_x^2 + k_y^2).$$

The growth rate of this mode is thus

$$\sigma_k = \Re e(\lambda_{\vec{k}}) = \mu - \alpha(\Omega - k^2)^2,$$

where $k = \sqrt{k_x^2 + k_y^2} = ||\vec{k}||$. So each Fourier mode grows (or decays) at a rate that only depends on the magnitude k of the associated wave vector $\vec{k}$, and not on its direction. If $\mu < 0$, all Fourier modes decay, and the solution $\psi = 0$ is linearly stable. As μ increases, some Fourier modes will become unstable.

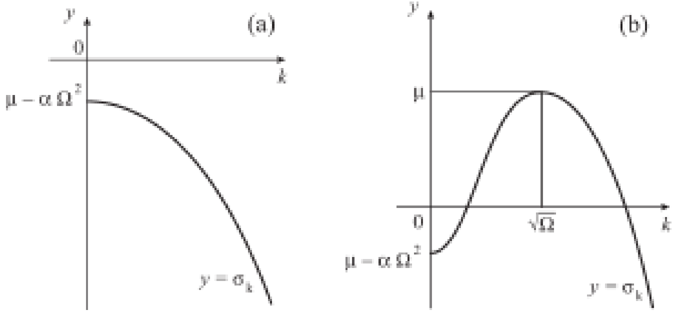


Figure 10.3. Sketch of the graph of the growth rate σk as a function of $k \geq 0$, for $0 < \mu < \alpha\ \Omega^2$. (a) Ω < 0; (b) Ω > 0$.

If $\Omega < 0$, the graph of σ_k as a function of $k \geq 0$ is monotonic (see Figure 10.3.a), and the first Fourier mode to become unstable is the $k = 0$ mode. This instability occurs for $\mu = \alpha\,\Omega^2$. If on the contrary $\Omega > 0$, then σ_k has a maximum for $k = \sqrt{\Omega} \equiv k_c$ (see Figure 10.3.b), and modes with $k = k_c$ become unstable when μ increases past zero. In this case, we expect a pattern to form above threshold with a characteristic length equal to $l_c = 2\pi/k_c$. The size of the system in which the pattern is to be observed should thus be much larger than l_c.

Above threshold, it is the nonlinear terms which decide which pattern is selected. Typically, if quadratic terms are present (i.e. if $\zeta \neq 0$ in Equation (10.1)), then hexagons are observed near threshold. On the contrary, if cubic terms dominate, then rolls (or stripes) prevail. If the imaginary parts of the coefficients in (10.1) are set to zero, the patterns that develop above threshold are stationary. If not, they are time-dependent. These and other aspects of the dynamics of Equation (10.1), including secondary instabilities and space-time disorder, may be explored with the MATLAB GUI interface called *Patterns*. This GUI allows the user to select the parameters and start a simulation from small, random initial conditions. Color-coded snapshots of the real part of the solution ψ are shown at successive times as the simulation progresses. After a simulation has ended, new parameters can be entered and a new simulation may then be restarted from the last solution.

The theory of pattern formation[10] provides means of describing the nonlinear dynamics of a pattern-forming system near and above threshold. A full discussion of this topic is however beyond the scope of these notes. In the next section, we briefly mention a reaction-diffusion model for the description of vegetation patterns in semiarid regions.

Vegetation patterns

In semiarid regions, plants (shrubs, grass, etc) tend to arrange themselves into stripes parallel to elevation contours on hilly terrain, and into patches on flat ground. Reaction-diffusion models have been proposed to explain these observations[11][12][13]. In one of these articles, C.A. Klausmeier introduces a simple model with two dependent variables, the plant biomass n and the amount of water w. Written in dimensionless form, the model reads

$$\frac{\partial w}{\partial t} = a - w - w\,n^2 + v\frac{\partial w}{\partial x},$$
$$\frac{\partial n}{\partial t} = w\,n^2 - m\,n + \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)n,$$

where a represents water input through rain, v is the speed at which water is transported downhill (in the $-x$ direction), $w\,n^2$ represents biomass increase due to water consumption, and m is the plant death rate. Moreover, water evaporates at rate w, and the diffusion of n accounts for the spreading of vegetation patches.

After finding equilibrium points (stationary and uniform solutions) of this model, the author investigates their linear stability. Figure 2 of the 1999 Klausmeier article shows a level set of the wavelength of the pattern that is expected to grow above threshold, in the (a, m) parameter plane, for a fixed value of v. More precisely, the (a, m) plane can be divided into three regions, one where the fixed point corresponding to no vegetation is stable, one where the fixed point corresponding to homogeneous vegetation is stable, and a region in between where linear stability analysis predicts the existence of vegetation patterns. A numerical simulation of the model in this latter regime confirms the existence of traveling stripes on hilly terrain, and of vegetation patches on flat ground. The speed v at which water flows downhill measures the steepness of the slope.

[10] See the articles in Footnote #7

[11] C.A. Klausmeier, *Regular and irregular patterns in semiarid vegetation*, Science **284**, 1826-1828(1999).

[12] R. Lefever, O. Lejeune and P. Couteron, *Generic modelling of vegetation patterns. A case study of Tiger bush in sub-saharian Sahel*, in Mathematical Models for Biological Pattern Formation, edited by P.K. Maini and H.G. Othmer, pp. 83-112, Springer, New York, 2001.

[13] J. von Hardenberg, E. Meron, M. Shachak, and Y. Zarmi, *Diversity of vegetation patterns and desertification*, Phys. Rev. Lett. **87**, 198101 (2001).

Summary

Different types of patterns may be obtained from the same, generic model, by choosing different sets of parameters. We have seen how to use linear stability analysis to predict whether a pattern can develop from a homogeneous solution as a *control parameter* (μ in the discussion of the Swift-Hohenberg equation) is varied, and if so, how to determine the typical length scale of the structure that will emerge. Although such a treatment is linear, it is nevertheless indicative of the kind of pattern one may expect. The same methodology, which consists in finding homogeneous solutions and studying their linear stability, can be applied to any pattern-forming system, as briefly illustrated in the case of the Klausmeier model.

Food for thought

Problem 1

Consider Equation (10.2) with $\mu = 0$. Find a solution to this equation in the form

$$\psi(x,t) = \exp[i(\omega t + qx + py)], \qquad \omega,\ q,\ p\ \in \mathbb{R}.$$

Try to express ω, q and p in terms of the parameters of this equation. Is this solution unique?

Problem 2

Find a solution of the form

$$\psi(x,t) = \exp[i(\omega t + qx + py)], \qquad \omega,\ q,\ p\ \in \mathbb{R}.$$

to Equation (10.2) with $\mu \neq 0$. Try to express ω, q and p in terms of the parameters of this equation. Is this solution unique?

Problem 3

Find a solution of the form

$$\psi(x,t) = \exp[\lambda t + i(\omega t + qx + py)], \qquad \lambda,\ \omega,\ q,\ p\ \in \mathbb{R}$$

to Equation (10.2) with $\mu \neq 0$. Try to express ω, q and p in terms of the parameters of this equation. Is this solution unique?

Problem 4

Read the 1999 article by C.A. Klausmeier. Explain how to go from their system of equations (1) to the dimensionless form (2).

Problem 5

Read the 1999 article by C.A. Klausmeier. What are the dimensions of the parameters A, L, R, V, J, M and D, and variables W and N in their system of equations (1)?

Problem 6

Read the 1999 article by C.A. Klausmeier.

- Explain how to find the fixed points of their system of equations (2).
- Find the Jacobian of (2).

PART V
APPENDICES

The rest of these notes is devoted to a review of typical concepts and methods presented in introductory courses on linear algebra (Chapter 11), vector calculus (Chapter 12), and ordinary differential equations (Chapter 13). Each section is meant to be used as a quick reference to support the various analyses conducted in the main chapters of the text. Results are stated without proof.

Chapter 14 gives examples of modeling projects that can be pursued while learning the material presented in this text.

11.

REFRESHER: LINEAR ALGEBRA

In this appendix, we state basic facts of linear algebra concerning matrices, eigenvalues and eigenvectors. No proofs are given and the reader should consult linear algebra texts for more details. The brief review presented below, although far from being complete, should however provide sufficient information for a reader to follow most of the linear stability arguments made in the previous chapters.

Vector spaces

Definitions

- $\mathcal{S}$ is a real (resp. complex) *vector space* if and only if it is closed under addition and under multiplication by a scalar. In other words,

$$\begin{aligned}&\forall x, y \in \mathcal{S}, x+y \in \mathcal{S}\\ &\forall x \in \mathcal{S}, \forall \alpha \in \mathbb{R}\ (\text{resp. } \mathbb{C}), \alpha x \in \mathcal{S}.\end{aligned}$$

- The vectors in $\{u_i \in S, i = 1 \ldots n\}$ are *linearly independent* if and only if any linear combination equal to zero must have all of its coefficients equal to zero. In other words,

$$\begin{aligned}&\forall \{\alpha_i, i = 1, \ldots n\} \subset \mathbb{R}\ (\text{or } \mathbb{C}),\\ &\sum_{i=1}^{n} \alpha_i\, u_i = 0 \Longrightarrow \alpha_i = 0, \forall i = 1, \ldots n.\end{aligned}$$

- $\mathcal{S}$ is *finite dimensional* if there exists a finite set of linearly independent vectors that *span* $\mathcal{S}$.
- Such a set is called a *basis* of $\mathcal{S}$. In what follows, we are only concerned with finite dimensional vector spaces.
- The *dimension* of a finite dimensional vector space $\mathcal{S}$ is the number of vectors in any basis of $\mathcal{S}$.

Linear mappings

We say that the mapping $\mathcal{T} : \mathcal{S} \to \mathcal{U}$ from a vector space $\mathcal{S}$ to a vector space $\mathcal{U}$ is *linear* if for every $x, y \in \mathcal{S}$

,

$$\begin{aligned} \mathcal{T}(x+y) &= \mathcal{T}(x) + \mathcal{T}(y) \\ \mathcal{T}(\alpha x) &= \alpha \mathcal{T}(x), \end{aligned}$$

where $\alpha \in \mathbb{R}$ (resp. $\alpha \in \mathbb{C}$) if $\mathcal{S}$ and $\mathcal{U}$ are vector spaces over $\mathbb{R}$ (resp. $\mathbb{C}$).

Properties of linear mappings

- The *range* $\mathcal{R}_\mathcal{T}$ of $\mathcal{T}$, which is the image of $\mathcal{S}$ under $\mathcal{T}$, is a linear subspace of $\mathcal{U}$.
- The *nullspace* or *kernel* of $\mathcal{T}$ is a linear subspace of $\mathcal{S}$. It is defined as the set $\mathcal{N}_\mathcal{T}$ of vectors of $\mathcal{S}$ whose image under $\mathcal{T}$ is zero,

$$\mathcal{N}_\mathcal{T} = \{x \in \mathcal{S} | \mathcal{T}(x) = 0\} .$$

- The dimensions of $\mathcal{R}_\mathcal{T}$ and $\mathcal{N}_\mathcal{T}$ are such that

$$\dim(\mathcal{R}_\mathcal{T}) + \dim(\mathcal{N}_\mathcal{T}) = \dim(\mathcal{S}).$$

Matrices

Every linear mapping

$$\begin{array}{cccc} \mathcal{T}: & \mathbb{R}^n & \to & \mathbb{R}^m \\ & x & \mapsto & u \end{array}$$

can be written as

$$u_i = \sum_{j=1}^{n} A_{ij} x_j \quad i = 1, \ldots, m; \qquad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{pmatrix}; \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix},$$

where $A \equiv (A_{ij})$ is a $m \times n$ (m rows, n columns) matrix with real entries. By convention, A_{ij} is the entry at the intersection of the i^{th} row and j^{th} column of A.

