Differential Equations

An Introduction for Engineers

by Matthew Charnley

August 12, 2022
(version 0.9)

Typeset in LaTeX.

The date is the main identifier of version. The major version / edition number is raised only if there have been substantial changes.

See https://sites.rutgers.edu/matthew.charnley for more information (including contact information).

Contents

Preface

Attributions

The main inspiration for this book, as well as the vast majority of the source material, is *Notes on Diffy Qs* by Jiří Lebl [JL]. The fact that the book is freely available and open-source provided the main motivation for creating this current text. It allowed this book to be put together in a timely manner to be useful. It significantly reduced the work needed to put together a free textbook that fit the course exactly.

Introduction to this Version

This text was originally designed for the Math 244 class at Rutgers University. This class is a first course in Differential Equations for Engineering majors. This class is taken immediately after Multivariable Calculus and does not assume any knowledge of linear algebra. Prior to the design of this book, the course used Boyce and DiPrima's *Elementary Differential Equations and Boundary Value Problems* [BD]. The course provided a very brief introduction to matrices in order to get to the information necessary to handle first order systems of differential equations. With the course being redesigned to include more linear algebra, I was pointed in the direction of Jiří Lebl's *Notes on Diffy Qs* [JL], which was meant to be a drop-in replacement for the Boyce and DiPrima text, and as of a more recent version of the text, contained an appendix on Linear Algebra.

In creating this book, I wanted to retain the style of *Notes on Diffy Qs* [JL] but shape the text into something that directly fit the course that we wanted to run. This included reorganizing some of the topics, extra contextualization of the concept of differential equations, sections devoted to modeling principles and how these equations can be derived, and guidance in using MATLAB to solve differential equations numerically. Specifically, the content added to this book is

- Appendix A that gives an introduction or review to coding in MATLAB, as well as references to sample MATLAB files that can be used to easily sketch slope fields and solution curves to differential equations.

- Section 1.10 on the accumulation equation and its use in mathematical models, and § 1.11 which contains a discussion of parameter estimation, with inspiration taken from SIMIODE.

- The work on the eigenvalue method was split into three sections to account for real, complex, and repeated eigenvalues.
- A discussion of the trace-determinant plane and applications to analysis of linear (and non-linear) systems was added in § 4.7.
- Appendix B on prerequisite material to be referred to when needed. Some of the material here was pulled from Stitz and Zeager's book *Precalculus* [SZ].
- Exercises were added at the end of most sections of the text.

Acknowledgements

I would like to acknowledge David Molnar, who initially referred me to the *Notes on Diffy Qs* text [JL], as well as the *Precalculus* text [SZ], and provided inspiration and motivation to work on designing this text. For feedback during the development of the text, I want to acknowledge David Herrera, Yi-Zhi Huang, and many others who have helped over the development and refinement of this text. Finally, I want to acknowledge the Rutgers Open and Affordable Textbook Program for supporting the development and implementation of this text.

Introduction to *Notes on Diffy Qs*

This book [JL] originated from my class notes for Math 286 at the University of Illinois at Urbana-Champaign (UIUC) in Fall 2008 and Spring 2009. It is a first course on differential equations for engineers. Using this book, I also taught Math 285 at UIUC, Math 20D at University of California, San Diego (UCSD), and Math 4233 at Oklahoma State University (OSU). Normally these courses are taught with Edwards and Penney, *Differential Equations and Boundary Value Problems: Computing and Modeling* [EP], or Boyce and DiPrima's *Elementary Differential Equations and Boundary Value Problems* [BD], and this book aims to be more or less a drop-in replacement. Other books I used as sources of information and inspiration are E.L. Ince's classic (and inexpensive) *Ordinary Differential Equations* [I], Stanley Farlow's *Differential Equations and Their Applications* [F], now available from Dover, Berg and McGregor's *Elementary Partial Differential Equations* [BM], and William Trench's free book *Elementary Differential Equations with Boundary Value Problems* [T]. See the Further Reading chapter at the end of the book.

Computer resources

The book's website https://www.jirka.org/diffyqs/ contains the following resources:

1. Interactive SAGE demos.
2. Online WeBWorK homeworks (using either your own WeBWorK installation or Edfinity) for most sections, customized for this book.
3. The PDFs of the figures used in this book.

I taught the UIUC courses using IODE (https://faculty.math.illinois.edu/iode/). IODE is a free software package that works with Matlab (proprietary) or Octave (free software). The graphs in the book were made with the Genius software (see https://www.jirka.org/genius.html). I use Genius in class to show these (and other) graphs.

Acknowledgments

Firstly, I would like to acknowledge Rick Laugesen. I used his handwritten class notes the first time I taught Math 286. My organization of this book through chapter 5, and the choice of material covered, is heavily influenced by his notes. Many examples and computations are taken from his notes. I am also heavily indebted to Rick for all the advice he has given me, not just on teaching Math 286. For spotting errors and other suggestions, I would also like to acknowledge (in no particular order): John P. D'Angelo, Sean Raleigh, Jessica Robinson, Michael Angelini, Leonardo Gomes, Jeff Winegar, Ian Simon, Thomas Wicklund, Eliot Brenner, Sean Robinson, Jannett Susberry, Dana Al-Quadi, Cesar Alvarez, Cem Bagdatlioglu, Nathan Wong, Alison Shive, Shawn White, Wing Yip Ho, Joanne Shin, Gladys Cruz, Jonathan Gomez, Janelle Louie, Navid Froutan, Grace Victorine, Paul Pearson, Jared Teague, Ziad Adwan, Martin Weilandt, Sönmez Şahutoğlu, Pete Peterson, Thomas Gresham, Prentiss Hyde, Jai Welch, Simon Tse, Andrew Browning, James Choi, Dusty

Grundmeier, John Marriott, Jim Kruidenier, Barry Conrad, Wesley Snider, Colton Koop, Sarah Morse, Erik Boczko, Asif Shakeel, Chris Peterson, Nicholas Hu, Paul Seeburger, Jonathan McCormick, David Leep, William Meisel, Shishir Agrawal, Tom Wan, Andres Valloud, and probably others I have forgotten. Finally, I would like to acknowledge NSF grants DMS-0900885 and DMS-1362337.

Chapter 0

Introduction

0.1 Introduction to differential equations

Attribution: [JL], §0.2.

Learning Objectives

After this section, you will be able to:

- Identify a differential equation and determine the order of a differential equation,
- Verify that a function is a solution to a differential equation, and
- Solve some fundamental differential equations.

0.1.1 Differential equations

Consider the following situation:

> An object falling through the air has its velocity affected by two factors: gravity and a drag force. The velocity downward is increased at a rate of 9.8 m/s^2 due to gravity, and it is decreased by a rate equation to 0.3 times the current velocity of the object. If the ball is initially thrown downwards at a speed of 2 m/s, what will the velocity be 10 seconds later?

There might be enough information here to determine the velocity at any later point in time (it turns out, there is) but the information given isn't really about the velocity. Rather, information is given about the rate of change of the velocity. We know that the velocity will be increased at a rate of $9.8m/s^2$ due to gravity. How can this be interpreted? The rate of change has been discussed previously way back in Calculus 1; this is the derivative. Thus, if we let the unknown function $v(t)$ represent the velocity of the object, the description above gives information about the derivative of this function for$v(t)$. Taking the two different factors (the increase and decrease of velocity) into account, we can write an expression for this derivative, giving that

$$\frac{dv}{dt} = 9.8 - 0.3v.$$

Even though we don't know what $v(t)$ is, we know that it must affect the derivative of the velocity in this particular way, so we can write this equation. That's why we give a name to this function, so that we can use it in writing this equestion, which, since it is an equation involving the derivative of an unknown function $v(t)$, we call this a differential equation. Our goal here would be to use this information, plus the fact that the velocity at time zero is $v(0) = 2$ m/s to find the value of $v(10)$, or, more generally, the function $v(t)$ for any t.

The laws of physics, beyond just that of simple velocity, are generally written down as differential equations. Therefore, all of science and engineering use differential equations to some degree. Understanding differential equations is essential to understanding almost anything you will study in your science and engineering classes. You can think of mathematics as the language of science, and differential equations are one of the most important parts of this language as far as science and engineering are concerned. As an analogy, suppose all your classes from now on were given half in Swahili and half in English. It would be important to first learn Swahili, or you would have a very tough time getting a good grade in your classes. Without it, you might be able to make sense of some of what is going on, but would definitely be missing an important part of the picture.

Definition 0.1.1

A *differential equation* is an equation that involves one or more derivatives of an unknown function. For a differential equation, the *order* of the differential equation is the highest order derivative that appears in the equation.

One example of a first order differential equation is

$$\frac{dx}{dt} + x = 2\cos t. \tag{1}$$

Here x is the *dependent variable* and t is the *independent variable*. Note that we can use any letter we want for the dependent and independent variables. This equation arises from Newton's law of cooling where the ambient temperature oscillates with time.

To make sure that everything is well-defined, we will assume that we can always write our differential equation with the highest order derivative written as a function of all lower derivatives and the independent variable. For the previous example, since we can write (1) as

$$\frac{dx}{dt} = 2\cos t - x$$

where the highest derivative x' is written as a function of t and x, we have a proper differential equation. On the other hand, something like

$$\left(\frac{dy}{dt}\right)^2 + y^2 = 1 \tag{2}$$

is not a proper differential equation because we can't solve for $\frac{dy}{dt}$. This expression could either be written as

$$\frac{dy}{dt} = \sqrt{1-y^2} \qquad \text{or} \qquad \frac{dy}{dt} = -\sqrt{1-y^2},$$

and while both of these are proper differential equations, the version in (2) is not.

For some equations, like $y' = y^2$, the independent variable is not explicitly stated. We could be looking for a function $y(t)$ or a function $y(x)$ (or y of any other variable) and without any other information, any of these is correct. Usually, there will be information in the problem statement to indicate that the independent variable is something like time, in which case everything should be written in terms of t. It is for this reason that Leibniz notation is preferred for derivatives; an equation like

$$\frac{dy}{dt} = y^2$$

is unambiguously looking for any answer $y(t)$.

Example 0.1.1: All of the below are differential equations

$$\frac{dy}{dt} = e^t y \qquad z'' + z^2 = t \sin z$$
$$\frac{d^4 f}{dx^4} - 3x\frac{d^2 f}{dx^2} = x \qquad y''' + (y'')^2 - 3y = t^4.$$

Note that any letter can be used for the unknown function and its dependent variable. From the context of the equations, we can see that the unknown functions we are looking for in these examples are $y(t)$, $z(t)$, $y(x)$, and $y(t)$ respectively. The order of these equations are 1, 2, 4, and 3 respectively.