Note that once a basis has been chosen, every linear vector space of dimension n is isomorphic to $\mathbb{R}^n$. We can then represent any linear mapping between two finite dimensional vector spaces by a matrix. In what follows, we will only consider matrices with real coefficients.

Definitions

- The *transpose* of the matrix A is A^T such that $(A^T_{ij}) = (A_{ji})$.
- The *rank* of the matrix A associated with the linear transformation $\mathcal{T}$ is the dimension of $\mathcal{R}_{\mathcal{T}}$. It is also equal to the rank of A^T.
- The *determinant* of a 2×2 matrix, $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is

$$\det A = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

- The *determinant* of an $n \times n$ matrix $A = (A_{ij})$ can be calculated by means of the formula below, where i is one row of A and M_{ij} is the matrix obtained from A by deleting row i and column j:

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} A_{ij} \det M_{ij}.$$

- A similar formula exists for expanding $\det A$ with respect to one column of A.
- The *trace* Tr(A) of a square matrix A is the sum of the diagonal entries of A.

Properties

- If A is an $m \times n$ matrix, the system $Ax = b$ has at least one solution for every b if and only if the columns of A span $\mathbb{R}^m$. Then the rank of A, r, is such that $r = m$, which implies $m \leq n$.
- The system $Ax = b$ has at most one solution for every b if and only if the columns of A are linearly independent, i.e. if and only if the nullspace of A is trivial. Then, $r = n$, which implies $n \leq m$.
- Let A be an $n \times n$ matrix. Then, the following statements are equivalent.
  - The equation $AX = b$ has exactly one solution.
  - The range of A is $\mathbb{R}^n$.
  - The nullspace of A is trivial.

  - The matrix A is invertible.
  - The determinant of A, $\det A$, is non-zero.

Eigenvalues and eigenvectors

Definitions

Let A be a real $n \times n$ matrix.

- The vector $h \in \mathcal{S}$ is an *eigenvector* of A with *eigenvalue* $a \in \mathbb{C}$ if

$$Ah = ah, \qquad h \neq 0.$$

- The vector $f \in \mathcal{S}$ is a *generalized eigenvector* of A with *eigenvalue* a if, for some positive integer $m \neq 1$, we have

$$(A - aI_n)f \neq 0, \qquad (A - aI_n)^m f = 0, \qquad f \neq 0.$$

In the above equation, I_n is the $n \times n$ identity matrix. To find the eigenvalues and eigenvectors of a matrix, first note that if u is an eigenvector of a matrix A with eigenvalue a, then the equation

$$(A - a\, I_n)u = 0, \qquad (A1.1)$$

has a non-trivial solution. This implies that

$$\det(A - aI) = 0, \qquad (A1.2)$$

and one can therefore find the eigenvalues of A by solving this equation.

Properties

- The left-hand-side of (A1.2) is a polynomial of degree n in a, called the *characteristic polynomial* of A.
- The characteristic polynomial of A has n complex roots, which are the eigenvalues of A.
- Since A has real entries, if a is an eigenvalue of A, so is its complex conjugate a^*. As a consequence, the eigenvalues of A are either real, or complex conjugate pairs.
- The trace of A is the sum of the eigenvalues of A.
- The determinant of A is the product of the eigenvalues of A.

Once an eigenvalue is found, one needs to solve (A1.1) in order to obtain a corresponding eigenvector. There is not one such eigenvector, but a linear subspace thereof. Each of these eigenspaces is an invariant subspace of the linear transformation $\mathcal{T}$ associated with the matrix $\mathcal{A}$. The vector space $\mathcal{S}$, or equivalently $\mathbb{R}^n$, can thus be viewed as the sum of the eigenspaces of $\mathcal{A}$, and this decomposition gives a geometric picture of how $\mathcal{T}$ acts on $\mathcal{S}$.

Food for thought

Problem 1

Show that eigenvectors $\boldsymbol{u}$ and $\boldsymbol{v}$ of a matrix $\mathcal{A}$ corresponding to different eigenvalues are linearly independent.

Problem 2

Find the determinant of the following matrix

$$C = \begin{bmatrix} 2 & 2 & 6 \\ 4 & 3 & 12 \\ 6 & 4 & 16 \end{bmatrix}.$$

Problem 3

Find the eigenvalues and eigenvectors of the following matrix

$$B = \begin{bmatrix} 4 & -1 & 1 \\ -1 & 4 & -1 \\ -1 & 1 & 2 \end{bmatrix}.$$

Problem 4

Consider the transformation from $\mathbb{R}^5$ to $\mathbb{R}^3$ defined by

$$\mathcal{T}(\vec{x}) = \begin{bmatrix} 0 \\ 2x_1 - 4x_2 + x_5 \\ x_2 + x_3 + x_5 \end{bmatrix}, \qquad \text{where} \qquad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}.$$

1. Is $\mathcal{T}$ a linear transformation ? Why or why not ?
2. Find the matrix of $\mathcal{T}$ relative to the standard bases of $\mathbb{R}^5$ and $\mathbb{R}^3$.

Problem 5

Consider the matrix

$$A = \begin{bmatrix} 1 & 2 & 2 & 4 & 4 \\ 2 & 4 & 3 & 4 & 2 \\ 1 & 1 & 3 & 2 & 3 \end{bmatrix}.$$

1. Find a basis for the column space (or range) of A. Justify your answer.
2. Find a basis for the null space of A. Justify your answer.
3. What is the rank of A?

Problem 6

Consider the space $\mathbb{P}_2$ of polynomials of degree less than or equal to 2, and let $\mathcal{S} = \{q_0, q_1, q_2, q_3, q_4\}$ be a set of polynomials in $\mathbb{P}_2$, where

$$q_0(t) = 6t - t^2, \qquad q_1(t) = 1 - t, \qquad q_2(t) = t + 1$$
$$q_3(t) = 4, \qquad q_4(t) = 2 - 2t + t^2.$$

1. Find the coordinates of the polynomial Q relative to the standard basis of $\mathbb{P}_2$, where $Q(t) = 4t^2 - 2t + 10$.

2. Give a basis of $\mathbb{P}_2$ which consists of vectors in $\mathcal{S}$. Explain how you choose the vectors.
3. Find the coordinates of the polynomial Q defined in Question #1 relative to the basis you found in Question #2.

Problem 7

Consider the following vectors in $\mathbb{R}^4$.

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \qquad \vec{v}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix}, \qquad \vec{v}_3 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix},$$
$$\vec{v}_4 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ -2 \end{bmatrix}, \qquad \vec{v}_5 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Show that $\{\vec{v}_1, \vec{v}_2, \vec{v}_3, \vec{v}_4, \vec{v}_5\}$ is a linearly dependent set.

12.

REFRESHER: VECTOR CALCULUS

The information below is an elementary overview of the basic properties of the gradient of a function and of the divergence and curl of a vector field. These topics are typically covered in a third semester calculus course.

The gradient

The gradient of a differentiable function of three variables $f(x, y, z)$ is defined by

$$\text{grad} f = \vec{\nabla} f = \frac{\partial f}{\partial x}\vec{\imath} + \frac{\partial f}{\partial y}\vec{\jmath} + \frac{\partial f}{\partial z}\vec{k}$$

and has the following properties.

- $\vec{\nabla} f$ points in the direction where f is increasing the fastest.
- The rate of change of f in that direction is equal to $||\vec{\nabla} f||$.
- $\vec{\nabla} f$ is perpendicular to the level surfaces of f (i.e. to surfaces of equation $f = C$ = constant).
- The directional derivative of f in the direction of the unit vector $\vec{u}$ is equal to $\vec{\nabla} f \cdot \vec{u}$.
- Critical points of f are such that $\vec{\nabla} f = \vec{0}$ at these points. Generic critical points are minima, maxima and saddle points of the function f.
- The extrema of f subject to the constraint $g(x, y, z) = C$ are points on the curve of equation $g(x, y, z) = C$ where the gradient of f is parallel to the gradient of g. The constant of proportionality is called a *Lagrange multiplier*.

Line integrals and gradient fields

Line integrals

The line integral of a continuous vector field $\vec{F}$ along the oriented path $\mathcal{C}$ is written

$$\int_{\mathcal{C}} \vec{F}(\vec{r}) \cdot d\vec{r},$$

where $\vec{r}$ is the position vector and $d\vec{r}$ is tangent to the path $\mathcal{C}$. If we know a parametrization of $\mathcal{C}$, i.e. if $\mathcal{C}$ is drawn by $\vec{r}(t)$ when t varies between t_0 and t_1, the above line integral reads

$$\int_{\mathcal{C}} \vec{F} \cdot d\vec{r} = \int_{t_0}^{t_1} \vec{F}(\vec{r}(t)) \cdot \frac{d\vec{r}}{dt}\, dt,$$

and can thus be computed.

Example 1

Consider the vector field $\vec{F}(x, y) = e^x \vec{\imath} + e^y \vec{\jmath}$. Let $\mathcal{C}$ be the part of the ellipse $x^2 + 4y^2 = 4$, joining $(0, 1)$ to $(2, 0)$ in the clockwise direction. A parametrization of $\mathcal{C}$ is $x = 2\sin(t)$, $y = \cos(t)$, with t varying between 0 and $\pi/2$. Therefore,

$$\begin{aligned}
\int_{\mathcal{C}} \vec{F} \cdot d\vec{r} &= \int_0^{\pi/2} \left[e^{2\sin(t)} \vec{\imath} + e^{\cos(t)} \vec{\jmath}\right] \cdot [2\cos(t)\vec{\imath} - \sin(t)\vec{\jmath}]\, dt \\
&= \int_0^{\pi/2} \left[2\cos(t)e^{2\sin(t)} - \sin(t)e^{\cos(t)}\right]\, dt \\
&= \left[e^{2\sin(t)} + e^{\cos(t)}\right]_0^{\pi/2} = e(e-1).
\end{aligned}$$

Example 2

The work done by a force $\vec{F}$ along the path $\mathcal{C}$ is given by $W = \displaystyle\int_{\mathcal{C}} \vec{F} \cdot d\vec{r}$. If $\vec{r}$ moves according to Newton's law, $\vec{F} = m\, d^2\vec{r}/dt^2$, so that

$$W = \int_{t_0}^{t_1} m\, \frac{d^2\vec{r}}{dt^2} \cdot \frac{d\vec{r}}{dt}\, dt = \left[\frac{1}{2}m\left(\frac{d\vec{r}}{dt}\right)^2\right]_{t_0}^{t_1} = \left[\frac{1}{2}mv^2\right]_{t_0}^{t_1},$$

i.e. the work done by $\vec{F}$ as a point mass moves along the path $\mathcal{C}$ is equal to the difference in kinetic energy of the point mass between the end points of $\mathcal{C}$.

Gradient fields

A vector field $\vec{F}$ is a *gradient field* if there is a function $f(x, y, z)$ such that $\vec{F} = \vec{\nabla} f$. The vector field $\vec{F}$ is then *conservative*, *path-independent*, and *circulation free*.

- A *path-independent* vector field $\vec{F}$ is such that the line integral of $\vec{F}$ along any path in the domain of $\vec{F}$ only depends on the end points of the path.
- A *circulation free* vector field is such that its circulation (i.e. its line integral along any closed curve) is zero everywhere in the domain of $\vec{F}$.

The fundamental theorem of calculus for line integrals tells us that a gradient field is path-independent, i.e.

$$\int_{\mathcal{C}} \vec{\nabla} f \cdot d\vec{r} = f(Q) - f(P),$$

where $\mathcal{C}$ is a (piecewise) smooth path joining P to Q and $\vec{\nabla} f$ is continuous on $\mathcal{C}$. Conversely, any path-independent vector field is a gradient field.