0.1.2 Solutions of differential equations

Solving the differential equation means finding the function that, when we plug it into the differential equation, gives a true statement. For example, take (1) from the previous section. In this case, this means that we want to find a function of t, which we call x, such that when we plug x, t, and $\frac{dx}{dt}$ into (1), the equation holds; that is, the left hand side equals the right hand side. It is the same idea as it would be for a normal (algebraic) equation of just x and t. We claim that

$$x = x(t) = \cos t + \sin t$$

is a *solution*. How do we check? We simply plug x into equation (1)! First we need to compute $\frac{dx}{dt}$. We find that $\frac{dx}{dt} = -\sin t + \cos t$. Now let us compute the left-hand side of (1).

$$\frac{dx}{dt} + x = \underbrace{(-\sin t + \cos t)}_{\frac{dx}{dt}} + \underbrace{(\cos t + \sin t)}_{x} = 2\cos t.$$

Yay! We got precisely the right-hand side. But there is more! We claim $x = \cos t + \sin t + e^{-t}$ is also a solution. Let us try,

$$\frac{dx}{dt} = -\sin t + \cos t - e^{-t}.$$

We plug into the left-hand side of (1)

$$\frac{dx}{dt} + x = \underbrace{(-\sin t + \cos t - e^{-t})}_{\frac{dx}{dt}} + \underbrace{(\cos t + \sin t + e^{-t})}_{x} = 2\cos t.$$

And it works yet again!

So there can be many different solutions. For this equation all solutions can be written in the form

$$x = \cos t + \sin t + Ce^{-t},$$

for some constant C. Different constants C will give different solutions, so there are really infinitely many possible solutions. See Figure 1 for the graph of a few of these solutions. We do not yet know how to find this solution, but we will get to that in the next chapter.

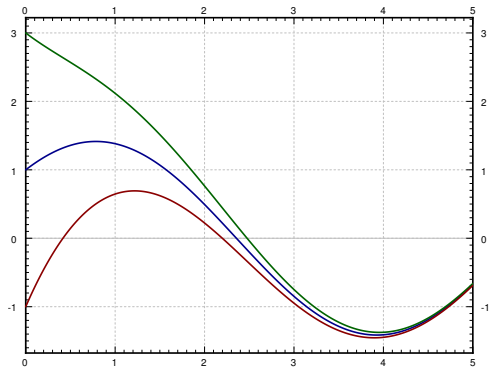


Figure 1: Few solutions of $\frac{dx}{dt} + x = 2\cos t$.

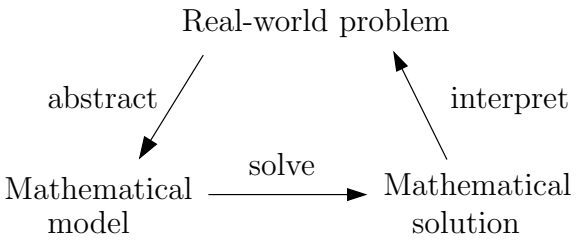
Solving differential equations can be quite hard. There is no general method that solves every differential equation. We will generally focus on how to get exact formulas for solutions of certain differential equations, but we will also spend a little bit of time on getting approximate solutions. And we will spend some time on understanding the equations without solving them.

Most of this book is dedicated to *ordinary differential equations* or ODEs, that is, equations with only one independent variable, where derivatives are only with respect to this one variable. If there are several independent variables, we get *partial differential equations* or PDEs.

Even for ODEs, which are very well understood, it is not a simple question of turning a crank to get answers. When you can find exact solutions, they are usually preferable to approximate solutions. It is important to understand how such solutions are found. Although in real applications you will leave much of the actual calculations to computers, you need to understand what they are doing. It is often necessary to simplify or transform your equations into something that a computer can understand and solve. You may even need to make certain assumptions and changes in your model to achieve this.

To be a successful engineer or scientist, you will be required to solve problems in your job that you have never seen before. It is important to learn problem solving techniques, so that you may apply those techniques to new problems. A common mistake is to expect to learn some prescription for solving all the problems you will encounter in your later career. This course is no exception.

0.1.3 Differential equations in practice



So how do we use differential equations in science and engineering? The main way this takes place is through the process of mathematical modeling. First, we have some *real-world problem* we wish to understand. We make some simplifying assumptions and create a *mathematical model*, which is a translation of this real-world problem into a set of differential equations. Think back to the example at the beginning of this section. We took a physical situation (a falling object) with some knowledge about how it behaves and turned that into a differential equation that describes the velocity over time. Then we apply mathematics to get some sort of a *mathematical solution*. Finally, we need to interpret our results, determining what this mathematical solution says about the real-world problem we started with. For instance, in the example at the start of the section, we could find the function $v(t)$, but then need to interpret that if we were to plug 10 into this function, we will get the velocity 10 seconds later.

Learning how to formulate the mathematical model and how to interpret the results is what your physics and engineering classes do. In this course, we will focus mostly on the mathematical analysis. This will be interspersed with discussions of this modeling process to give some context to what we are doing, and give practice for what will be seen in future physics and engineering classes.

Let us look at an example of this process. One of the most basic differential equations is the standard *exponential growth model*. Let P denote the population of some bacteria on a Petri dish. We assume that there is enough food and enough space. Then the rate of growth of bacteria is proportional to the population—a large population grows quicker. Let t denote time (say in seconds) and P the population. Our model is

$$\frac{dP}{dt} = kP,$$

for some positive constant $k > 0$.

Example 0.1.2: Suppose there are 100 bacteria at time 0 and 200 bacteria 10 seconds later. How many bacteria will there be 1 minute from time 0 (in 60 seconds)?

Solution: First we need to solve the equation. We claim that a solution is given by

$$P(t) = Ce^{kt},$$

where C is a constant. Let us try:

$$\frac{dP}{dt} = Cke^{kt} = kP.$$

And it really is a solution.

OK, now what? We do not know C, and we do not know k. But we know something. We know $P(0) = 100$, and we know

$P(10) = 200$. Let us plug these conditions in and see what happens.

$$100 = P(0) = Ce^{k0} = C,$$
$$200 = P(10) = 100\, e^{k10}.$$

Therefore, $2 = e^{10k}$ or $\frac{\ln 2}{10} = k \approx 0.069$. So

$$P(t) = 100\, e^{(\ln 2)t/10} \approx 100\, e^{0.069t}.$$

At one minute, $t = 60$, the population is $P(60) = 6400$. See Figure 2.

Let us talk about the interpretation of the results. Does our solution mean that there must be exactly 6400 bacteria on the plate at 60s? No! We made assumptions that might not be true exactly, just approximately. If our assumptions are reasonable, then there will be approximately 6400 bacteria. Also, in real life P is a discrete quantity, not a real number. However, our model has no problem saying that for example at 61 seconds, $P(61) \approx 6859.35$.

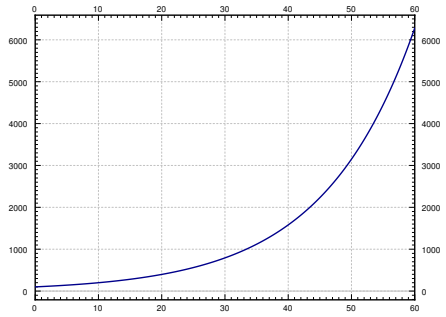


Figure 2: Bacteria growth in the first 60 seconds.

Normally, the k in $P' = kP$ is known, and we want to solve the equation for different *initial conditions*. What does that mean? Take $k = 1$ for simplicity. Suppose we want to solve the equation $\frac{dP}{dt} = P$ subject to $P(0) = 1000$ (the initial condition). Then the solution turns out to be (exercise)

$$P(t) = 1000\, e^t.$$

We call $P(t) = Ce^t$ *the general solution*, as every solution of the equation can be written in this form for some constant C. We need an initial condition to find out what C is, in order to find the *particular solution* we are looking for. Generally, when we say "particular solution," we just mean some solution.

0.1.4 Four fundamental equations

A few equations appear often and it is useful to know what their solutions are. Let us call them the four fundamental equations. Their solutions are reasonably easy to guess by recalling properties of exponentials, sines, and cosines. They are also simple to check, which is something that you should always do. No need to wonder if you remembered the solution correctly. It is good to have these as solutions that you "know" to build from when we learn solutions to other differential equations later on. In Chapter 1 we will cover the first two, and the last two will be discussed in Chapter 2.

First such equation is

$$\frac{dy}{dx} = ky,$$

for some constant $k > 0$. Here y is the dependent and x the independent variable. The general solution for this equation is

$$y(x) = Ce^{kx}.$$

We saw above that this function is a solution, although we used different variable names.

Next,

$$\frac{dy}{dx} = -ky,$$

for some constant $k > 0$. The general solution for this equation is

$$y(x) = Ce^{-kx}.$$

***Exercise* 0.1.1:** *Check that the y given is really a solution to the equation.*

Next, take the *second order differential equation*

$$\frac{d^2y}{dx^2} = -k^2y,$$

for some constant $k > 0$. The general solution for this equation is

$$y(x) = C_1 \cos(kx) + C_2 \sin(kx).$$

Since the equation is a second order differential equation, we have two constants in our general solution.

***Exercise* 0.1.2:** *Check that the y given is really a solution to the equation.*

Finally, consider the second order differential equation

$$\frac{d^2y}{dx^2} = k^2y,$$

for some constant $k > 0$. The general solution for this equation is

$$y(x) = C_1e^{kx} + C_2e^{-kx},$$

or

$$y(x) = D_1 \cosh(kx) + D_2 \sinh(kx).$$

For those that do not know, cosh and sinh are defined by

$$\cosh x = \frac{e^x + e^{-x}}{2}, \qquad \sinh x = \frac{e^x - e^{-x}}{2}.$$

They are called the *hyperbolic cosine* and *hyperbolic sine*. These functions are sometimes easier to work with than exponentials. They have some nice familiar properties such as $\cosh 0 = 1$, $\sinh 0 = 0$, and $\frac{d}{dx} \cosh x = \sinh x$ (no that is not a typo) and $\frac{d}{dx} \sinh x = \cosh x$.

***Exercise* 0.1.3:** *Check that both forms of the y given are really solutions to the equation.*

Example 0.1.3: In equations of higher order, you get more constants you must solve for to get a particular solution. The equation $\frac{d^2y}{dx^2} = 0$ has the general solution $y = C_1x + C_2$; simply integrate twice and don't forget about the constant of integration. Consider the initial conditions $y(0) = 2$ and $y'(0) = 3$. We plug in our general solution and solve for the constants:

$$2 = y(0) = C_1 \cdot 0 + C_2 = C_2, \qquad 3 = y'(0) = C_1.$$

In other words, $y = 3x + 2$ is the particular solution we seek.