Example

Consider the force $\vec{F} = m\dfrac{d^2\vec{r}}{dt^2}$ discussed above. If $\vec{F}$ has a potential function, i.e. if there exists a function $V(r)$ such that $\vec{F} = -\vec{\nabla} V$, we have

$$\begin{aligned} W &= \int_{\mathcal{C}} \vec{F} \cdot d\vec{r} = -\int_{\mathcal{C}} \vec{\nabla} V \cdot d\vec{r} = V(\vec{r}(t_0)) - V(\vec{r}(t_1)) \\ &= \left[\frac{1}{2}mv^2\right]_{t_0}^{t_1}. \end{aligned}$$

Thus, $\dfrac{1}{2}mv^2(t_0) + V\left(\vec{r}(t_0)\right) = \dfrac{1}{2}mv^2(t_1) + V\left(\vec{r}(t_1)\right)$, i.e. the total energy, which is the sum of the kinetic and potential energies, is conserved.

- If $\vec{F} = F_1\vec{\imath} + F_2\vec{\jmath} + F_3\vec{k}$ is a gradient field with continuous partial derivatives, then

$$\frac{\partial F_1}{\partial y} = \frac{\partial F_2}{\partial x}, \qquad \frac{\partial F_2}{\partial z} = \frac{\partial F_3}{\partial y}, \qquad \frac{\partial F_3}{\partial x} = \frac{\partial F_1}{\partial z},$$

i.e. $\text{curl}\vec{F} = \vec{0}$.

The curl

The curl of a vector field $\vec{F} = F_1\vec{\imath} + F_2\vec{\jmath} + F_3\vec{k}$ with continuous partial derivatives is a vector given by

$$\begin{aligned}\text{curl}\vec{F} = \vec{\nabla} \times \vec{F} &= \begin{vmatrix} \vec{\imath} & \vec{\jmath} & \vec{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{vmatrix} \\ &= \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right)\vec{\imath} + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right)\vec{\jmath} + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right)\vec{k}.\end{aligned}$$

It has the following properties.

- The direction of $\text{curl}\vec{F}$ at a point P is the direction around which the circulation density of $\vec{F}$ is the greatest.
- The magnitude of $\text{curl}\vec{F}$ is the circulation density around that direction.
- The circulation density of $\vec{F}$ around any unit vector $\vec{n}$ is equal to $\text{curl}\vec{F} \cdot \vec{n}$.
- For any function f with continuous second partial derivatives,

$$\text{curl grad} f = \vec{\nabla} \times \vec{\nabla} f = \vec{0}.$$

- Conversely, any smooth vector field defined on a domain with no codimension-one holes and whose curl is zero everywhere is a gradient field (this is the *curl test*).

Stokes's and Green's theorems

The flux of a vector field $\vec{F}$ through an oriented surface S is $\iint_S \vec{F} \cdot d\vec{S}$.

- If the surface S is the graph of a function $f(x, y)$, then, if R is the domain of $f(x, y)$, we may write

$$\iint_S \vec{F} \cdot d\vec{S} = \iint_R \vec{F}\left(x, y, z(x, y)\right) \cdot \left[-\frac{\partial f}{\partial x}\vec{\imath} - \frac{\partial f}{\partial y}\vec{\jmath} + \vec{k}\right] dxdy,$$

- If S is a parametric surface parametrized by $\vec{r} = \vec{r}(s, t)$ where s and t vary in a region R, and if we assume that $\frac{\partial \vec{r}}{\partial s} \times \frac{\partial \vec{r}}{\partial t}$ points in the direction of the normal $\vec{n}$ to the surface S everywhere, then we may write

$$\int_S \vec{F} \cdot d\vec{S} = \int_R \vec{F}\left(\vec{r}(s,t)\right) \cdot \left(\frac{\partial \vec{r}}{\partial s} \times \frac{\partial \vec{r}}{\partial t}\right) dsdt.$$

Stokes's theorem

Stokes's theorem links the circulation of a smooth vector field around a (piecewise) smooth closed curve $\mathcal{C}$ to the flux of $\mathrm{curl}\vec{F}$ through any smooth surface S whose boundary is equal to $\mathcal{C}$. It reads:

$$\int_{\mathcal{C}} \vec{F} \cdot d\vec{r} = \int_S \mathrm{curl}\vec{F} \cdot d\vec{A},$$

where the orientation of $\mathcal{C}$ is determined from the orientation of S (or vice-versa) by *the right hand-rule*. Note that Stokes's theorem is valid only if $\vec{F}$ is defined everywhere on S and $\mathcal{C}$.

Green's theorem

Green's theorem is a planar version of Stokes's theorem and reads

$$\int_{\mathcal{C}} \vec{F} \cdot d\vec{r} = \int_R \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right)\, dx\, dy,$$

where $\vec{F}(x,y) = F_1(x,y)\vec{\imath} + F_2(x,y)\vec{\jmath}$ is a smooth vector field defined at every point of $\mathcal{C}$ as well as inside $\mathcal{C}$, and $\mathcal{C}$ is a (simple) closed curve oriented such that its interior R is on the left as one moves along $\mathcal{C}$.

The divergence theorem

The divergence of a vector field $\vec{F} = F_1\vec{\imath} + F_2\vec{\jmath} + F_3\vec{k}$ with continuous partial derivatives is given by

$$\mathrm{div}\vec{F} = \vec{\nabla} \cdot \vec{F} = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}.$$

Note that it is a scalar quantity. It has the following properties.

- $\mathrm{div}(\vec{\nabla} f) = \vec{\nabla}^2 f = \Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}.$
- $\mathrm{div}(f\vec{\nabla} g) = f\vec{\nabla}^2 g + \vec{\nabla} f \cdot \vec{\nabla} g.$
- $\mathrm{div}(f\vec{g}) = f\mathrm{div}(\vec{g}) + \vec{\nabla} f \cdot \vec{g}.$

- The divergence of $\vec{F}$ at a point P is equal to the limit, when the surface S surrounding P is shrunk to zero, of the flux of $\vec{F}$ through S (oriented outward) divided by the volume of the region W bounded by S. In other words,

$$\text{div}\vec{F} = \lim_{W \to 0} \frac{\int_S \vec{F} \cdot d\vec{A}}{\text{volume of } W}.$$

- For any vector field $\vec{F}$ with continuous second partial derivatives,

$$\text{div curl}\vec{F} = \vec{\nabla} \cdot (\vec{\nabla} \times \vec{F}) = 0.$$

- Conversely, any smooth vector field $\vec{F}$ whose domain is closed and has no holes, and whose divergence is zero everywhere is a *curl field*, i.e. there exists a vector field $\vec{G}$ such that $\vec{F} = \text{curl}\vec{G}$ (this is the *divergence test*).

The divergence theorem relates the flux of a smooth vector field through a (piecewise) smooth closed surface S oriented outward to the volume integral of its divergence over the region W bounded by S, and reads

$$\int_S \vec{F} \cdot d\vec{A} = \int_W \text{div}\vec{F}\, dV,$$

assuming that $\vec{F}$ *is defined at every point* in W and on S.

Example: the continuity equation

Consider the flow $\vec{v}(\vec{r})$ of a fluid of density $\rho(\vec{r}, t)$. Call V a fixed region of the fluid domain. The mass of fluid in V is given by $m = \int_V \rho dV$, and its rate of change is

$$\frac{dm}{dt} = \int_V \frac{\partial \rho}{\partial t}\, dV,$$

if one assumes that ρ is smooth. On the other hand, this rate of change is given by the negative of the flux of matter $\rho\vec{v}$ through the boundary S of V, and by virtue of the divergence theorem,

$$\frac{dm}{dt} = -\int_S \rho\, \vec{v} \cdot d\vec{A} = -\int_V \text{div}(\rho\, \vec{v})\, dV.$$

By comparison of the two expressions of dm/dt, which are equal for every region V, we get the continuity equation, which reads

$$\frac{\partial \rho}{\partial t} + \mathrm{div}(\rho\,\vec{v}) = 0.$$

13.

REFRESHER: ORDINARY DIFFERENTIAL EQUATIONS

This appendix briefly reviews the most common methods for solving ordinary differential equations (ODEs). Many ODEs or systems of ODEs introduced in these notes are nonlinear and cannot therefore be easily solved with such methods. However, linear stability analysis of fixed points relies on solving linear systems with constant coefficients, which are discussed here. Moreover, it is essential for a modeler to be able to recognize ODEs that can be solved exactly, and to solve them if necessary. The information below is meant to be used as a quick reference; theorems are given without proof and the reader should consult classical texts on differential equations for details.

Definitions and basic existence theorems

Definitions

- An *ordinary differential equation* of order n is an equation of the form

$$\frac{d^n y}{dx^n} = F\left(x, y, \frac{dy}{dx}, \dots, \frac{d^{n-1}y}{dx^{n-1}}\right). \qquad (13.1)$$

- A *solution* to this differential equation is an n-times differentiable function $y(x)$ of a real variable x that satisfies Equation (13.1).
- An *initial condition* is the prescription of the values of y and of its $(n-1)$st derivatives at a point x_0. It takes the following form, where y_0, y_1, ... y_{n-1} are given numbers:

$$y(x_0) = y_0, \frac{dy}{dx}(x_0) = y_1, \dots \frac{d^{n-1}y}{dx^{n-1}}(x_0) = y_{n-1}, \qquad (13.2)$$

- *Boundary conditions* prescribe the values of linear combinations of y and its derivatives at two different values of x.

Existence and uniqueness theorems

We list below the main existence and uniqueness theorems for solutions to first-order systems of ordinary differential equations. The reader should note that Equation (13.1) may be written as a first-order system

$$\frac{dY}{dx} = f(x, Y) \qquad (13.3)$$

by setting $Y = \left(y, \frac{dy}{dx}, \frac{d^2y}{dx}, \cdots, \frac{d^{n-1}y}{dx^{n-1}}\right)$.

The Cauchy-Peano theorem

If f is continuous on the rectangle $\mathcal{R} = \{|x - x_0| \leq a, ||Y - Y_0|| \leq b\}$ where $a > 0$ and $b > 0$, then there exists a continuously differentiable solution Y of (13.3) on $|x - x_0| \leq \alpha$ for which $Y(x_0) = Y_0$, where

$$\alpha = \min\left(a, \frac{b}{M}\right), \quad M = \max_{(x,Y)\in\mathcal{R}} ||f(x, Y)||.$$

The Picard-Lindelöf Theorem

If f is Lipschitz on $\mathcal{R}$, i.e. if there exists a constant $k > 0$ such that

$$\forall(x, Y_1) \in \mathcal{R}, \forall(x, Y_2) \in \mathcal{R}, ||f(x, Y_1) - f(x, Y_2)|| < k||Y_1 - Y_2||,$$

and if f is continuous on $\mathcal{R}$, then there exists a unique solution Y to (13.3) on $|x - x_0| \leq \alpha$, such that $Y(x_0) = Y_0$.

Together, these theorems imply that for continuously differentiable dynamical systems of the form (13.3), there is a unique solution to every initial value problem. In other words, if f in (13.3) is continuously differentiable, then trajectories in the phase space of the dynamical system (13.3) cannot cross.

In the following, we list various common techniques used to solve differential equations. Initial or boundary conditions should be imposed after the general solution has been found.

First order differential equations

We start with a first order differential equation, which we write as

$$M(x, y)dx + N(x, y)dy = 0.$$

Separable equations

If $\dfrac{M(x,y)}{N(x,y)} = -f(x)g(y)$, then the equation is *separable*, and reads

$$f(x)dx = \frac{1}{g(y)}dy,$$

which can be integrated easily.

Example

The solution to the initial value problem $\dfrac{dy}{dx} = \dfrac{3x^2 + 4x + 2}{2(y-1)}$, $y(0) = -1$ is $y = 1 - \sqrt{x^3 + 2x^2 + 2x + 4}$.

Equations where *M* and *N* are both homogeneous of degree *n*

If $M(x, y)$ and $N(x, y)$ are both homogeneous of degree n, then the change of variable $v = y/x$ (or $u = x/y$) will make the ODE separable.

Example

The general solution to $\dfrac{dy}{dx} = \dfrac{y^2 + 2xy}{x^2}$ is $y = \dfrac{Cx^2}{1 - Cx}$, where C is an arbitrary constant.

Exact equations

If $\partial M/\partial y$ and $\partial N/\partial x$ are continuous and if $\dfrac{\partial M}{\partial y} = \dfrac{\partial N}{\partial x}$, then the ODE is *exact*, i.e. there exists $u(x, y)$ such that

$$\frac{\partial u}{\partial x} = M(x,y) \qquad \text{and} \qquad \frac{\partial u}{\partial y} = N(x,y).$$

Then the ODE can be written in the form

$$\frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy = 0,$$

which gives $d\left(u(x,y)\right) = 0$, i.e. $u(x,y) = C$. The function u can be found by integrating the two equations

$$\frac{\partial u}{\partial x} = M(x,y) \quad \text{and} \quad \frac{\partial u}{\partial y} = N(x,y).$$

Example

The general solution to $(y\cos(x) + 2xe^y) + (\sin(x) + x^2e^y - 1)y' = 0$ is implicitly given by $y\sin(x) + x^2e^y - y = C$, where C is an arbitrary constant.

Integrating factor

An integrating factor is a function $\rho(x,y)$ such that the differential equation

$$\rho(x,y)\left[M(x,y)dx + N(x,y)dy\right]$$

is exact. If $\dfrac{\partial M}{\partial y} \neq \dfrac{\partial N}{\partial x}$, but if

- $\dfrac{\partial M}{\partial y} - \dfrac{\partial N}{\partial x} = -n\dfrac{M}{y} + m\dfrac{N}{x}$, try the integrating factor $\rho(x,y) = x^m y^n$.
- If $\dfrac{1}{N}\left[\dfrac{\partial M}{\partial y} - \dfrac{\partial N}{\partial x}\right]$ is a function of x only, try the integrating factor $\rho(x,y) = f(x)$.
- If $\dfrac{1}{M}\left[\dfrac{\partial M}{\partial y} - \dfrac{\partial N}{\partial x}\right]$ is a function of y only, try the integrating factor $\rho(x,y) = f(y)$.

Example

The general solution to $3xy + y^2 + (x^2 + xy)\dfrac{dy}{dx} = 0$ is implicitly given by $x^3y + \dfrac{1}{2}x^2y^2 = C$.

Linear equations

If the differential equation can be written in the form $y' + p(x)y = q(x)$, i.e. if

$$-\frac{M(x,y)}{N(x,y)} = -p(x)y + q(x),$$

then the ODE is *linear*. Let $\rho(x) = \exp\left[\int p(x)dx\right]$. We have

$$\rho'(x) = p(x)\exp\left[\int p(x)dx\right] = p(x)\rho(x),$$

and by multiplying the ODE by $\rho(x)$, we get

$$\begin{aligned} & y'\rho(x) + p(x)\rho(x)y = q(x)\rho(x) \\ \Longleftrightarrow & \frac{d}{dx}[\rho(x)y] = q(x)\rho(x), \\ \Longleftrightarrow & y\rho(x) = \int q(x)\rho(x)dx + C, \\ \Longleftrightarrow & y = \frac{1}{\rho(x)}\int q(x)\rho(x)dx + \frac{C}{\rho(x)}, \\ \Longleftrightarrow & y = \left[\int q(x)\rho(x)dx + C\right]\exp\left[-\int p(x)\,dx\right]. \end{aligned}$$

Example 1

The general solution to $y' + 2y = e^{-x}$ is $y = e^{-x} + Ce^{-2x}$.

Example 2

The solution to $y' - 2xy = x, y(0) = 0$ is $y = -\frac{1}{2} + \frac{1}{2}e^{x^2}$.

Bernoulli's equation

If $-\frac{M(x,y)}{N(x,y)} = -p(x)y + q(x)y^n$, with $n \neq 0, 1$, we obtain the following Bernoulli's equation

$$y' + p(x)y = q(x)y^n.$$

This nonlinear equation can be brought to the form of a linear equation by the change of variable $u = y^{1-n}$.

Example

The general solution to $x^2y' + 2xy - y^3 = 0, x > 0$ is $y = \pm\sqrt{\frac{5x}{2 + Cx^5}}$.

Riccati's equation

If $-\frac{M(x,y)}{N(x,y)} = a_0(x) + a_1(x)y + a_2(x)y^2$, we have the following Riccati's equation

$$y' = a_0(x) + a_1(x)y + a_2(x)y^2.$$

To solve it, proceed as follows.

1. Find a particular solution $y = y_1(x)$ by trying simple functions such as ax^b or $a\exp(bx)$.
2. Note that $u = y - y_1$ satisfies the following Bernoulli's equation (where $n = 2$):
   $$\begin{aligned} u' &= a_1(x)u + a_2(x)u(2y_1(x) + u) \\ &= [a_1(x) + 2a_2(x)y_1(x)]\, u + a_2(x)u^2, \end{aligned}$$
3. Convert the above equation into a linear equation by applying the change of variable $w = u^{1-2} = 1/u$, and solve for w. In terms of the original variable, the solution y is given by $y = y_1 + 1/w$.

Example

The general solution to $y' = 1 + x^2 - 2xy + y^2$ is $y = x + 1/(C - x)$.

Clairaut's equation

The differential equation $y = xy' + f(y')$ has for solution the family of straight lines

$$y = cx + f(c), \qquad c = \text{ constant}$$

and may also have a singular solution in the parametric form

$$\begin{cases} x = -f'(p) \\ y = px + f(p) \end{cases}.$$

Example

The general solution to $y = y'x + \frac{1}{y'}$ is $y = Cx + 1/C$ and a singular solution is $y^2 = 4x$.

Second order differential equations

Equations without *y*

Second order differential equations without y are of the form $F(y'', y', x) = 0$. With $w = y'$, this equation reads $F(w', w, x) = 0$, which is a first order equation. It can first be solved for w and then for y, using $w = dy/dx$.

Example

The general solution to $x^2y'' + y'^2 - 2xy' = 0$ is

$$y = \frac{1}{2}x^2 - C_1x + C_1^2 \ln(|x + C_1|) + C_2,$$

where C_1 and C_2 are arbitrary constants.

Equations without *x*

Such ODEs read $F(y'', y', y) = 0$. Let $w = dy/dx$ be a function of y. Then,

$$\frac{d^2y}{dx^2} = \frac{dw}{dx} = \frac{dw}{dy}\frac{dy}{dx} = \frac{dw}{dy}w,$$

and the ode becomes $F(w\frac{dw}{dy}, w, y) = 0$, which is a first order differential equation where the independent variable is y. One can solve this equation for w in terms of y and then solve the first order separable equation $\frac{dy}{dx} = w(y)$.

Example

The general solution to $yy'' + (y')^2 + 4 = 0$ is implicitly given by

$$x^2 = \left(C_2 + \frac{1}{4C_1}\sqrt{1 - 4C_1^2 y^2}\right)^2.$$

Linear equations

The general solution to $a_2(x)y'' + a_1(x)y' + a_0(x)y = h(x)$ reads

$$y(x) = C_1 y_1(x) + C_2 y_2(x) + y_p(x),$$

where

- $y_p(x)$ is a particular solution to $a_2(x)y'' + a_1(x)y' + a_0(x)y = h(x)$,
- y_1 and y_2 are two linearly independent solutions of the associated homogeneous equation $a_2(x)y'' + a_1(x)y' + a_0(x)y = 0$,
- C_1 and C_2 are two arbitrary constants.

Therefore, the above linear equation is solved in two steps:

1. Solve the homogeneous equation (i.e. find two linearly independent solutions y_1 and y_2),
2. Find a particular solution to the full equation.

How to solve the homogeneous equation

The general solution y to the homogeneous equation

$$a_2(x)y'' + a_1(x)y' + a_0(x)y = 0$$

is a linear combination of two of its linearly independent solutions y_1 and y_2, i.e. $y = C_1 y_1 + C_2 y_2$ where C_1 and C_2 are constants.

Equations with constant coefficients

Assume that the homogeneous differential equation reads

$$a_2 y'' + a_1 y' + a_0 y = 0,$$

where a_2, a_1 and a_0 are constants. The associated characteristic equation is $a_2 r^2 + a_1 r + a_0 = 0$.

- If the characteristic equation has 2 real distinct roots r_1 and r_2, then two linearly independent solutions to the homogeneous equation are

$$y_1(x) = \exp(r_1 x) \qquad \text{and} \qquad y_2(x) = \exp(r_2 x).$$

- If the characteristic equation has 2 complex roots $\alpha \pm i\beta$, two linearly independent solutions are

$$y_1(x) = \exp(\alpha x)\cos(\beta x) \qquad y_2(x) = \exp(\alpha x)\sin(\beta x).$$

- If the characteristic equation has 1 real repeated root r, two linearly independent solutions are

$$y_1(x) = \exp(rx) \qquad y_2(x) = x\exp(rx).$$

Example 1

The solution to $y'' + y' - 2y = 0$, $y(0) = 0$ and $y'(0) = 3$ is $y = e^x - e^{-2x}$.

Example 2

The general solution to $y'' - 4y' + 4y = 0$ is $y = C_1 e^{2x} + C_2 x e^{2x}$.

Cauchy-Euler equation

If the ODE reads

$$a_2 x^2 y'' + a_1 x y' + a_0 y = 0, \quad a_2, a_1, \text{and } a_0 \text{ constants},$$

we look for solutions in the form $y = x^r = \exp(r\ln(x))$ (defined for $x > 0$). Then, r must satisfy

$$a_2 r(r-1) + a_1 r + a_0 = 0.$$

- If this equation has 2 real distinct roots r_1 and r_2, then two linearly independent solutions are

$$y_1 = x^{r_1} \qquad y_2 = x^{r_2},$$

- If this equation has 2 complex roots $\alpha \pm i\beta$, two linearly independent solutions are

$$y_1(x) = x^\alpha\cos(\beta\ln(x)) \qquad y_2(x) = x^\alpha\sin(\beta\ln(x)),$$

- If this equation has 1 real repeated root r, two linearly independent solutions are

$$y_1(x) = x^r \qquad y_2(x) = x^r \ln(x).$$

Note: If you have to solve for $x < 0$, first make the change of variable $t = -x$.

Example

$y = C_1 x^2$ solves $x^2 y'' - 3xy' + 4y = 0, y(0) = 0, y'(0) = 0.$

Other equations

If you have a particular solution y_{ph} (for instance found by inspection) to the homogeneous equation, you may apply the method of *variation of constants*: look for another solution in the form $y = v(x)y_{ph}(x)$, substitute into the homogeneous equation and solve the first order equation for v. This procedure is called *reduction of order*.

How to find a particular solution to the full equation

We now turn to $a_2(x)y'' + a_1(x)y' + a_0(x)y = h(x)$ and look for a particular solution y_p to this equation. As before, there are a few of special cases for which a systematic method of solution exists.

Equations with constant coefficients: Method of undetermined coefficients

If the ode has constant coefficients and if

$$h(x) = P_m(x)\exp(\alpha x)\cos(\beta x) + Q_m(x)\exp(\alpha x)\sin(\beta x),$$

where P_m and Q_m are polynomials of degree m, then try the particular solution given below, where K_m and L_m are polynomials of degree m in each case.