0.1.5 Exercises

*Note: Exercises marked with * have answers in the back of the book.*

***Exercise* 0.1.4:** *Show that $x = e^{4t}$ is a solution to $x''' - 12x'' + 48x' - 64x = 0$.*

Exercise* 0.1.5: *Show that $x = e^{-2t}$ is a solution to $x'' + 4x' + 4x = 0$.*

***Exercise* 0.1.6:** *Show that $x = e^t$ is not a solution to $x''' - 12x'' + 48x' - 64x = 0$.*

***Exercise* 0.1.7:** *Is $y = \sin t$ a solution to $\left(\frac{dy}{dt}\right)^2 = 1 - y^2$? Justify.*

Exercise* 0.1.8: *Is $y = x^2$ a solution to $x^2y'' - 2y = 0$? Justify.*

***Exercise* 0.1.9:** *Let $y'' + 2y' - 8y = 0$. Now try a solution of the form $y = e^{rx}$ for some (unknown) constant r. Is this a solution for some r? If so, find all such r.*

Exercise* 0.1.10: *Let $xy'' - y' = 0$. Try a solution of the form $y = x^r$. Is this a solution for some r? If so, find all such r.*

***Exercise* 0.1.11:** *Verify that $x = Ce^{-2t}$ is a solution to $x' = -2x$. Find C to solve for the initial condition $x(0) = 100$.*

***Exercise* 0.1.12:** *Verify that $x = C_1e^{-t} + C_2e^{2t}$ is a solution to $x'' - x' - 2x = 0$. Find C_1 and C_2 to solve for the initial conditions $x(0) = 10$ and $x'(0) = 0$.*

Exercise* 0.1.13: *Verify that $x = C_1e^t + C_2$ is a solution to $x'' - x' = 0$. Find C_1 and C_2 so that x satisfies $x(0) = 10$ and $x'(0) = 100$.*

***Exercise* 0.1.14:** *Find a solution to $(x')^2 + x^2 = 4$ using your knowledge of derivatives of functions that you know from basic calculus.*

Exercise* 0.1.15: *Solve $\frac{d\varphi}{ds} = 8\varphi$ and $\varphi(0) = -9$.*

***Exercise* 0.1.16:** *Solve:*

a) $\dfrac{dA}{dt} = -10A, \quad A(0) = 5$

b) $\dfrac{dH}{dx} = 3H, \quad H(0) = 1$

c) $\dfrac{d^2y}{dx^2} = 4y, \quad y(0) = 0, \quad y'(0) = 1$

d) $\dfrac{d^2x}{dy^2} = -9x, \quad x(0) = 1, \quad x'(0) = 0$

Exercise* 0.1.17: *Solve:*

a) $\dfrac{dx}{dt} = -4x, \quad x(0) = 9$

b) $\dfrac{d^2x}{dt^2} = -4x, \quad x(0) = 1, \quad x'(0) = 2$

c) $\dfrac{dp}{dq} = 3p, \quad p(0) = 4$

d) $\dfrac{d^2T}{dx^2} = 4T, \quad T(0) = 0, \quad T'(0) = 6$

***Exercise* 0.1.18:** *Is there a solution to $y' = y$, such that $y(0) = y(1)$?*

***Exercise* 0.1.19:** *The population of city X was 100 thousand 20 years ago, and the population of city X was 120 thousand 10 years ago. Assuming constant growth, you can use the exponential population model (like for the bacteria). What do you estimate the population is now?*

***Exercise* 0.1.20:** *Suppose that a football coach gets a salary of one million dollars now, and a raise of 10% every year (so exponential model, like population of bacteria). Let s be the salary in millions of dollars, and t is time in years.*

a) *What is $s(0)$ and $s(1)$.*

b) *Approximately how many years will it take for the salary to be 10 million.*

c) *Approximately how many years will it take for the salary to be 20 million.*

d) *Approximately how many years will it take for the salary to be 30 million.*

0.2 Classification of differential equations

Attribution: [JL], §0.3.

Learning Objectives

After this section, you will be able to:

- Classify equation as ordinary or partial differential equations,
- Identify whether an equation is linear or non-linear, and
- Classify linear equations as homogenoeus, non-homogenoeus, or constant coefficient, as appropriate.

There are many types of differential equations, and we classify them into different categories based on their properties. Let us quickly go over the most basic classification. We already saw the distinction between ordinary and partial differential equations:

Definition 0.2.1

- *Ordinary differential equations* or (ODE) are equations where the derivatives are taken with respect to only one variable. That is, there is only one independent variable.
- *Partial differential equations* or (PDE) are equations that depend on partial derivatives of several variables. That is, there are several independent variables.

Let us see some examples of ordinary differential equations:

$$\frac{dy}{dt} = ky, \qquad \text{(Exponential growth)}$$

$$\frac{dy}{dt} = k(A - y), \qquad \text{(Newton's law of cooling)}$$

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = f(t). \qquad \text{(Mechanical vibrations)}$$

And of partial differential equations:

$$\frac{\partial y}{\partial t} + c\frac{\partial y}{\partial x} = 0, \qquad \text{(Transport equation)}$$

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad \text{(Heat equation)}$$

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}. \qquad \text{(Wave equation in 2 dimensions)}$$

If there are several equations working together, we have a so-called *system of differential equations*. For example,

$$y' = x, \qquad x' = y$$

is a simple system of ordinary differential equations. Maxwell's equations for electromagnetics,

$$\nabla \cdot \vec{D} = \rho, \qquad \nabla \cdot \vec{B} = 0,$$
$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}, \qquad \nabla \times \vec{H} = \vec{J} + \frac{\partial \vec{D}}{\partial t},$$

are a system of partial differential equations. The divergence operator $\nabla\cdot$ and the curl operator $\nabla\times$ can be written out in partial derivatives of the functions involved in the x, y, and z variables.

In the first chapter, we will start attacking first order ordinary differential equations, that is, equations of the form $\frac{dy}{dx} = f(x, y)$. In general, lower order equations are easier to work with and have simpler behavior, which is why we start with them.

We also distinguish how the dependent variables appear in the equation (or system).

Definition 0.2.2

We say an equation is *linear* if the dependent variable (or variables) and their derivatives appear linearly, that is only as first powers, they are not multiplied together, and no other functions of the dependent variables appear. Otherwise, the equation is called *nonlinear*.

Another way to determine if a differential equation is linear is if the equation is a sum of terms, where each term is some function of the independent variables or some function of the independent variables multiplied by a dependent variable or its derivative. That is, an ordinary differential equation is linear if it can be put into the form

$$a_n(x)\frac{d^n y}{dx^n} + a_{n-1}(x)\frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_1(x)\frac{dy}{dx} + a_0(x)y = b(x). \tag{3}$$

The functions a_0, a_1, ..., a_n are called the *coefficients*. The equation is allowed to depend arbitrarily on the independent variable. So

$$e^x \frac{d^2 y}{dx^2} + \sin(x)\frac{dy}{dx} + x^2 y = \frac{1}{x} \tag{4}$$

is still a linear equation as y and its derivatives only appear linearly. The equation

$$\cos(x)\frac{d^2 y}{dx^2} - xy + \frac{e^x}{x} = 0$$

is also linear, even though it is not initially in the correct form. From this equation, we can move the last term over to the right-hand side as a $-\frac{e^x}{x}$, and then it is in the correct form, with the $\frac{dy}{dx}$ term missing (or has coefficient zero).

All the equations and systems above as examples are linear. It may not be immediately obvious for Maxwell's equations unless you write out the divergence and curl in terms of partial derivatives. Let us see some nonlinear equations. For example Burger's equation,

$$\frac{\partial y}{\partial t} + y\frac{\partial y}{\partial x} = \nu \frac{\partial^2 y}{\partial x^2},$$

is a nonlinear second order partial differential equation. It is nonlinear because y and $\frac{\partial y}{\partial x}$ are multiplied together. The equation

$$\frac{dx}{dt} = x^2 \tag{5}$$

is a nonlinear first order differential equation as there is a second power of the dependent variable x.

Definition 0.2.3

A linear equation may further be called *homogeneous* if all terms depend on the dependent variable. That is, if no term is a function of the independent variables alone. Otherwise, the equation is called *nonhomogeneous* or *inhomogeneous*.

For example, the exponential growth equation, the wave equation, or the transport equation above are homogeneous. The mechanical vibrations equation above is nonhomogeneous as long as $f(t)$ is not the zero function. Similarly, if the ambient temperature A is nonzero, Newton's law of cooling is nonhomogeneous. A homogeneous linear ODE can be put into the form

$$a_n(x)\frac{d^n y}{dx^n} + a_{n-1}(x)\frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_1(x)\frac{dy}{dx} + a_0(x)y = 0.$$

Compare to (3) and notice there is no function $b(x)$.

If the coefficients of a linear equation are actually constant functions, then the equation is said to have *constant coefficients*. The coefficients are the functions multiplying the dependent variable(s) or one of its derivatives, not the function $b(x)$ standing alone. A constant coefficient nonhomogeneous ODE is an equation of the form

$$a_n\frac{d^n y}{dx^n} + a_{n-1}\frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_1\frac{dy}{dx} + a_0 y = b(x),$$

where $a_0, a_1, \ldots, a_n$ are all constants, but b may depend on the independent variable x. The mechanical vibrations equation above is a constant coefficient nonhomogeneous second order ODE. The same nomenclature applies to PDEs, so the transport equation, heat equation and wave equation are all examples of constant coefficient linear PDEs.

Finally, an equation (or system) is called *autonomous* if the equation does not explicitly depend on the independent variable. For autonomous ordinary differential equations, the independent variable is then thought of as time. Autonomous equation means an equation that does not change with time. For example, Newton's law of cooling is autonomous, so is equation (5). On the other hand, mechanical vibrations or (4) are not autonomous.

0.2.1 Exercises

***Exercise* 0.2.1:** *Classify the following equations. Are they ODE or PDE? Is it an equation or a system? What is the order? Is it linear or nonlinear, and if it is linear, is it homogeneous, constant coefficient? If it is an ODE, is it autonomous?*

a) $\sin(t)\dfrac{d^2x}{dt^2} + \cos(t)x = t^2$

b) $\dfrac{\partial u}{\partial x} + 3\dfrac{\partial u}{\partial y} = xy$

c) $y'' + 3y + 5x = 0, \quad x'' + x - y = 0$

d) $\dfrac{\partial^2 u}{\partial t^2} + u\dfrac{\partial^2 u}{\partial s^2} = 0$

e) $x'' + tx^2 = t$

f) $\dfrac{d^4x}{dt^4} = 0$

Exercise* 0.2.2: *Classify the following equations. Are they ODE or PDE? Is it an equation or a system? What is the order? Is it linear or nonlinear, and if it is linear, is it homogeneous, constant coefficient? If it is an ODE, is it autonomous?*

a) $\dfrac{\partial^2 v}{\partial x^2} + 3\dfrac{\partial^2 v}{\partial y^2} = \sin(x)$

b) $\dfrac{dx}{dt} + \cos(t)x = t^2 + t + 1$

c) $\dfrac{d^7F}{dx^7} = 3F(x)$

d) $y'' + 8y' = 1$

e) $x'' + tyx' = 0, \quad y'' + txy = 0$

f) $\dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial s^2} + u^2$

***Exercise* 0.2.3:** *If $\vec{u} = (u_1, u_2, u_3)$ is a vector, we have the divergence $\nabla \cdot \vec{u} = \frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} + \frac{\partial u_3}{\partial z}$ and curl $\nabla \times \vec{u} = \left(\frac{\partial u_3}{\partial y} - \frac{\partial u_2}{\partial z}, \ \frac{\partial u_1}{\partial z} - \frac{\partial u_3}{\partial x}, \ \frac{\partial u_2}{\partial x} - \frac{\partial u_1}{\partial y}\right)$. Notice that curl of a vector is still a vector. Write out Maxwell's equations in terms of partial derivatives and classify the system.*

***Exercise* 0.2.4:** *Suppose F is a linear function, that is, $F(x, y) = ax + by$ for constants a and b. What is the classification of equations of the form $F(y', y) = 0$.*

***Exercise* 0.2.5:** *Write down an explicit example of a third order, linear, nonconstant coefficient, nonautonomous, nonhomogeneous system of two ODE such that every derivative that could appear, does appear.*