- $y_p = K_m(x)\exp(\alpha x)\cos(\beta x) + L_m(x)\exp(\alpha x)\sin(\beta x)$ if $\alpha \pm i\beta$ are not roots of the associated characteristic equation,
- $y_p = x^h\left[K_m(x)\exp(\alpha x)\cos(\beta x) + L_m(x)\exp(\alpha x)\sin(\beta x)\right]$ if $\alpha \pm i\beta$ are roots of multiplicity h of the associated characteristic equation (h can be 1 or 2).

Note: Since the equation is linear, it might be useful to use the principle of superposition: if $h(x) = h_1(x) + h_2(x)$, and if y_{p_1} and y_{p_2} are particular solutions to $a_2(x)y'' + a_1(x)y' + a_0(x)y = h_1(x)$ and $a_2(x)y'' + a_1(x)y' + a_0(x)y = h_2(x)$ respectively, then $y_p = y_{p_1} + y_{p_2}$ is a particular solution to $a_2(x)y'' + a_1(x)y' + a_0(x)y = h(x)$.

Example 1

A particular solution to $y'' + y' - 2y = 4\sin(2x)$ is $y_p = -\frac{1}{5}\cos(2x) - \frac{3}{5}\sin(2x)$.

Example 2

A particular solution to $y'' - 4y' + 4y = 12xe^{2x}$ is $y_p = 2x^3 e^{2x}$.

Cauchy-Euler equations

The change of variable $x = e^t$ will turn a Cauchy-Euler equation into an equation with constant coefficients, which can then be solved as described above.

Other equations

If the ODE does not have constant coefficients, or if $h(x)$ is not of the form discussed above, you may want to try using the method of variation of constants: look for a particular solution $y_p = v_1(x)y_1(x) + v_2(x)y_2(x)$, where y_1 and y_2 are two linearly independent solutions to the associated homogeneous equation and solve for v_1 and v_2 after imposing

$$v_1'(x)y_1(x) + v_2'(x)y_2(x) = 0.$$

Example

The general solution to $y'' + 2y' + y = 2x^{-2}e^{-x}, x > 0$ is

$$C_1 e^{-x} + C_2 x e^{-x} - 2\left(1 + \ln(x)\right)e^{-x}.$$

Note: There is a convenient way to check that two functions y_1 and y_2 are linearly independent: their Wronskian

$$W[y_1, y_2] = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix}$$

must be nonzero.

Linear differential equations of order higher than two

The general solution to

$$a_n(x)y^{(n)} + a_{n-1}(x)y^{(n-1)} + \cdots + a_1(x)y' + a_0(x)y = h(x),$$
$$a_n(x) \neq 0$$

reads

$$y(x) = C_n y_n(x) + C_{n-1}y_{n-1}(x) + \cdots + C_1 y_1(x) + y_p(x),$$

where

- $y_p(x)$ is a particular solution to the full equation

$$a_n(x)y^{(n)} + a_{n-1}(x)y^{(n-1)} + \cdots + a_2(x)y'' + a_1(x)y' + a_0(x)y = h(x),$$

- $y_i, i = 1, \ldots, n$, are n linearly independent solutions to the associated homogeneous equation

$$a_n(x)y^{(n)} + a_{n-1}(x)y^{(n-1)} + \cdots + a_2(x)y'' + a_1(x)y' + a_0(x)y = 0,$$

- $C_i, i = 1, \ldots, n$, are n arbitrary constants.

Therefore, the method to solve the linear equation is as follows.

1. Solve the homogeneous equation (i.e. find n linearly independent solutions $y_1 \ldots, y_n$),
2. Find a particular solution to the full equation.

How to solve the homogeneous equation

The general solution y to the homogeneous equation

$$a_n(x)y^{(n)} + a_{n-1}(x)y^{(n-1)} + \cdots + a_2(x)y'' + a_1(x)y' + a_0(x)y = 0$$

is a linear combination of n of its linearly independent solutions $y_n, \ldots, y_1$, i.e. $y = C_n y_n + \cdots + C_1 y_1$ where the C_i are constants.

Equations with constant coefficients

Suppose the homogeneous equation reads

$$a_n y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_2 y'' + a_1 y' + a_0 y = 0,$$

where the a_i are real constants. The associated characteristic equation is

$$a_n r^n + \cdots + a_1 r + a_0 = 0.$$

- If the characteristic equation has only simple roots r_i, you can find n linearly independent solutions by using $y_j(x) = \exp(r_j x)$ for real roots, and the following solutions for any set of complex conjugate roots $r_i = \alpha_i + i\beta_i$ and $r_{i+1} = \alpha_i - i\beta_i$:

$$\begin{cases} y_i(x) & = & \exp(\alpha_i x)\cos(\beta_i x) \\ y_{i+1}(x) & = & \exp(\alpha_i x)\sin(\beta_i x) \end{cases}$$

- If the characteristic equation has one or more roots of multiplicity greater than one, use the above formula for roots of multiplicity one, and for any root r_i of multiplicity h, use

$$\begin{aligned} y_i &= \exp(r_i x), \\ y_{i+1} &= x\exp(r_i x), \\ y_{i+2} &= x^2\exp(r_i x), \\ &\vdots \\ y_{i+h-1} &= x^{h-1}\exp(r_i x), \end{aligned}$$

Again, if $r_i = \alpha_i \pm i\beta_i$, it is customary to use $x^k \exp(\alpha_i x)\cos(\beta_i x)$ and $x^k \exp(\alpha_i x)\sin(\beta_i x)$ instead of $x^k \exp(r_i x)$ (since the ODE has real coefficients).

Example 1

The general solution to $y^{(4)} - 8y'' - 9y = 6x + 12\sin(x)$ is

$$y = C_1\cos(x) + C_2\sin(x) + C_3 e^{3x} + C_4 e^{-3x} - \frac{2}{3}x + \frac{3}{5}x\cos(x).$$

Example 2

The solution to $y^{(4)} - 2y^{(3)} = 0$ with $y(0) = 0, y'(0) = 0, y''(0) = 0$ and $y^{(3)}(0) = 4$ is

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

Cauchy-Euler equation

To solve

$$a_n x^n y^{(n)} + a_{n-1} x^{n-1} y^{(n-1)} + \cdots + a_2 x^2 y'' + a_1 x y' + a_0 y = 0, \qquad x > 0$$

where the a_i are real constants, make the change of variable $x = \exp(t) = e^t$. This gives an equation with constant coefficients for y (as a function of t), which can be solved as described above. Then, the change of variable $t = \ln(x)$ gives y in terms of x.

Note: If you have to solve for $x < 0$, make the change of variable $x = -\exp(t)$, and proceed as before (but now, $t = \ln(-x)$).

Example

The general solution to $x^3 y^{(3)} - 2x^2 y'' + 8xy' - 8y = 0$, $x > 0$ is

$$C_1 x + C_2 x^2 \cos(2\ln(x)) + C_3 x^2 \sin(2\ln(x)).$$

How to find a particular solution to the full equation

We now look for a particular solution to

$$a_n(x) y^{(n)} + a_{n-1}(x) y^{(n-1)} + \cdots + a_2(x) y'' + a_1(x) y' + a_0(x) y = h(x).$$

If the above equation has constant coefficients and if

$$h(x) = P_m(x)\exp(\alpha x)\cos(\beta x) + Q_m(x)\exp(\alpha x)\sin(\beta x),$$

where P_m and Q_m are polynomials of degree m, then try the following particular solution, where K_m and L_m are polynomials of degree m in each case.

- $y_p = K_m(x)\exp(\alpha x)\cos(\beta x) + L_m(x)\exp(\alpha x)\sin(\beta x)$ if $\alpha \pm i\beta$ are not roots of the associated characteristic equation.
- $y_p = x^h \left[K_m(x)\exp(\alpha x)\cos(\beta x) + L_m(x)\exp(\alpha x)\sin(\beta x)\right]$ if $\alpha \pm i\beta$ are roots of multiplicity h of the associated characteristic equation.

If the ODE does not have constant coefficients, or if $h(x)$ is not of the form discussed above, use the method of variation of constants: look for a particular solution $y_p = v_n(x) y_n(x) + \cdots + v_1(x) y_1(x)$ where the y_i are n linearly independent solutions of the associated homogeneous equation, and solve for the v_i after imposing the $n-1$ following conditions:

$$\begin{cases} v_n'(x)y_n(x) & + & \dots & + & v_1'(x)y_1(x) & =0 \\ v_n'(x)y_n'(x) & + & \dots & + & v_1'(x)y_1'(x) & =0 \\ \vdots & & \ddots & & \vdots & \\ v_n'(x)y_n^{(n-2)}(x) & + & \dots & + & v_1'(x)y_1^{(n-2)}(x) & =0. \end{cases}$$

Note: If the Wronskian

$$W[y_1,\dots,y_n]=\begin{vmatrix} y_1 & y_2 & \dots & y_n \\ y_1' & y_2' & \dots & y_n' \\ \vdots & \vdots & \ddots & \vdots \\ y_1^{(n-1)} & y_2^{(n-1)} & \dots & y_n^{(n-1)} \end{vmatrix}$$

of $y_1,\dots,y_n$ is nonzero, then these functions are linearly independent.

Systems of first order linear differential equations

We consider systems of the form

$$\begin{cases} \dfrac{dx_1}{dt} & =a_{11}(t)x_1(t) & +a_{12}(t)\,x_2(t) & +\dots & +a_{1n}(t)x_n(t) & +b_1(t) \\ \dfrac{dx_2}{dt} & =a_{21}(t)x_1(t) & +a_{22}(t)x_2(t) & +\dots & +a_{2n}(t)x_n(t) & +b_2(t) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \dfrac{dx_n}{dt} & =a_{n1}(t)x_1(t) & +a_{n2}(t)x_2(t) & +\dots & +a_{nn}(t)x_n(t) & +b_n(t) \end{cases},$$

which can also be written as $DX=\dfrac{d}{dt}X=\dot{X}=AX+B$,

where

$$X = X(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix}, \qquad B = \begin{pmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_n(t) \end{pmatrix},$$

and

$$A = A(t) = \begin{pmatrix} a_{11}(t) & a_{12}(t) & \dots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \dots & a_{2n}(t) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(t) & a_{n2}(t) & \dots & a_{nn}(t) \end{pmatrix}.$$

If B and the a_{ij} are continuous on an interval $a < t < b$, and if $t_0 \in (a, b)$, then there is a unique solution to $\dot{X} = AX + B$ which satisfies $x_1(t) = \alpha_1, x_2(t) = \alpha_2, \dots, x_n(t) = \alpha_n$, valid on the interval (a, b).

The general solution to $\dot{X} = AX + B$ reads

$$\begin{aligned} X(t) &= U(t)C + X_p(t), \\ &= C_1 X_1(t) + C_2 X_2(t) + \dots + C_n X_n(t) + X_p(t), \end{aligned}$$

where

- $X_p(t)$ is a particular solution to $\dot{X} = AX + B$,
- $X_i, i = 1, \dots, n$, are n linearly independent solutions to the associated homogeneous system $\dot{X} = AX$,
- $C_i, i = 1, \dots, n$, are n constants.

The matrix $U(t) = [X_1(t), X_2(t), \dots, X_n(t)]$ is called a fundamental matrix of the homogeneous system $\dot{X} = AX$. Therefore, to solve the differential system, proceed as follows.

1. Solve the homogeneous system (i.e. find n linearly independent solutions $X_1, \dots, X_n$),
2. Find a particular solution to the full system.

How to solve the homogeneous system

The general solution $X(t)$ to the homogeneous system $\dot{X} = AX$ is a linear combination of n linearly independent solutions $X_1, \dots, X_n$, i.e. $X(t) = C_1 X_1(t) + \dots + C_n X_n(t)$ where the C_i are constants. In other words, it reads

$$X(t) = U(t)C$$

where $U(t)$ is a fundamental matrix and C a constant vector.

In the following, we consider systems with constant coefficients only, that is we assume the matrix A does not depend on time. The procedure is then as follows.

1. Find the eigenvalues and eigenvectors of A.
2. For each eigenvector ξ belonging to the eigenvalue λ, we know that $X(t) = \exp(\lambda t)\xi$ is a solution to $\dot{X} = AX$. Therefore,

    - If A has n distinct eigenvalues, we already know n linearly independent solutions to the homogeneous system, namely $X_i(t) = \exp(\lambda_i t)\xi_i, i = 1, \dots, n$.
    - If an eigenvalue λ has a multiplicity h higher than one, find as many linearly independent eigenvectors as you can, and write the corresponding solutions $X_j(t) = \exp(\lambda t)\xi_j$. Then, look for the missing solutions in the form $X_i = (t^{h-1}Y_{h-1} + t^{h-2}Y_{h-2} + \dots + tY_1 + Y_0)\exp(\lambda t)$, where the Y_i are constant vectors. Substitute this solution in the homogeneous system and solve for the Y_i. Always make sure that you have found n linearly independent solutions to the homogeneous system, and write the corresponding fundamental matrix.

Note: If one eigenvalue is complex, and since (in general) A has real coefficients, its complex conjugate is also an eigenvalue. In the same way, if ξ is a (complex) eigenvector that belongs to the complex eigenvalue λ, then its complex conjugate ξ^* belongs to the eigenvalue λ^*. Thus, if one eigenvalue $\lambda = \alpha + i\beta$ is complex, you may use the real functions $X_i(t) = \Re e[\exp(\alpha + i\beta)\xi]$ and $X_{i+1}(t) = \Im m[\exp(\alpha + i\beta)\xi]$ instead of the complex solutions $X_i(t) = \exp(\lambda t)\xi$ and $X_{i+1}(t) = X_i^*(t) = \exp(\lambda^* t)\xi^*$.

Example 1

A fundamental matrix for $\dot{X} = AX$ with $A = \begin{pmatrix} 1 & 3 \\ 2 & 2 \end{pmatrix}$ is $\begin{pmatrix} 3e^{-t} & e^{4t} \\ -2e^{-t} & e^{4t} \end{pmatrix}$.

Example 2

A fundamental matrix for $\dot{X} = AX$ with $A = \begin{pmatrix} 1 & 1 \\ -1 & 3 \end{pmatrix}$ is $\begin{pmatrix} e^{2t} & te^{2t} \\ e^{2t} & (t+1)e^{2t} \end{pmatrix}$.

How to find a particular solution to the full system

If $B(t) = b\exp(\omega t)$, where b is a constant vector, and if ω is not an eigenvalue of A, then try the particular solution $X_p(t) = k\exp(\omega t)$, where k is a constant vector. Substitute the expression of X_p back into the differential system $\dot{X} = AX + B$ and solve for k.

Note: It might be useful to use the principle of superposition. Indeed, since A has real coefficients, if X_p is a solution to $\dot{X} = AX + b\exp(\alpha t + i\beta t)$, (where b is a constant vector and is real), then $\Re e(X_p)$ and $\Im m(X_p)$ are respectively solutions to $\dot{X} = AX + b\exp(\alpha t)\cos(\beta t)$ and to $\dot{X} = AX + b\exp(\alpha t)\sin(\beta t)$. Thus, if $B(t)$ has the form $b\exp(\alpha t)\cos(\beta t)$ or $b\exp(\alpha t)\sin(\beta t)$, and if $\alpha + i\beta$ is not an eigenvalue of A, the above method can be used to find a particular solution to $\dot{X} = AX + b\exp(\alpha t + i\beta t)$. Its real or its imaginary part (depending on the form of $B(t)$) is then a particular solution to the original differential system.

If the above method cannot be applied, use the method of variation of constants: look for a solution in the form $X_p = U(t)V(t)$, where $U(t)$ is a fundamental matrix associated with the homogeneous system, and $V(t)$ is a vector to be determined. Then $V(t)$
satisfies $\dot{V} = U^{-1}(t)B(t)$, i.e. $V(t) = \int U^{-1}(t)B(t)dt$.

Example 1

A particular solution to $\dot{X} = AX + H$ with $A = \begin{pmatrix} -2 & 4 \\ 1 & 1 \end{pmatrix}$ and $H = \begin{pmatrix} 2 \\ 0 \end{pmatrix} e^t$ is $X_p = \begin{pmatrix} 0 \\ -1/2 \end{pmatrix} e^t$.

Example 2

A particular solution to $\dot{X} = AX + H$ with $A = \begin{pmatrix} 2 & 1 \\ -3 & -2 \end{pmatrix}$ and $H = \begin{pmatrix} 2 \\ 4 \end{pmatrix} e^t$ is $X_p = \begin{pmatrix} 5t - \frac{3}{2} \\ -5t + \frac{9}{2} \end{pmatrix} e^t$.

General remark: Sometimes, the elimination method (which consists in making successive operations on the rows of the linear system) may be easier to apply, or may simply go faster. It may also be useful if the matrix A depends on time.

Phase plane analysis

We now apply the information presented above to the linear stability analysis of fixed points of two-dimensional dynamical systems. Suppose that the nonlinear system of differential equations

$$\frac{dY}{dt} = F(Y), \qquad (13.4)$$

where $Y \in \mathbb{R}^2$ has a fixed point at $Y = Y_0$. The linearization of the above system near $Y = Y_0$ is obtained by setting $Y = Y_0 + X$, where $||X||$ is small, substituting this expression into Equation (13.4), and keeping only the terms that are linear in X. One thus writes

$$\frac{dX}{dt} = \frac{dY}{dt} = F(Y_0 + X) = F(Y_0) + DF(Y_0)\,X + \mathcal{O}(X^2) \simeq DF(Y_0)X,$$

where $DF(Y_0)$ is the Jacobian of F at $Y = Y_0$, $\mathcal{O}(X^2)$ represents nonlinear terms, and we used the fact that $F(Y_0) = 0$ since Y_0 is a fixed point of (13.4). The Jacobian of F at $Y = Y_0$ is defined as

$$DF(Y_0) = \begin{pmatrix} \left.\frac{\partial F_1}{\partial x}\right|_{(x_0,y_0)} & \left.\frac{\partial F_1}{\partial y}\right|_{(x_0,y_0)} \\ \left.\frac{\partial F_2}{\partial x}\right|_{(x_0,y_0)} & \left.\frac{\partial F_2}{\partial y}\right|_{(x_0,y_0)} \end{pmatrix},$$

where we used the following notation

$$Y = \begin{pmatrix} x \\ y \end{pmatrix}, \qquad F(X) = \begin{pmatrix} F_1(x,y) \\ F_2(x,y) \end{pmatrix}, \qquad Y_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}.$$

The dynamics of the linear system

$$\frac{dX}{dt} = DF(Y_0)\,X \qquad (13.5)$$

depends on the eigenvalues and eigenvectors of the matrix $A \equiv DF(Y_0)$. Indeed, we know that if A has distinct eigenvalues, the general solution of the linear system is of the form

$$X = C_1 \exp(\lambda_1 t)\xi_1 + C_2 \exp(\lambda_2 t)\xi_2, \qquad (13.6)$$

where ξ_1 and ξ_2 are eigenvectors of A associated with eigenvalues λ_1 and λ_2, and where C_1 and C_2 are arbitrary constants.

The fixed point $Y = Y_0$ of (13.4) is *linearly stable* if all solutions of the linearized system (13.5) converge towards the origin as $t \to +\infty$. This only occurs if the real parts of λ_1 and λ_2 are both negative, which implies that the trace of A is also negative. If we assume that F is real, as we do from now on, then A has real coefficients and its eigenvalues are either both real or complex conjugate of one another. In this case, a linearly stable fixed point is such that both $\mathrm{Tr}(A) = \lambda_1 + \lambda_2 < 0$ and $\det(A) = \lambda_1\lambda_2 > 0$.

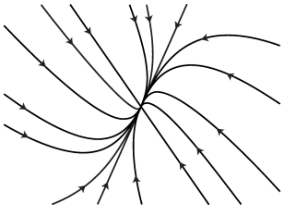


Figure 13.1

If the eigenvalues of A are both negative, real and distinct, then the fixed point is a *stable node*. The phase portrait of the linear system (13.5) looks like the sketch of Figure 13.1. Note that the trajectories are tangent at the origin to the eigendirection associated with the slowest eigenvalue.

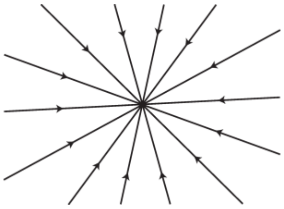


Figure 13.2

If A is proportional to the identity matrix (i.e. if the eigenvalues of A are the same but the corresponding eigenspace is two-dimensional), then the fixed point $Y = Y_0$ is called a *star*.

If the eigenvalues of A are the same but the corresponding eigenspace is one-dimensional, then $Y = Y_0$ is called a *degenerate node*.

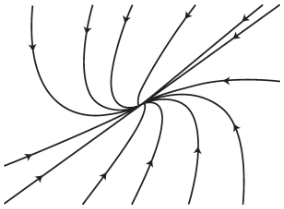


Figure 13.3

Figures 13.2 and 13.3 show the typical phase portrait of the linear system (13.5) when the fixed point is a stable star and a stable degenerate node, respectively.

If the eigenvalues of A have non-zero imaginary parts, then $Y = Y_0$ is a stable spiral. The corresponding phase portrait of (13.5) is shown in Figure 13.4.

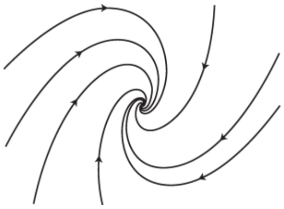


Figure 13.4

If one of the eigenvalues of A has positive real part, then the fixed point $Y = Y_0$ is linearly unstable. This occurs if either $\det(A) < 0$, in which case the fixed point is a *saddle point*, or if $\det(A) > 0$ and $\mathrm{Tr}(A) > 0$.

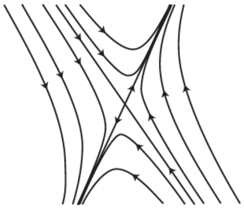


Figure 13.5

The phase portrait of (13.5) when the origin is a saddle point is shown in Figure 13.5. The phase portraits of (13.5) when the origin is an unstable node, an unstable star, an unstable degenerate node, or unstable spiral are similar to those of Figures 13.1, 13.2, 13.3, and 13.4, but with the direction of the arrows reversed.

If $\det(A) = 0$ and $\mathrm{Tr}(A) \neq 0$, then one of the eigenvalues of A is zero, and system (13.5) has a line of fixed points.

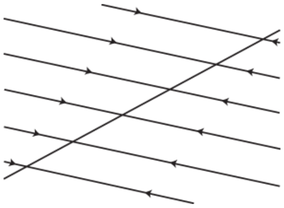


Figure 13.6

This is a degenerate situation and in most cases the dynamical system (13.4) has a single fixed point, even though its linearization has a continuous family of fixed points. The phase portrait of (13.5) in a case where $\lambda_1 < 0$ and $\lambda_2 = 0$ is shown in Figure 13.6.