Exercise* 0.2.6: *Write down the general zeroth order linear ordinary differential equation. Write down the general solution.*

Exercise* 0.2.7: *For which k is $\frac{dx}{dt} + x^k = t^{k+2}$ linear. Hint: there are two answers.*

***Exercise* 0.2.8:** *Write out an explicit example of a non-homogeneous fourth order, linear, constant coefficient differential equation. where all possible derivatives of the unknown function y appear.*

Exercise* 0.2.9: *Let x, y, and z be three functions of t defined by the system of differential equations*

$$x' = y \quad y' = z \quad z' = 3x - 2y + 5z + e^t$$

with initial conditions $x(0) = 3$, $y(0) = -2$ and $z(0) = 1$, and let $u(t)$ be the function defined by the solution to

$$u''' - 5u'' + 2u' - 3u = e^t$$

with initial conditions $u(0) = 3$, $u'(0) = -2$, and $u''(0) = 1$.

a) *Use the substitution $u = x$, $u' = y$, and $u'' = z$ to verify that $x(t) = u(t)$ because they solve the same initial value problem.*

b) *What is the order of the system defining x, y, and z and how many components does it have?*

c) *What is the order of the equation defining u? How many components does that have?*

d) *How do these numbers relate to each other?*

Chapter 1

First Order Differential Equations

In this chapter, we begin by discussing first order differential equations. As they have the lowest possible order, only containing one derivative of the unknown function, they tend to be the simplest equations to try to analyze and solve. This doesn't mean that we'll be able to solve all of them, but we can make a decent effort at a fair number of them. These equations are also very common in modeling problems, as most principles from science and engineering give us a way to express the rate of change of a given quantity. This setup gives rise to a first order differential equation involving that quantity, which, if we can solve it, will tell us how the quantity evolves over time. Even if we can't solve the equation analytically, a numerical solution may allow us to predict the behavior of a system over time and design it to best fit our needs.

1.1 Integrals as solutions

Attribution: [JL], §1.1.

Learning Objectives

After this section, you will be able to:

- Solve a first order differential equation by direct integration and
- Understand the difference between a general solution and the solution to an initial value problem.

A first order ODE is an equation of the form

$$\frac{dy}{dx} = f(x, y),$$

or just

$$y' = f(x, y).$$

Some examples that fit this form are

$$y' = x^2 y - e^x \sin x$$

and

$$y' = e^y(x^2 + 1) - \cos(y).$$

Looking back at the last section, the first of these is linear and the second is not. In general, there is no simple formula or procedure one can follow to find solutions. In the next few sections we will look at special cases where solutions are not difficult to obtain. In this section, let us assume that f is a function of x alone, that is, the equation is

$$y' = f(x). \tag{1.1}$$

We could just integrate (antidifferentiate) both sides with respect to x.

$$\int y'(x)\,dx = \int f(x)\,dx + C,$$

that is

$$y(x) = \int f(x)\,dx + C.$$

This $y(x)$ is actually the general solution. So to solve (1.1), we find some antiderivative of $f(x)$ and then we add an arbitrary constant to get the general solution.

Now is a good time to discuss a point about calculus notation and terminology. One of the final keystone concepts in Calculus 1 is that of the fundamental theorem of calculus, which ties together two mathematical ideas: definite integrals (defined as the area under a curve) and indefinite integrals or antidifferentiation (undoing the operation of differentiation). This theorem says that these two ideas are in some sense the same; in order to compute a definite integral, one can first find an antiderivative and plug in the endpoints (the most common use of the theorem), and that the derivative of a definite integral gives back the function inside (something that will be useful in this course).

The main distinction between these two is the type of object that they are. Definite integrals evaluate to numbers, so they are functions, which means they are the object we want to deal with in this course. Indefinite integrals are families of functions, and while they have their uses (motivating the idea of a general solution), their main use is the process of antidifferentiation which leads us to solutions in the form of definite integrals. Provided that you can evaluate the antiderivative in question, these two processes will end up at exactly the same solution. If you end up confused about the terminology, the goal for this class is always a definite integral, but we can use antiderivatives to get there. Hence the terminology *to integrate* when we may really mean *to antidifferentiate*. Integration is just one way to compute the antiderivative (and it is a way that always works, see the following examples). Integration is defined as the area under the graph and it also happens to also compute antiderivatives. For sake of consistency, we will keep using the indefinite integral notation when we want an antiderivative, and you should *always* think of the definite integral as a way to write it.

Example 1.1.1: Find the general solution of $y' = 3x^2$.

Solution: Elementary calculus tells us that the general solution must be $y = x^3 + C$. Let us check by differentiating: $y' = 3x^2$. We got *precisely* our equation back. □

Normally, we will also have an initial condition such as $y(x_0) = y_0$ for some two numbers x_0 and y_0 (x_0 is often 0, but not always). If we do, the combination of a differential equation and an initial condition is called an initial value problem. We can then write the solution as a definite integral in a nice way. Suppose our problem is $y' = f(x)$, $y(x_0) = y_0$. Then the solution is

$$y(x) = \int_{x_0}^{x} f(s)\,ds + y_0. \tag{1.2}$$

Let us check! We compute

$$y'(x) = \frac{d}{dx}\left[\int_{x_0}^{x} f(s)\,ds + y_0\right].$$

Since y_0 is a constant, it's derivative is zero, and by the fundamental theorem of calculus

$$\frac{d}{dx}\int_{x_0}^{x} f(s)\;dx = f(x).$$

Therefore $y' = f(x)$, and by Jupiter, y is a solution. Is it the one satisfying the initial condition? Well,

$$y(x_0) = \int_{x_0}^{x_0} f(x)\,dx + y_0$$

and since f is a nice function, we know that the integral of f with matching endpoints is 0. Therefore $y(x_0) = y_0$. It is!

Do note that the definite integral and the indefinite integral (antidifferentiation) are completely different beasts. The definite integral always evaluates to a number. Therefore, (1.2) is a formula we can plug into the calculator or a computer, and it will be happy to calculate specific values for us. We will easily be able to plot the solution and work with it just like with any other function. It is not so crucial to always find a closed form for the antiderivative.

Example 1.1.2: Solve

$$y' = e^{-x^2}, \qquad y(0) = 1.$$

Solution: By the preceding discussion, the solution must be

$$y(x) = \int_{0}^{x} e^{-s^2}\,ds + 1.$$

Here is a good way to make fun of your friends taking second semester calculus. Tell them to find the closed form solution. Ha ha ha (bad math joke). It is not possible (in closed form). There is absolutely nothing wrong with writing the solution as a definite integral. This particular integral is in fact very important in statistics. ⌟

While there is nothing wrong with writing solutions as a definite integral, they should be simplified and evaluated if possible. Given the differential equation

$$y' = 3x^2, \qquad y(2) = 6,$$

the solution can be written as

$$y(x) = \int_2^x 3s^2 \, ds + 6.$$

However, it is much more convenient, both for human reasoning and computers, to write this solution as

$$y(x) = x^3 - 2.$$

So, if integrals can be evaluated and simplified to explicit functions, then they should be worked out. If it is not possible, then answers in integral form are completely fine.

Classical problems leading to differential equations solvable by integration are problems dealing with velocity, acceleration and distance. You have surely seen these problems before in your calculus class.

Example 1.1.3: Suppose a car drives at a speed $e^{t/2}$ meters per second, where t is time in seconds. How far did the car get in 2 seconds (starting at $t = 0$)? How far in 10 seconds?

Solution: Let x denote the distance the car traveled. The equation is

$$x' = e^{t/2}.$$

We just integrate this equation to get that

$$x(t) = 2e^{t/2} + C.$$

We still need to figure out C. We know that when $t = 0$, then $x = 0$. That is, $x(0) = 0$. So

$$0 = x(0) = 2e^{0/2} + C = 2 + C.$$

Thus $C = -2$ and

$$x(t) = 2e^{t/2} - 2.$$

Now we just plug in to get where the car is at 2 and at 10 seconds. We obtain

$$x(2) = 2e^{2/2} - 2 \approx 3.44 \text{ meters}, \qquad x(10) = 2e^{10/2} - 2 \approx 294 \text{ meters}.$$

Example 1.1.4: Suppose that the car accelerates at a rate of t^2 m/s^2. At time $t = 0$ the car is at the 1 meter mark and is traveling at 10 m/s. Where is the car at time $t = 10$?

Solution: Well this is actually a second order problem. If x is the distance traveled, then x' is the velocity, and x'' is the acceleration. The initial value problem for this situation is

$$x'' = t^2, \qquad x(0) = 1, \qquad x'(0) = 10.$$

What if we say $x' = v$. Then we have the problem

$$v' = t^2, \qquad v(0) = 10.$$

Once we solve for v, we can integrate and find x.

***Exercise* 1.1.1:** *Solve for v, and then solve for x. Find $x(10)$ to answer the question.*

1.1.1 Exercises

***Exercise* 1.1.2:** *Solve* $\frac{dy}{dx} = x^2 + x$ *with* $y(1) = 3$.

***Exercise* 1.1.3:** *Solve* $\frac{dy}{dx} = \sin(5x)$ *with* $y(0) = 2$.

Exercise* 1.1.4: *Solve* $\frac{dy}{dx} = e^x + x$ *with* $y(0) = 10$.

***Exercise* 1.1.5:** *Solve* $\frac{dy}{dx} = 2xe^{3x}$ *with* $y(0) = 1$.

***Exercise* 1.1.6:** *Solve* $\frac{dx}{dt} = e^t \cos(2t) + t$ *with* $y(0) = 3$.

***Exercise* 1.1.7:** *Solve* $\frac{dy}{dx} = \frac{1}{x^2+1} + 3e^{2x}$ *with* $y(0) = 2$.

***Exercise* 1.1.8:** *Solve* $\frac{dy}{dx} = \frac{1}{x^2-1}$ *for* $y(0) = 0$. *(This requires partial fractions or hyperbolic trigonometric functions.)*

***Exercise* 1.1.9** (harder)**:** *Solve* $y'' = \sin x$ *for* $y(0) = 0$, $y'(0) = 2$.

***Exercise* 1.1.10:** *A spaceship is traveling at the speed* $2t^2 + 1$ *km/s (*t *is time in seconds). It is pointing directly away from earth and at time* $t = 0$ *it is 1000 kilometers from earth. How far from earth is it at one minute from time* $t = 0$*?*

Exercise* 1.1.11: *Sid is in a car traveling at speed* $10t + 70$ *miles per hour away from Las Vegas, where* t *is in hours. At* $t = 0$*, Sid is 10 miles away from Vegas. How far from Vegas is Sid 2 hours later?*

***Exercise* 1.1.12:** *Solve* $\frac{dx}{dt} = \sin(t^2) + t$, $x(0) = 20$. *It is OK to leave your answer as a definite integral.*

***Exercise* 1.1.13:** *Solve* $\frac{dy}{dt} = e^{t^2} + \sin(t)$, $y(0) = 4$. *The answer can be left as a definite integral.*

***Exercise* 1.1.14:** *A dropped ball accelerates downwards at a constant rate* 9.8 *meters per second squared. Set up the differential equation for the height above ground* h *in meters. Then supposing* $h(0) = 100$ *meters, how long does it take for the ball to hit the ground.*

Exercise* 1.1.15: *The rate of change of the volume of a snowball that is melting is proportional to the surface area of the snowball. Suppose the snowball is perfectly spherical. The volume (in centimeters cubed) of a ball of radius* r *centimeters is* $(4/3)\pi r^3$*. The surface area is* $4\pi r^2$*. Set up the differential equation for how the radius* r *is changing. Then, suppose that at time* $t = 0$ *minutes, the radius is 10 centimeters. After 5 minutes, the radius is 8 centimeters. At what time* t *will the snowball be completely melted?*

Exercise* 1.1.16: *Find the general solution to* $y'''' = 0$*. How many distinct constants do you need?*

1.2 Slope fields

Attribution: [JL], §1.2.