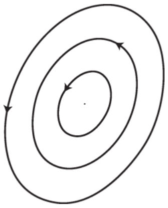


Figure 13.7

Finally, if all of the eigenvalues of A have zero real part, then $Y = Y_0$ is *marginally stable*. Since we ignore the non-generic situation where $A = 0$, this implies that A has purely imaginary complex conjugate eigenvalues, and the fixed point $Y = Y_0$ is called a *linear center*. In this case, $\det(A) > 0$ but $\mathrm{Tr}(A) = 0$. Figure 13.7 shows the phase portrait of the linear system (13.5) when the origin is a (linear) center. Determining whether the fixed point $Y = Y_0$ is a nonlinear center for the dynamical system (13.4) requires further analysis.

The above results can be summarized in the diagram of Figure 13.8, which shows the classification of the fixed point $X = 0$ of the linear system (13.5), as a function of the trace and determinant of A.

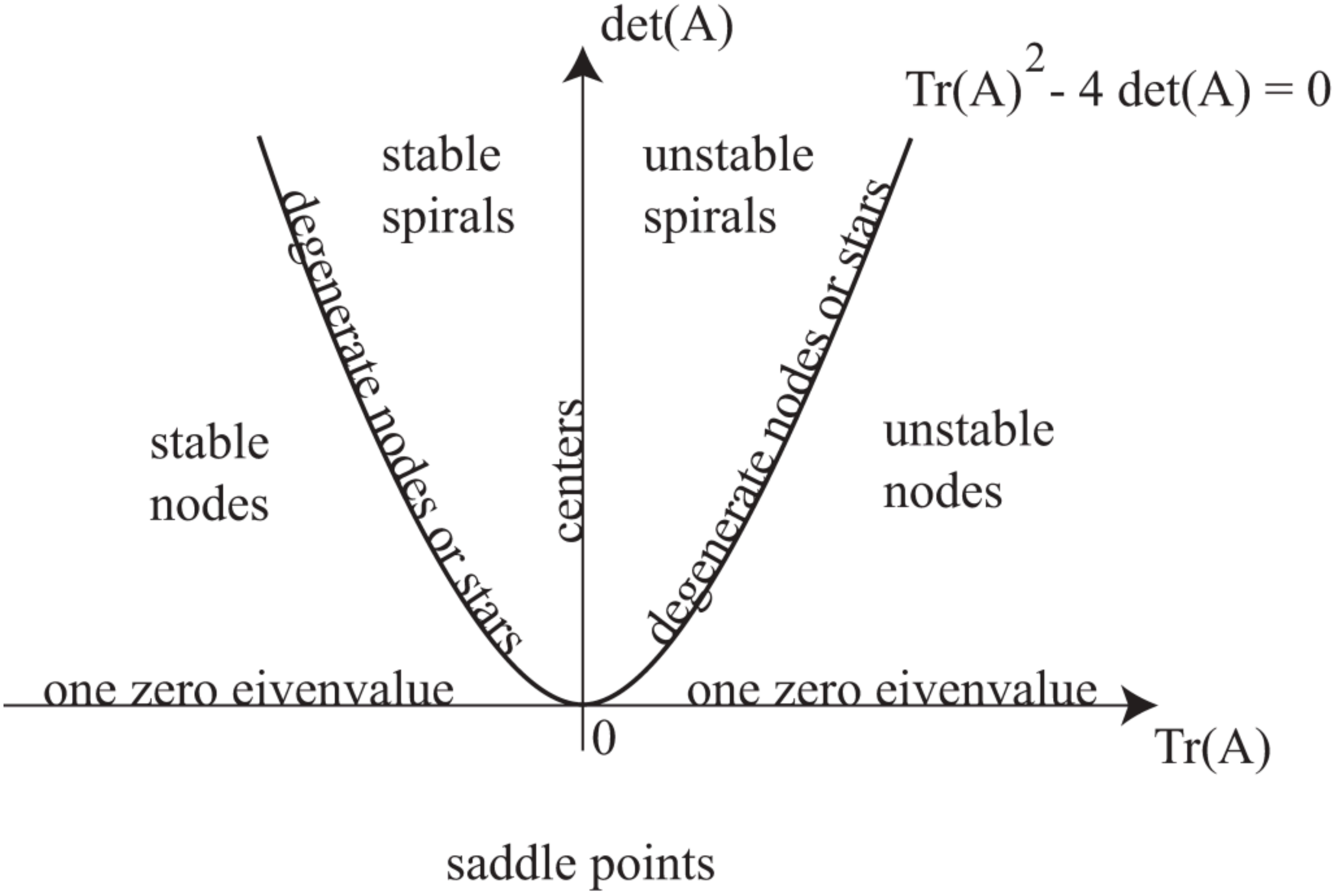


Figure 13.8. Classification of the fixed point $X = 0$ of the linear system (13.5), as a function of the trace and determinant of its Jacobian A.

The Hartman-Grobman theorem indicates that for a smooth dynamical system, the nature of a fixed point is not changed by nonlinear terms provided the fixed point is hyperbolic, i.e. provided the eigenvalues of the Jacobian of its linearization all have non-zero real parts. Linear stability analysis can therefore be used to classify hyperbolic fixed points of nonlinear dynamical systems. Further analysis is required to determine the nonlinear stability of non-hyperbolic fixed points. In particular, centers may remain centers or become stable or unstable spirals. Methods to investigate the effect of nonlinearities on hyperbolic and non-hyperbolic fixed points are typically discussed in an introductory text on dynamical systems.

Food for thought

Problem 1

What is the type (linear, separable, exact, solvable after a change of variable, Riccati's, Bernoulli's, Clairaut's) of the following first order equations? (Do not try to solve the equations).

1. $(1-y^3)\sinh(3x)y' + 7\cosh(y)\exp(x) = 0.$
2. $(4y + 2x + 4xy + x^2 + 3y^2)dx + (4x + 6y)dy = 0.$
3. $\left(x^2 \ln\left(\frac{x}{y}\right)\right) dx + \frac{1}{x}\left(x^3 + 3yx^2 + \frac{y^4}{x}\right) dy = 0.$
4. $x^3 \sinh(x)y' + 5x^2 y - \frac{3}{x}\cos(2x) = \sin(x).$
5. $(2x^2 y - 2y + 1)dx - xdy = 0.$
6. $\left(\cos(x)\sin(y) - x\sin(x)\sin(y) + y^2\right) dx + (x\cos(x)\cos(y) + 2xy)\, dy = 0.$
7. $yy' + \cosh(x)y^2 = \exp(x).$

Problem 2

Solve equations (5) and (6) above.

Problem 3

1. Show that $\frac{1}{3}\ln\left[(y-1)^2|y+2|\right] = \exp(-x) - x + C$ is an implicit solution to $\exp(x)(1+y)\frac{dy}{dx} + (1+\exp(x))(y^2+y-2) = 0.$
2. Find a solution to the above differential equation which satisfies $y(0) = 3$. Does the

corresponding initial value problem have a unique solution ? Explain.

Problem 4

1. Find the general solution to $y'' - 3y' + 2y = 0$.
2. Solve the initial value problem consisting of the above ODE and the initial condition $y(0) = 0$ and $y'(0) = 1$.
3. Is there a solution to the boundary value problem $y(0) = 0$ and $y(1) = 0$? If so, what is it ?

Problem 5

Solve the differential equation

$$x^3 y^{(3)} - 2x^2 y'' + 8xy' - 8y = 7x + x^2 \cos(2\ln(x)) \qquad x > 0.$$

Problem 6

Solve $xy'' + y' = 0$.

Problem 7

Solve the differential equation

$$y'' + 6y' + 9y = \frac{1}{x}\exp(-3x) \qquad x \neq 0.$$

Problem 8

Solve the following system of differential equations

$$\begin{cases} \dfrac{dx}{dt} = 2x + y + 3e^t \\ \dfrac{dy}{dt} = -x + 2y - e^t \end{cases} \qquad x(0) = 0; y(0) = 1.$$

Problem 9

Solve the system $\dot{X} = AX + B$ where

$$A = \begin{pmatrix} 7 & 6 \\ 2 & 6 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} -70 \\ 35 \end{pmatrix} \exp(3t).$$

Problem 10

Solve the differential equation

$$\frac{d^3y}{dt^3} - 5\frac{d^2y}{dt^2} + 12\frac{dy}{dt} - 8y = \exp(3t)\cos(2t) + 7t\exp(t).$$

Problem 11

Solve $\left[5x^4y^3 + 3y\sin(x) + 3xy\cos(x)\right]dx + \left[3x^5y^2 + 3x\sin(x)\right]dy = 0$, with the initial condition $y(\pi) = 5$.

Answers

Problem 1

1. Separable.

2. e^x is an integrating factor.
3. Homogeneous of degree 2.
4. Linear.
5. Linear.
6. Exact.
7. Bernoulli with $n = -1$.

Problem 2

- (5): $y = -\frac{1}{2x^2} + \frac{C}{x^2}e^{x^2}$.
- (6): $K = x\cos(x)\sin(y) + xy^2$.

Problem 3

$$\frac{1}{3}\ln\left[(y-1)^2|y+2|\right] = e^{-x} - x + \frac{1}{3}\ln(20) - 1.$$

Problem 4

1. $y = C_1 e^{2x} + C_2 e^x$.
2. $y = e^{2x} - e^x$.
3. Yes, $y = 0$.

Problem 5

$$y = C_1 x + C_2 x^2 \cos(2\ln(x)) + C_3 x^2 \sin(2\ln(x)) + \frac{7x}{5}\ln(x)$$
$$- \frac{x^2}{10}\ln(x)\cos(2\ln(x)) + \frac{x^2}{20}\ln(x)\sin(2\ln(x)).$$

Problem 6

$$y = K_1 \ln(|x|) + K_2.$$

Problem 7

$$y = C_1 e^{-3x} + C_2 x e^{-3x} + x \ln(|x|) e^{-3x}.$$

Problem 8

$$\begin{cases} x = C_1 \cos(t) e^{2t} + C_2 \sin(t) e^{2t} - 2e^t \\ y = -C_1 \sin(t) e^{2t} + C_2 \cos(t) e^{2t} - e^t \end{cases}.$$

Problem 9

$$X = \begin{pmatrix} 3e^{3t} & 2e^{10t} \\ -2e^{3t} & e^{10t} \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} + \begin{pmatrix} -60t + \frac{10}{7} \\ 40t + \frac{5}{7} \end{pmatrix} e^{3t}.$$

Problem 10

$$y = C_1 e^t + C_2 e^{2t} \cos(2t) + C_3 e^{2t} \sin(2t) - \frac{3}{68} e^{3t} \cos(2t)$$
$$+ \frac{5}{68} e^{3t} \sin(2t) + \left(\frac{7}{10} t^2 + \frac{14}{25} t \right) e^t.$$

Problem 11

$$x^5 y^3 + 3xy \sin(x) = 125\pi^5.$$

14.

MODELING PROJECTS

This section lists modeling projects, each based on a recently published research article, that can be explored in conjunction with the material presented in this text. At the University of Arizona, students enrolled in the Mathematical Modeling course form teams of 4 or 5 individuals and work together on one such project for the entire semester. Each group has their own project. They aim to understand the modeling approach described in the article, reproduce its results, and present the main ideas and conclusions to the rest of the class. Students are encouraged to formulate modeling questions that extend the work discussed in each article and, time permitting, to address some of them.

1. Collective Intelligence

> Understand and quantify whether different cooperative strategies between group members lead to increased problem-solving performance.

Project Information

- **Article**: *Agent-based models of collective intelligence* by S.M. Reia, A.C. Amado, J.F. Fontanari, Physics of Life Reviews **31**, 320–331 (2019).
- **Relevant Course Sections**: The modeling process (Chapter 1). Agent-based models (Chapter 2).
- **Useful Advanced Knowledge**: Logical reasoning and critical thinking. Coding.

Project Expectations

- Synthesize the different models introduced in the article and contrast the problem-solving approaches they represent.
- Develop a code to reproduce the authors' results on the performance of each approach as a function

group size.

- Interpret the outcome of your exploration of each model, including the role of relevant parameters.
- Evaluate the hypotheses made in support of the overall modeling strategy and assess the validity of the conclusions reached by the authors.
- Interpret any additional explorations of the model and defend your conclusions.

2. A Model for the Early Spread of COVID-19

Explore a simple mechanistic description of the zoonotic transmission of SARS-CoV-2 at the Wuhan market, the subsequent spread to humans, and associated mitigation measures.

Project Information

- **Article**: *A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action* by Q. Lin *et al.*, Int. J. Infectious Diseases **93**, 211–216 (2020).
- **Relevant Course Topics**: The modeling process (Chapter 1). Compartmental models and epidemiology (Chapter 7).
- **Useful Advanced Knowledge**: Dynamical Systems. Numerical simulations of systems of ODEs.

Project Expectations

- Summarize the model hypotheses.
- Construct the compartmental model and analyze its behavior.
- Synthesize the modeling approach and explore different scenarios.
- Evaluate the contributions of the model in light of what was known about the disease in early 2020.
- Assess how to revise the model hypotheses given what is currently known about SARS-CoV-2 and its variants.

3. Non-pharmaceutical Interventions for the Mitigation of Disease Spread

Develop a compartmental model to assess the potential of non-pharmaceutical interventions to mitigate the spread of COVID-19 in Italy, before vaccines became available.

Project Information

- **Article**: *Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy* by G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo, and M. Colaneri, Nature Medicine **26**, 855–860 (2020).
- **Relevant Course Sections**: The modeling process (Chapter 1). Fixed points and their stability (Chapter 3). Epidemics (Chapter 7).
- **Useful Advanced Knowledge**: Dynamical Systems. Numerical simulations of systems of ODEs (requires some coding experience). Proofs.

Project Expectations

- Summarize the model hypotheses, construct the compartmental model, and analyze its behavior.
- Write a code to simulate the coupled ODEs, explore different mitigation scenarios, and reproduce the results of the article.
- Synthesize the modeling approach, and evaluate the contributions of the model in light of what was known about COVID-19 in early 2020.

4. Vaccination Campaigns, Non-pharmaceutical Interventions, and the Burden of COVID-19

Develop a compartmental model to assess the effect of non-pharmaceutical interventions and COVID-19 vaccination campaigns on the health-care system in Italy.

Project Information

- **Article**: *Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in Italy* by G. Giordano, M. Colaneri, A. Di Filippo, F. Blanchini, P. Bolzern, G. De Nicolao, P. Sacchi, P. Colaneri, and R. Bruno, Nature Medicine **27**, 993–998 (2021).
- **Relevant Course Sections**: The modeling process (Chapter 1). Epidemics (Chapter 7).
- **Useful Advanced Knowledge**: Dynamical Systems. Numerical simulations of systems of ODEs (requires some coding experience).

Project Expectations

- Summarize the model hypotheses, construct the compartmental model, and discuss the significance and importance of each term, especially those related to vaccination.
- Write a code to simulate the coupled ODEs, explore different vaccine rollout scenarios, and reproduce the results of the article.
- Quantify how vaccination campaigns and non-pharmaceutical strategies affect the health-care system and the number of disease-related deaths.
- Synthesize the modeling approach and evaluate the contributions of model.

5. Glucose–Insulin Dynamics

Develop an ODE model of diabetes and study its dynamics, including the analysis of disease management strategies.

Project Information

- **Article**: *Dynamics of a Glucose–Insulin Model* by M. Ma & J. Li, Journal of Biological Dynamics **16**, 733-745 (2022).
- **Relevant Course Sections**: The modeling process (Chapter 1). Phase plane analysis (Chapter 3). Chemical reactions (Chapter 8).
- **Useful Advanced Knowledge**: Dynamical Systems. Numerical simulations of systems of ODEs. Proofs. Michaelis-Menten kinetics.

Project Expectations

- Summarize the model hypotheses, compare current and previous approaches to model glucose-insulin dynamics discussed in the article, and explain how the current model is constructed.
- Develop a complete phase plane analysis of the model by finding all of its fixed points and analyzing their stability, as well as through numerical simulations (no coding necessary if you use the Phase Plane app).
- Assess various disease control strategies based on your interpretation of how parameters affect the dynamics of the system.
- Evaluate the contributions of the model and interpret any of its limitations.

6. Collective Behaviors in Crowds

Develop and agent-based models to explain how visual cues can lead to the emergence of collective behaviors in crowds.

Project Information

- **Article**: *The visual coupling between neighbours explains local interactions underlying human 'flocking'* by G.C. Dachner, T.D. Wirth, E. Richmond, and W.H. Warren, Proc. R. Soc. B **289**, 20212089 (2022).
- **Relevant Course Sections**: The modeling process (Chapter 1). Agent-based models (Chapter 2).
- **Useful Advanced Knowledge**: Coding. Advanced Applied Analysis.

Project Expectations

- Synthesize the motivations of the authors and judge (critique, argue against or in favor of) their modeling choices, based on your understanding of the problem and of the data discussed in the article.
- Develop a code to simulate the dynamics between agents and reproduce the "simulation experiments" described in the article.
- Explore the effect of different parameter choices and compile your results, including any limitations of the model.
- Decide how you might improve the model and explore some of these improvements.
- Appraise the modeling approach and its contributions.

7. Trade-off Between Model Complexity and Parameter Identification

Develop and compare two different compartmental models for the transmission of dengue, a vector-borne disease.

Project Information

- **Article**: *Comparing vector-host and SIR models for dengue transmission* by A. Pandey, A. Mubayi, J. Medlock, Mathematical Biosciences **246**, 252–259 (2013).
- **Relevant Course Sections**: The modeling process (Chapter 1). Epidemics (Chapter 7).
- **Useful Advanced Knowledge**: Dynamical Systems. Numerical simulations of systems of ODEs (requires some coding experience). Theory of Probability. Markov-Chain Monte-Carlo (MCMC) meth-

ods.

Project Expectations

- Describe how vector-borne diseases are transmitted, summarize the modeling hypotheses, and build the compartmental models introduced in the article.
- Discuss the significance of the different terms in each of the models, and justify their form in light of the hypotheses that were made.
- Develop a numerical simulation of each model and simulate trajectories for different initial conditions and parameter choices, including those identified in the article.
- Describe the MCMC method for parameter estimation and reflect on the trade-off between model complexity and difficulties associated with parameter identification.
- Conclude with your own reflection on what modelers should take into account during the model selection step of the modeling process.

8. Predator-Prey Models

Develop predator-prey models, analyze their behavior, and consider their relevance in ecology.

Project Information

- **Article**: *Asymptotic stability of a modified Lotka-Volterra model with small immigrations* by T. Tahara, M.K. Areja Gavina, T. Kawano, J.M. Tubay, J.F. Rabajante, H. Ito, S. Morita, G. Ichinose, T. Okabe, T. Togashi, K. Tainaka, A. Shimizu, T. Nagatani & J. Yoshimura, Scientific Reports **8**, 7029 (2018).
- **Context Article**: *Long-term cyclic persistence in an experimental predator–prey system* by B. Blasius, L. Rudolf, G. Weithoff, U. Gaedke & G.F. Fussmann, Nature **577**, 226-230 (2020). The context article describes recent experimental results on the persistence of cyclic dynamics a predator-prey system.
- **Relevant Course Sections**: The modeling process (Chapter 1). Fixed points and their stability (Chapter 3). Two-Species Models (Chapter 6).
- **Useful Advanced Knowledge**: Dynamical Systems. Advanced Applied Analysis. Proofs.

Project Expectations

- Describe the modeling approach and the goals of the study.
- Analyze the linear model and describe the dynamics of the classical Lotka-Volterra model (no coding necessary if you use the Phase Plane app).
- Examine and justify the different modifications proposed in the article.
- Implement the corresponding models and analyze the resulting dynamics: use phase plane analysis techniques (find the fixed points and study their stability) as well as numerical simulations (e.g. with the Phase Plane app).
- Compare the results to the discussion in Chapter 7 of this text.
- Decide whether exploring additional variations of the model is warranted.
- Synthesize the results and critically evaluate the contributions of the work, especially in light of the recent discoveries presented in the context article.

9. Predator-Prey Interactions when the Prey Fears the Predator

A predator-prey system that takes fear into account.

Project Information

- **Article**: *Fear factor in a prey–predator system in deterministic and stochastic environment* by J. Roy & S. Alam, Physica A **541**, 123359 (2020).
- **Relevant Course Topics**: The modeling process (Chapter 1). Stability analysis (Chapter 3). Competing species (Chapter 6).
- **Useful Advanced Knowledge**: Dynamical Systems. Numerical simulations of systems of ODEs. Proofs. Real analysis.

Project Expectations

- Synthesize the modeling approach.
- Construct the deterministic autonomous model.

- Find the fixed points and analyze their stability.
- Explore the dynamics of the deterministic model, both with and without seasonal forcing.
- Assess the contributions of the model.

10. Communication in Honeybee Swarms

Understand how honeybees create a dynamic network of scents that allows them to locate their queen from afar.

Project Information

- **Article**: *Flow-mediated olfactory communication in honeybee swarms* by D.M.T. Nguyen *et al.*, PNAS **118**, e2011916118 (2021).
- **Relevant Course Topics**: The modeling process (Chapter 1). Agent-based models (Chapter 2). Diffusion (Chapter 9).
- **Useful Advanced Knowledge**: Applied Mathematical Analysis. Partial Differential Equations. Coding. Stochastic Processes.
- **Additional Information**: Movies and a detailed description of the model are provided in the online supplementary materials.

Project Expectations

- Summarize the data-collection process and subsequent analysis.
- Explain how the agent-based model works.
- Reproduce some of its results.
- Synthesize the modeling approach.
- Appraise the contributions of model.

11. Whale Migration

How sea water temperature and krill density affect whale migration along the coast of California.

Project Information

- **Article**: *Disentangling the biotic and abiotic drivers of emergent migratory behavior using individual-based models* by S. Dodson *et al.*, Ecological Modelling **432**, 109225 (2020).
- **Relevant Course Topics**: The modeling process (Chapter 1). Agent-based models (Chapter 2). Diffusion (Chapter 9).
- **Useful Advanced Knowledge**: Applied Mathematical Analysis. Theory of Probability. Coding. Stochastic Processes.

Project Expectations

- Synthesize the modeling approach.
- Construct approximate seasonal maps of sea water temperature and krill density.
- Build and run the individual-based model.
- Appraise the contributions of the model in light of our current knowledge of climate change.

12. Melt Ponds in the Arctic

A simple model that reproduces the geometric and scaling properties of melt ponds in the Arctic.

Project Information

- **Article**: *Ising model for melt ponds on Arctic sea ice* by Y-P. Ma et al., New J. Phys. **21**, 063029 (2019).
- **Relevant Course Topics**: The modeling process (Chapter 1). Scalings and Energy Minimization Methods (Chapter 3). Agent-based models (Chapter 2).
- **Useful Advanced Knowledge**: Physics (phase transitions, the Ising model). Applied Mathematical Analysis. Coding.
- **Additional Information**: for helicopter images of sea ice, see *Aerial observations of the evolution of ice surface conditions during summer* by D.K. Perovich, W.B. Tucker III, K.A. Ligett, as well as the SHEBA Reconnaissance Imagery, Version 1 database.

Project Expectations

- Synthesize the modeling approach.
- Construct the model and explore its behavior.
- Reproduce the results of the article.
- Assess the contributions of the model.