Learning Objectives

After this section, you will be able to:

- Identify or sketch a slope field for a first order differential equation and
- Use the slope field to determine the trajectory of a solution to a differential equation.

As we said, the general first order equation we are studying looks like

$$y' = f(x, y).$$

A lot of the time, we cannot simply solve these kinds of equations explicitly, because our direct integration method only works when the equation is of the form $y' = f(x)$, which we could integrate directly. In these more complicated cases, it would be nice if we could at least figure out the shape and behavior of the solutions, or find approximate solutions.

1.2.1 Slope fields

Suppose that we have a solution to the equation $y' = f(x, y)$ with $y(x_0) = y_0$. What does the fact that this solves the differential equation tell us about the solution? It tells us that the derivative of the solution at this point will be $f(x_0, y_0)$. Graphically, the derivative gives the slope of the solution, so it means that the solution will pass through the point (x_0, y_0) and will have slope $f(x_0, y_0)$. For example, if $f(x, y) = xy$, then at point $(2, 1.5)$ we draw a short line of slope $xy = 2 \times 1.5 = 3$. So, if $y(x)$ is a solution and $y(2) = 1.5$, then the equation mandates that $y'(2) = 3$. See Figure 1.1.

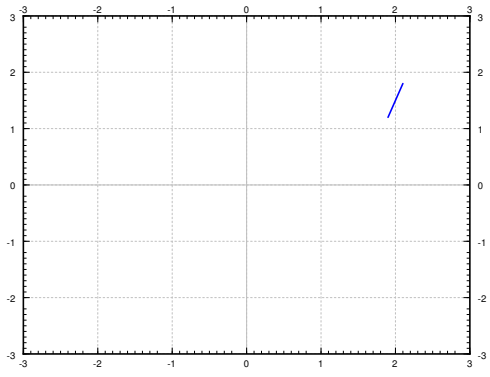


Figure 1.1: The slope $y' = xy$ at $(2, 1.5)$.

To get an idea of how solutions behave, we draw such lines at lots of points in the plane, not just the point $(2, 1.5)$. We would ideally want to see the slope at every point, but that is just not possible. Usually we pick a grid of points fine enough so that it shows the behavior, but not too fine so that we can still recognize the individual lines. We call this picture the *slope field* of the equation. See Figure 1.2 on the next page for the slope field of the equation $y' = xy$. Usually in practice, one does not do this by hand, but has a computer do the drawing.

The idea of a slope field is that it tells us how the graph of the solution should be sloped, or should curve, if it passed through a given point. Having a wide variety of slopes plotted in our slope field gives an idea of how all of the solutions behave for a bunch of different initial

conditions. Which curve we want in particular, and where we should start the curve, depends on the initial condition.

Suppose we are given a specific initial condition $y(x_0) = y_0$. A solution, that is, the graph of the solution, would be a curve that follows the slopes we drew, starting from the point (x_0, y_0). For a few sample solutions, see Figure 1.3. It is easy to roughly sketch (or at least imagine) possible solutions in the slope field, just from looking at the slope field itself. You simply sketch a line that roughly fits the little line segments and goes through your initial condition. The graph should "flow" along the little slopes that are on the slope field.

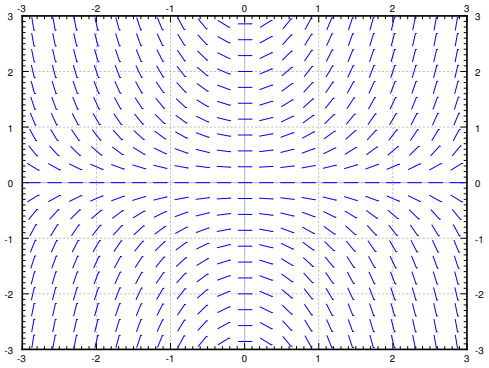


Figure 1.2: Slope field of $y' = xy$.

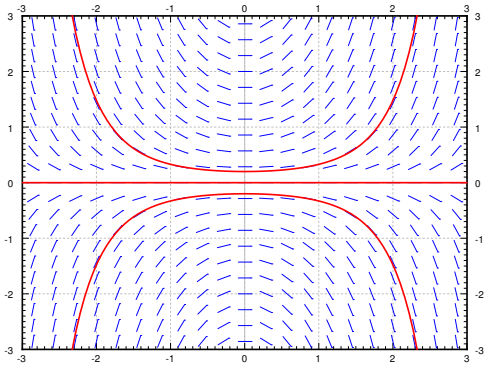


Figure 1.3: Slope field of $y' = xy$ with a graph of solutions satisfying $y(0) = 0.2$, $y(0) = 0$, and $y(0) = -0.2$.

By looking at the slope field we get a lot of information about the behavior of solutions without having to solve the equation. For example, in Figure 1.3 we see what the solutions do when the initial conditions are $y(0) > 0$, $y(0) = 0$ and $y(0) < 0$. A small change in the initial condition causes quite different behavior. We see this behavior just from the slope field and imagining what solutions ought to do.

We see a different behavior for the equation $y' = -y$. The slope field and a few solutions is in see Figure 1.4 on the following page. If we think of moving from left to right (perhaps x is time and time is usually increasing), then we see that no matter what $y(0)$ is, all solutions tend to zero as x tends to infinity. Again that behavior is clear from simply looking at the slope field itself.

1.2.2 Exercises

Exercise 1.2.1: *Sketch slope field for $y' = e^{x-y}$. How do the solutions behave as x grows? Can you guess a particular solution by looking at the slope field?*

Exercise 1.2.2:* Sketch the slope field of $y' = y^3$. Can you visually find the solution that satisfies $y(0) = 0$?*

Exercise 1.2.3: *Sketch slope field for $y' = x^2$.*

Exercise 1.2.4: *Sketch slope field for $y' = y^2$.*

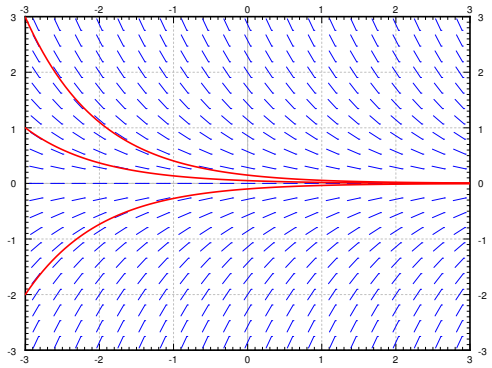
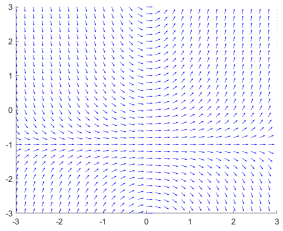
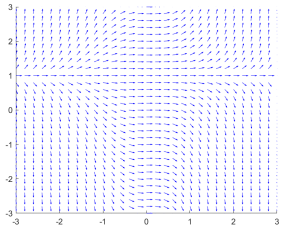
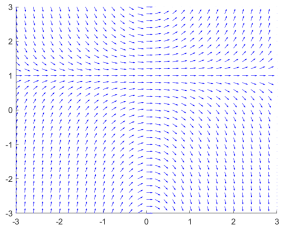


Figure 1.4: Slope field of $y' = -y$ with a graph of a few solutions.

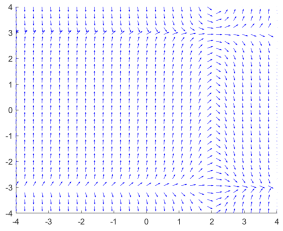
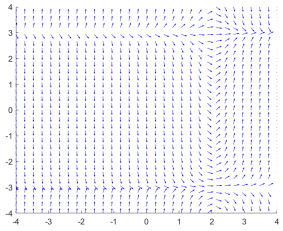
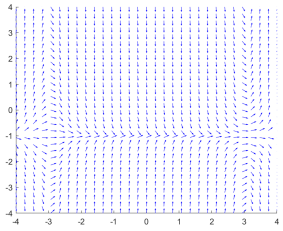
***Exercise* 1.2.5:** *For each of the following differential equations, sketch out a slope field on $-3 < x < 3$ and $-3 < y < 3$ and determine the overall behavior of the solutions to the equation as $t \to \infty$. If this fact depends on the value of the solution at $t = 0$, explain how it changes.*

a) $\dfrac{dy}{dx} = 3 - 2y$ b) $\dfrac{dy}{dx} = 1 + y$ c) $\dfrac{dy}{dx} = y - 1$ d) $\dfrac{dy}{dx} = -2 - y$

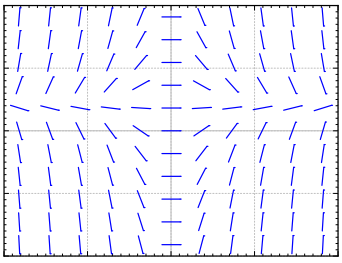
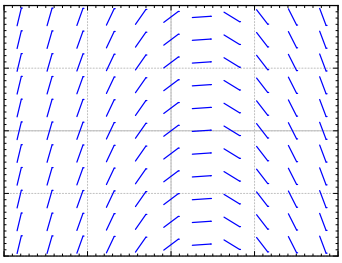
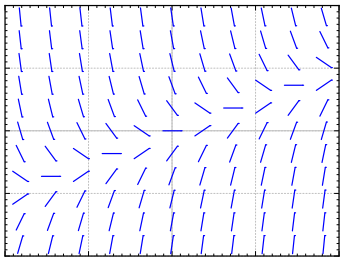
***Exercise* 1.2.6:** *Which of the following slope fields corresponds to the differential equation $\frac{dy}{dt} = t(y-1)$. Explain your reasoning.*

a)  b)  c) 

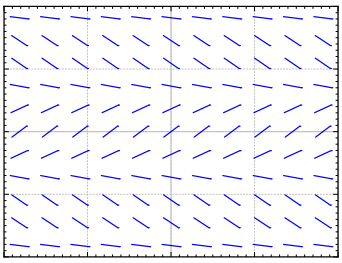
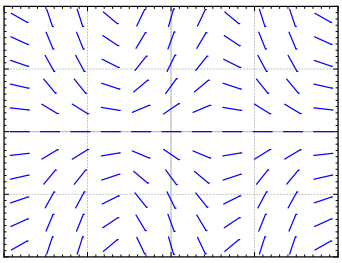
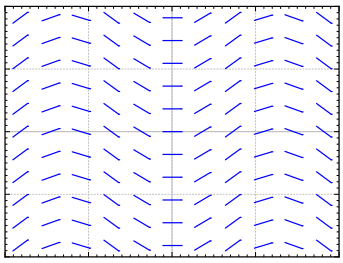
***Exercise* 1.2.7:** *Which of the following slope fields corresponds to the differential equation $\frac{dy}{dt} = (2-t)(y^2-9)$. Explain your reasoning.*

a)  b)  c) 

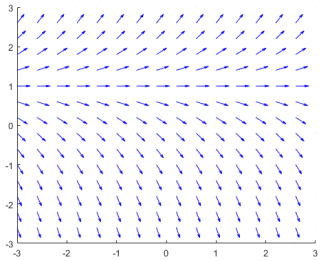
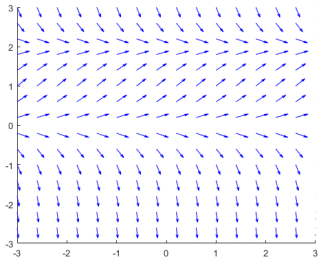
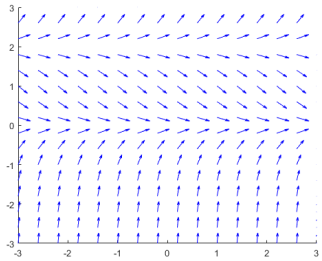
***Exercise* 1.2.8:** *Match equations $y' = 1 - x$, $y' = x - 2y$, $y' = x(1 - y)$ to slope fields. Justify.*

a)  b)  c) 

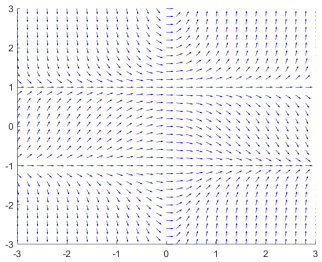
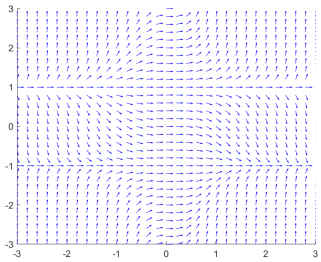
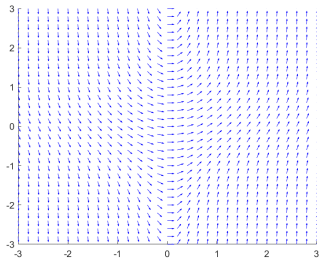
Exercise* 1.2.9: *Match equations $y' = \sin x$, $y' = \cos y$, $y' = y\cos(x)$ to slope fields. Justify.*

a)  b)  c) 

***Exercise* 1.2.10:** *Match equations $y' = y(y - 2)$, $y' = y - 1$, $y' = y(2 - y)$ to slope fields. Justify.*

a)  b)  c) 

***Exercise* 1.2.11:** *Match equations $y' = t(y^2 + 1)$, $y' = t(y^2 - 1)$, $y' = t^2(y^2 - 1)$ to slope fields. Justify.*

a)  b)  c) 

***Exercise* 1.2.12:** *The slope field for the differential equation $y' = (3 - y)(y + 2)$ is below. If we find the solution to this differential equation with initial condition, $y(0) = 1$, what will happen to the solution as $t \to \infty$? Use the slope field and your knowledge of the equation to determine the long-time behavior of this solution.*

***Exercise* 1.2.13:** *The slope field for the differential equation $y' = (t-2)(y+4)(y-3)$ is below. If we find the solution to this differential equation with initial condition, $y(0) = 1$, what will happen to the solution as $t \to \infty$? Use the slope field and your knowledge of the equation to determine the long-time behavior of this solution.*

***Exercise* 1.2.14:** *The slope field for the differential equation $y' = (y+1)(y+4)$ is below. If we find the solution to this differential equation with initial condition, $y(0) = 1$, what will happen to the solution as $t \to \infty$? Use the slope field and your knowledge of the equation to determine the long-time behavior of this solution.*

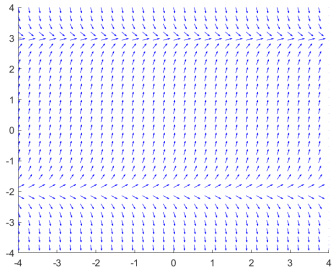


Figure 1.5: Exercise 1.2.12

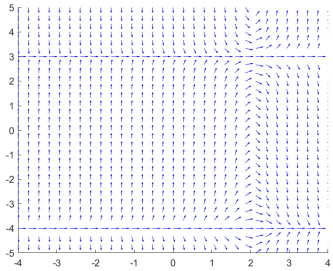


Figure 1.6: Exercise 1.2.13

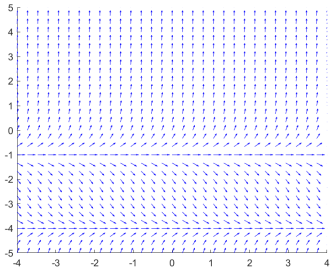


Figure 1.7: Exercise 1.2.14

***Exercise* 1.2.15** (challenging)**:** *Take $y' = f(x,y)$, $y(0) = 0$, where $f(x,y) > 1$ for all x and y. If the solution exists for all x, can you say what happens to $y(x)$ as x goes to positive infinity? Explain.*

***Exercise* 1.2.16:** *Suppose $y' = f(x,y)$. What will the slope field look like, explain and sketch an example, if you know the following about $f(x,y)$:*

a) f does not depend on y.

b) f does not depend on x.

c) $f(t,t) = 0$ for any number t.

d) $f(x,0) = 0$ and $f(x,1) = 1$ for all x.

***Exercise* 1.2.17:** *Describe what each of the following facts about the function $f(x,y)$ tells you about the slope field for the differential equation $y' = f(x,y)$.*

a) $f(2,y) = 0$ for all y

b) $f(x,-x) = 0$ for all x

c) $f(x,x) = 1$ for all x

d) $f(x,-1) = 0$ for all x

1.3 Separable equations

Attribution: [JL], §1.3.

Learning Objectives

After this section, you will be able to:

- Identify when a differential equation is separable,
- Find the general solution of a separable differential equation, and
- Solve initial value problems for separable differential equations.

As mentioned in § 1.1, when a differential equation is of the form $y' = f(x)$, we can just integrate: $y = \int f(x)\,dx + C$. Unfortunately this method no longer works for the general form of the equation $y' = f(x, y)$. Integrating both sides yields

$$y = \int f(x, y)\,dx + C.$$

Notice the dependence on y in the integral. Since y is a function of x, this expression is really of the form

$$y = \int f(x, y(x))\ dx + C$$

and without knowing what $y(x)$ is in advance (which we don't, because that's what we are trying to solve for) we can't compute this integral. Note that while you may have seen integrals of the form

$$\int f(x, y)\ dx$$

in Calculus 3, this is not the same situation. In that class, x and y were both independent variables, so we could integrate this expression in x, treating y as a constant. However, here y is a function of x, so they are not both independent variables and y can not be treated like a constant. If y is a function of x and any y shows up in the integral, you can not compute it.

1.3.1 Separable equations

One particular type of differential equation that we can evaluate using a technique very similar to direct integration is separable equations.

Definition 1.3.1

We say a differential equation is *separable* if we can write it as

$$y' = f(x)g(y),$$

for some functions $f(x)$ and $g(y)$.

Let us write the equation in the Leibniz notation

$$\frac{dy}{dx} = f(x)g(y).$$

Then we rewrite the equation as

$$\frac{dy}{g(y)} = f(x)\,dx.$$

It looks like we just separated the derivative as a fraction. The actual reasoning here is the differential from Calculus 1. This is the fact that for y a function of x, we know that

$$dy = \frac{dy}{dx}dx.$$

This means that we can take the equation

$$\frac{dy}{dx} = f(x)g(y),$$

rearrange it as

$$\frac{1}{g(y)}\frac{dy}{dx} = f(x)$$

and then multiply both sides by dx to get

$$\frac{1}{g(y)}\frac{dy}{dx}\,dx = f(x)dx$$

which leads to the rewritten equation above. Both sides look like something we can integrate. We obtain

$$\int \frac{dy}{g(y)} = \int f(x)\,dx + C.$$

If we can find closed form expressions for these two integrals, we can, perhaps, solve for y.

Example 1.3.1: Solve the equation

$$y' = xy.$$

Solution: Note that $y = 0$ is a solution. We will remember that fact and assume $y \neq 0$ from now on, so that we can divide by y. Write the equation as $\frac{dy}{dx} = xy$. Then

$$\int \frac{dy}{y} = \int x\,dx + C.$$

We compute the antiderivatives to get

$$\ln|y| = \frac{x^2}{2} + C,$$

or

$$|y| = e^{\frac{x^2}{2}+C} = e^{\frac{x^2}{2}}e^C = De^{\frac{x^2}{2}},$$

where $D > 0$ is some constant. Because $y = 0$ is also a solution and because of the absolute value we can write:

$$y = De^{\frac{x^2}{2}},$$

for any number D (including zero or negative).

We check:

$$y' = Dxe^{\frac{x^2}{2}} = x\left(De^{\frac{x^2}{2}}\right) = xy.$$

Yay! □

One particular case in which this method works very well is if the function $f(x, y)$ is only a function of y. With this, we can explicitly complete the solution to equations like

$$y' = ky,$$

reaching the solution $y(x) = e^{kx}$.

We should be a little bit more careful with this method. You may be worried that we integrated in two different variables. We seemingly did a different operation to each side. Let us work through this method more rigorously. Take

$$\frac{dy}{dx} = f(x)g(y).$$

We rewrite the equation as follows. Note that $y = y(x)$ is a function of x and so is $\frac{dy}{dx}$!

$$\frac{1}{g(y)}\frac{dy}{dx} = f(x).$$

We integrate both sides with respect to x:

$$\int \frac{1}{g(y)}\frac{dy}{dx}\,dx = \int f(x)\,dx + C.$$

We use the change of variables formula (substitution) on the left hand side:

$$\int \frac{1}{g(y)}\,dy = \int f(x)\,dx + C.$$

And we are done.

However, in some cases there are some special solutions to these problems as well that don't fit the same formula. Assume we have

$$\frac{dy}{dx} = f(x)g(y)$$

and we have a value y_0 such that $g(y_0) = 0$. Then, the function $y(x) = y_0$ is a solution, provided $f(x)$ is defined everywhere. (Plug this in and check!) This fills in the issue for having $\frac{1}{g(y)}$ in our integral expression, which is not defined when $g(y) = 0$. These are called *singular solutions*, and the next example will showcase one of them.

1.3.2 Implicit solutions

We sometimes get stuck even if we can do the integration. Consider the separable equation

$$y' = \frac{xy}{y^2+1}.$$

We separate variables,

$$\frac{y^2+1}{y}\,dy = \left(y + \frac{1}{y}\right)\,dy = x\,dx.$$

We integrate to get

$$\frac{y^2}{2} + \ln|y| = \frac{x^2}{2} + C,$$

or perhaps the easier looking expression (where $D = 2C$)

$$y^2 + 2\ln|y| = x^2 + D.$$

It is not easy to find the solution explicitly as it is hard to solve for y. We, therefore, leave the solution in this form and call it an *implicit solution*. It is still easy to check that an implicit solution satisfies the differential equation. In this case, we differentiate with respect to x, and remember that y is a function of x, to get

$$y'\left(2y + \frac{2}{y}\right) = 2x.$$

Multiply both sides by y and divide by $2(y^2+1)$ and you will get exactly the differential equation. We leave this computation to the reader.

If you have an implicit solution, and you want to compute values for y, you might have to be tricky. You might get multiple solutions y for each x, so you have to pick one. Sometimes you can graph x as a function of y, and then flip your paper. Sometimes you have to do more.

Computers are also good at some of these tricks. More advanced mathematical software usually has some way of plotting solutions to implicit equations, which makes these solutions just as good for visualizing or graphing as explicit solutions. For example, for $C = 0$ if you plot all the points (x, y) that are solutions to $y^2 + 2\ln|y| = x^2$, you find the two curves in Figure 1.8 on the next page. This is not quite a graph of a function. For each x there are two choices of y. To find a function you would have to pick one of these two curves. You pick the one that satisfies your initial condition if you have one. For example, the top curve satisfies the condition $y(1) = 1$. So for each C we really got two solutions. As you can see, computing values from an implicit solution can be somewhat tricky, but sometimes, an implicit solution is the best we can do.

The equation above also has the solution $y = 0$. Since our function

$$g(y) = \frac{y}{y^2+1}$$

is zero at $y = 0$, and gives an additional solution to the problem. The function $y(x) = 0$ satisfies $y'(x) = 0$ and $\frac{xy}{y^2+1} = 0$ for all x, which is the right-hand side of the equation. So the general solution is

$$y^2 + 2\ln|y| = x^2 + C, \qquad \text{and} \qquad y = 0.$$

These outlying solutions such as $y = 0$ are sometimes called *singular solutions*, as mentioned previously.

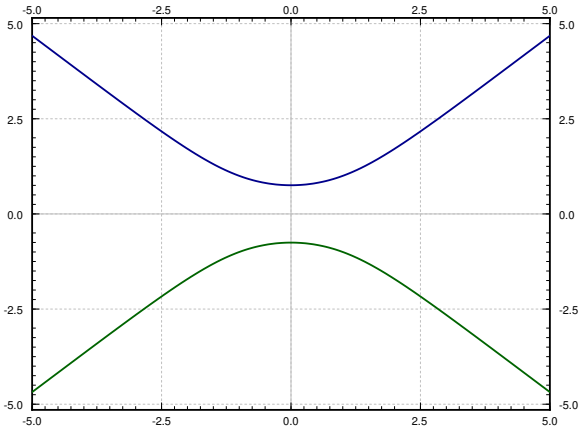


Figure 1.8: The implicit solution $y^2 + 2\ln|y| = x^2$ *to* $y' = \frac{xy}{y^2+1}$.

1.3.3 Examples of separable equations

Example 1.3.2: Solve $x^2y' = 1 - x^2 + y^2 - x^2y^2$, $y(1) = 0$.

Solution: Factor the right-hand side

$$x^2y' = (1 - x^2)(1 + y^2).$$

Separate variables, integrate, and solve for y:

$$\begin{aligned}
\frac{y'}{1+y^2} &= \frac{1-x^2}{x^2}, \\
\frac{y'}{1+y^2} &= \frac{1}{x^2} - 1, \\
\arctan(y) &= \frac{-1}{x} - x + C, \\
y &= \tan\left(\frac{-1}{x} - x + C\right).
\end{aligned}$$

Solve for the initial condition, $0 = \tan(-2 + C)$ to get $C = 2$ (or $C = 2 + \pi$, or $C = 2 + 2\pi$, etc.). The particular solution we seek is, therefore,

$$y = \tan\left(\frac{-1}{x} - x + 2\right).$$

Example 1.3.3: Bob made a cup of coffee, and Bob likes to drink coffee only once reaches 60 degrees Celsius and will not burn him. Initially at time $t = 0$ minutes, Bob measured the temperature and the coffee was 89 degrees Celsius. One minute later, Bob measured the coffee again and it had 85 degrees. The temperature of the room (the ambient temperature) is 22 degrees. When should Bob start drinking?

Solution: Let T be the temperature of the coffee in degrees Celsius, and let A be the ambient (room) temperature, also in degrees Celsius. Newton's law of cooling states that the rate at which the temperature of the coffee is changing is proportional to the difference between the ambient temperature and the temperature of the coffee. That is,

$$\frac{dT}{dt} = k(A - T),$$

for some constant k. For our setup $A = 22$, $T(0) = 89$, $T(1) = 85$. We separate variables and integrate (let C and D denote arbitrary constants):

$$\begin{aligned} \frac{1}{T-A}\frac{dT}{dt} &= -k, \\ \ln(T-A) &= -kt + C, \qquad \text{(note that } T - A > 0\text{)} \\ T - A &= D\,e^{-kt}, \\ T &= A + D\,e^{-kt}. \end{aligned}$$

That is, $T = 22 + D\,e^{-kt}$. We plug in the first condition: $89 = T(0) = 22 + D$, and hence $D = 67$. So $T = 22 + 67\,e^{-kt}$. The second condition says $85 = T(1) = 22 + 67\,e^{-k}$. Solving for k we get $k = -\ln\frac{85-22}{67} \approx 0.0616$. Now we solve for the time t that gives us a temperature of 60 degrees. Namely, we solve

$$60 = 22 + 67e^{-0.0616t}$$

to get $t = -\frac{\ln\frac{60-22}{67}}{0.0616} \approx 9.21$ minutes. So Bob can begin to drink the coffee at just over 9 minutes from the time Bob made it. That is probably about the amount of time it took us to calculate how long it would take. See Figure 1.9.

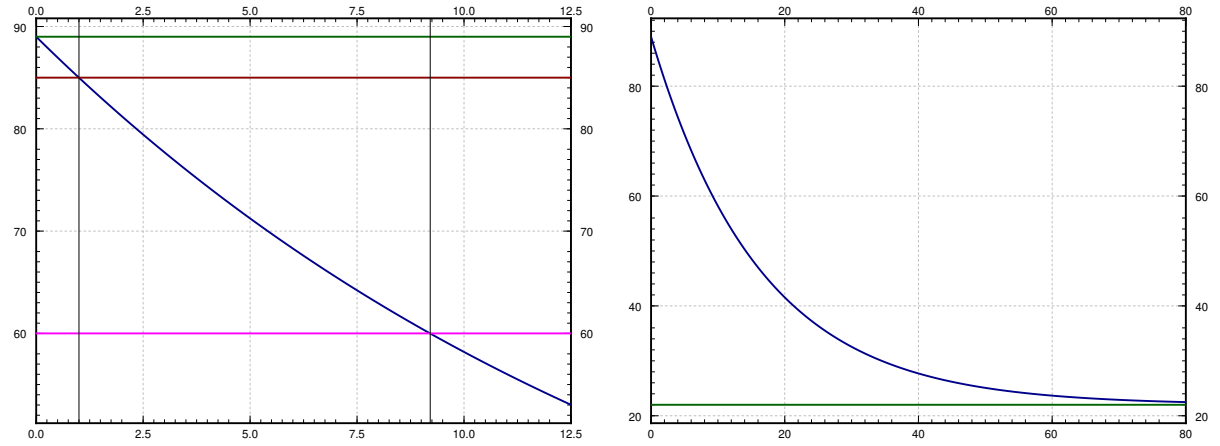


Figure 1.9: Graphs of the coffee temperature function $T(t)$. On the left, horizontal lines are drawn at temperatures 60, 85, and 89. Vertical lines are drawn at $t = 1$ and $t = 9.21$. Notice that the temperature of the coffee hits 85 at $t = 1$, and 60 at $t \approx 9.21$. On the right, the graph is over a longer period of time, with a horizontal line at the ambient temperature 22.

Example 1.3.4: Find the general solution to $y' = \frac{-xy^2}{3}$ (including singular solutions).

Solution: First note that $y = 0$ is a solution (a singular solution). Now assume that $y \neq 0$.

$$\begin{aligned} \frac{-3}{y^2}y' &= x, \\ \frac{3}{y} &= \frac{x^2}{2} + C, \\ y &= \frac{3}{x^2/2 + C} = \frac{6}{x^2 + 2C}. \end{aligned}$$

So the general solution is,

$$y = \frac{6}{x^2 + 2C}, \qquad \text{and} \qquad y = 0.$$

Example 1.3.5: Find the general solution to

$$\frac{dy}{dx} = (x^2 + e^x)(y^2 - 3y - 4).$$

Solution: Using the methods of separable equations, we can rewrite this differential equation as

$$\frac{dy}{y^2 - 3y - 4} = (x^2 + e^x)\ dx$$

and we can integrate both sides to solve. This leads to

$$\int \frac{dy}{y^2 - 3y + 4} = \int x^2 + e^x\ dx.$$

The right-hand side of this can be integrated normally to give

$$\int x^2 + e^x\ dx = \frac{x^3}{3} + e^x + C$$

and the left-hand side requires partial fractions in order to integrate correctly. If you are not familiar with this technique of partial fractions, it is reviewed in § B.3.

Using the method of partial fractions, we want to rewrite

$$\frac{1}{y^2 - 3y - 4} = \frac{A}{y - 4} + \frac{B}{y + 1}$$

and solve for A and B, which gives

$$\frac{1}{y^2 - 3y - 4} = \frac{1/5}{y - 4} - \frac{1/5}{y + 1}.$$

Therefore, we can compute the integral

$$\int \frac{dy}{y^2 - 3y - 4} = \int \frac{1/5}{y - 4} - \frac{1/5}{y + 1} dy = \frac{1}{5}\ln(|y - 4|) - \frac{1}{5}\ln(|y + 1|) + C.$$

Therefore, we can write the general solution as

$$\frac{1}{5}\ln\left(\frac{|y-4|}{|y+1|}\right) = \frac{x^3}{3} + e^x + C.$$

We could solve this out for y as an explicit function, but that is not necessary for a problem like this.

There are also two singular solutions here at $y = 4$ and $y = -1$. Notice that the implicit solution that we found previously is not defined at either of these values, because they involve taking the natural log of 0, which is not defined. ⌟

1.3.4 Exercises

***Exercise* 1.3.1:** *Solve $y' = y^3$ for $y(0) = 1$.*

Exercise* 1.3.2: *Solve $x' = \frac{1}{x^2}$, $x(1) = 1$.*

Exercise* 1.3.3** (little harder):*** *Solve $y' = (y-1)(y+1)$ for $y(0) = 3$. (Note: Requires partial fractions)*

Exercise* 1.3.4: *Solve $x' = \frac{1}{\cos(x)}$, $x(0) = \frac{\pi}{2}$.*

***Exercise* 1.3.5:** *Solve $\frac{dy}{dx} = \frac{1}{y+1}$ for $y(0) = 0$.*

***Exercise* 1.3.6:** *Solve $y' = x/y$.*

***Exercise* 1.3.7:** *Solve $y' = x^2 y$.*

Exercise* 1.3.8: *Consider the differential equation*

$$\frac{dy}{dx} = \frac{2x}{y}$$

a) Find the general solution as an implicit function.

b) Find the solution to this differential equation as an explicit *function with $y(1) = 4$.*

c) Find the solution to this differential equation as an explicit *function with $y(0) = -2$.*

Exercise* 1.3.9: *Solve $y' = y^n$, $y(0) = 1$, where n is a positive integer. Hint: You have to consider different cases.*

***Exercise* 1.3.10:** *Solve $\dfrac{dx}{dt} = (x^2 - 1)\,t$, for $x(0) = 0$. (Note: Requires partial fractions)*

***Exercise* 1.3.11:** *Solve $\dfrac{dx}{dt} = x\,\sin(t)$, for $x(0) = 1$.*

Exercise* 1.3.12: *Solve $y' = 2xy$.*

***Exercise* 1.3.13:** *Solve $y' = ye^{2x}$ with $y(0) = 4$.*

Exercise 1.3.14: *Solve* $\dfrac{dy}{dx} = xy + x + y + 1$. *Hint: Factor the right-hand side.*

Exercise 1.3.15:* *Solve* $x' = 3xt^2 - 3t^2$, $x(0) = 2$.

Exercise 1.3.16: *Find the general solution of* $y' = e^x$, *and then* $y' = e^y$.

Exercise 1.3.17: *Solve* $xy' = y + 2x^2y$, *where* $y(1) = 1$.

Exercise 1.3.18:* *Find an implicit solution for* $x' = \frac{1}{3x^2+1}$, $x(0) = 1$.

Exercise 1.3.19: *Solve* $\dfrac{dy}{dx} = \dfrac{y^2+1}{x^2+1}$, *for* $y(0) = 1$.

Exercise 1.3.20: *Find an implicit solution for* $\dfrac{dy}{dx} = \dfrac{x^2+1}{y^2+1}$, *for* $y(0) = 1$.

Exercise 1.3.21:* *Find an implicit solution to* $y' = \frac{\sin(x)}{\cos(y)}$.

Exercise 1.3.22: *Find an implicit solution for* $xy' = \frac{x^2+1}{y^2-1}$ *with* $y(3) = 2$.

Exercise 1.3.23: *Find an explicit solution for* $y' = xe^{-y}$, $y(0) = 1$.

Exercise 1.3.24:* *Find an explicit solution to* $xy' = y^2$, $y(1) = 1$.

Exercise 1.3.25: *Find an explicit solution for* $xy' = e^{-y}$, *for* $y(1) = 1$.

Exercise 1.3.26: *Find an explicit solution for* $y' = y^2(x^4 + 1)$ *with* $y(1) = 2$.

Exercise 1.3.27: *Find an explicit solution for* $y' = \frac{\cos(x)+1}{y}$ *with* $y(0) = 4$.

Exercise 1.3.28: *Find an explicit solution for* $y' = ye^{-x^2}$, $y(0) = 1$. *It is alright to leave a definite integral in your answer.*

Exercise 1.3.29: *Is the equation* $y' = x + y + 1$ *separable? If so, find the general solution, if not, explain why.*

Exercise 1.3.30: *Is the equation* $y' = ty^2 + t$ *separable? If so, find the general solution, if not, explain why.*

Exercise 1.3.31: *Is the equation* $y' = xy^2 + 3y^2 - 4x - 12$ *separable? If so, find the general solution, if not, explain why. (Note: Requires partial fractions)*

Exercise 1.3.32: *Suppose a cup of coffee is at 100 degrees Celsius at time* $t = 0$, *it is at 70 degrees at* $t = 10$ *minutes, and it is at 50 degrees at* $t = 20$ *minutes. Compute the ambient temperature.*

Exercise 1.3.33:* *Take Example 1.3.3 with the same numbers: 89 degrees at* $t = 0$, *85 degrees at* $t = 1$, *and ambient temperature of 22 degrees. Suppose these temperatures were measured with precision of* ± 0.5 *degrees. Given this imprecision, the time it takes the coffee to cool to (exactly) 60 degrees is also only known in a certain range. Find this range. Hint: Think about what kind of error makes the cooling time longer and what shorter.*

Exercise* 1.3.34: *A population x of rabbits on an island is modeled by $x' = x - \left(1/1000\right)x^2$, where the independent variable is time in months. At time $t = 0$, there are 40 rabbits on the island.*

a) Find the solution to the equation with the initial condition.

b) How many rabbits are on the island in 1 month, 5 months, 10 months, 15 months (round to the nearest integer).

1.4 Linear equations and the integrating factor

Attribution: [JL], §1.4.

Learning Objectives

After this section, you will be able to:

- Identify a linear first-order differential equation and write a first-order linear equation in standard form,
- Solve initial value problems for first-order linear differential equations by integrating factors, and
- Write solutions to first-order linear initial value problems in integral form if needed.

One of the most important types of equations we will learn how to solve are the so-called *linear equations*. In fact, the majority of the course is about linear equations. In this section we focus on the *first order linear equation*.

Definition 1.4.1

A first order equation is *linear* if we can put it into the form:

$$y' + p(x)y = f(x). \tag{1.3}$$

The word "linear" means linear in y and y'; no higher powers nor functions of y or y' appear. The dependence on x can be more complicated.

Solutions of linear equations have nice properties. For example, the solution exists wherever $p(x)$ and $f(x)$ are defined, and has the same regularity (read: it is just as nice). We'll see this in detail in § 1.5. But most importantly for us right now, there is a method for solving linear first order equations. In § 1.1, we saw that we could easily solve equations of the form

$$\frac{dy}{dx} = f(x)$$

because we could directly integrate both sides of the equation, since the left hand side was the derivative of something (in this case, y) and the right side was only a function of x. We want to do the same here, but the something on the left will not be the derivative of just y.

The trick is to rewrite the left-hand side of (1.3) as a derivative of a product of y with another function. Let $r(x)$ be this other function, and we can compute, by the product rule, that

$$\frac{d}{dx}\Big[r(x)y\Big] = r(x)y' + r'(x)y.$$

Now, if we multiply (1.3) by the function $r(x)$ on both sides, we get

$$r(x)y' + p(x)r(x)y = f(x)r(x)$$

and the first term on the left here matches the first term from our product rule derivative. To make the second terms match up as well, we need that

$$r'(x) = p(x)r(x).$$

This equation is separable! We can solve for the $r(x)$ here by separating variables to get that

$$\frac{dr}{r} = p(x)\,dx$$

so that

$$\ln|r| = \int p(x)\,dx$$

or

$$r(x) = e^{\int p(x)\,dx}.$$

With this choice of $r(x)$, we get that

$$r(x)y' + r(x)p(x)y = \frac{d}{dx}\Big[r(x)y\Big],$$

so that if we multiply (1.3) by $r(x)$, we obtain $r(x)y' + r(x)p(x)y$ on the left-hand side, which we can simplify using our product rule derivative above to obtain

$$\frac{d}{dx}\Big[r(x)y\Big] = r(x)f(x).$$

Now we integrate both sides. The right-hand side does not depend on y and the left-hand side is written as a derivative of a function. Afterwards, we solve for y. The function $r(x)$ is called the *integrating factor* and the method is called the *integrating factor method*.

This method works for any first order linear equation, no matter what $p(x)$ and $f(x)$ are. In general, we can compute:

$$\begin{aligned}
y' + p(x)y &= f(x),\\
e^{\int p(x)\,dx}y' + e^{\int p(x)\,dx}p(x)y &= e^{\int p(x)\,dx}f(x),\\
\frac{d}{dx}\left[e^{\int p(x)\,dx}y\right] &= e^{\int p(x)\,dx}f(x),\\
e^{\int p(x)\,dx}y &= \int e^{\int p(x)\,dx}f(x)\,dx + C,\\
y &= e^{-\int p(x)\,dx}\left(\int e^{\int p(x)\,dx}f(x)\,dx + C\right).
\end{aligned}$$

Advice: Do not try to remember the formula itself, that is way too hard. It is easier to remember the process and repeat it.

Of course, to get a closed form formula for y, we need to be able to find a closed form formula for the integrals appearing above.

Example 1.4.1: Solve

$$y' + 2xy = e^{x-x^2}, \qquad y(0) = -1.$$

Solution: First note that $p(x) = 2x$ and $f(x) = e^{x-x^2}$. The integrating factor is $r(x) = e^{\int p(x)\,dx} = e^{x^2}$. We multiply both sides of the equation by $r(x)$ to get

$$\begin{aligned} e^{x^2}y' + 2xe^{x^2}y &= e^{x-x^2}e^{x^2}, \\ \frac{d}{dx}\left[e^{x^2}y\right] &= e^x. \end{aligned}$$

We integrate

$$\begin{aligned} e^{x^2}y &= e^x + C, \\ y &= e^{x-x^2} + Ce^{-x^2}. \end{aligned}$$

Next, we solve for the initial condition $-1 = y(0) = 1 + C$, so $C = -2$. The solution is

$$y = e^{x-x^2} - 2e^{-x^2}.$$

Note that we do not care which antiderivative we take when computing $e^{\int p(x)dx}$. You can always add a constant of integration, but those constants will not matter in the end.

***Exercise* 1.4.1:** *Try it! Add a constant of integration to the integral in the integrating factor and show that the solution you get in the end is the same as what we got above.*

Since we cannot always evaluate the integrals in closed form, it is useful to know how to write the solution in definite integral form. A definite integral is something that you can plug into a computer or a calculator. Suppose we are given

$$y' + p(x)y = f(x), \qquad y(x_0) = y_0.$$

Look at the solution and write the integrals as definite integrals.

$$\boxed{y(x) = e^{-\int_{x_0}^{x} p(s)\,ds}\left(\int_{x_0}^{x} e^{\int_{x_0}^{t} p(s)\,ds} f(t)\,dt + y_0\right).} \tag{1.4}$$

You should be careful to properly use dummy variables here. If you now plug such a formula into a computer or a calculator, it will be happy to give you numerical answers.

***Exercise* 1.4.2:** *Check that $y(x_0) = y_0$ in formula (1.4).*

Example 1.4.2: Solve the initial value problem

$$ty' + 4y = t^2 - 1 \qquad y(1) = 3.$$

Solution: In order to solve this equation, we want to put the equation in standard form, which is

$$y' + \frac{4}{t}y = t - \frac{1}{t}.$$

In this form, the coefficient $p(t)$ of y is $p(t) = \frac{4}{t}$, so that the integrating factor is

$$r(t) = e^{\int p(t)\ dt} = e^{\int \frac{4}{t}\ dt} = e^{4\ln(t)}.$$

Since $4\ln(t) = \ln(t^4)$, we have that $r(t) = t^4$. Multiplying both sides of the equation by t^4 gives

$$t^4y' + 4t^3y = t^5 - t^3$$

where the left hand side is $\frac{d}{dt}(t^4y)$. Therefore, we can integrate both sides of the equation in t to give

$$t^4y = \frac{t^6}{6} - \frac{t^4}{4} + C$$

and we can solve out for y as

$$y(t) = \frac{t^2}{6} - \frac{1}{4} + \frac{C}{t^4}.$$

To solve for C using the initial condition, we plug in $t = 1$ to get that we need

$$3 = \frac{1}{6} - \frac{1}{4} + C \qquad C = \frac{37}{12}.$$

Therefore, the solution to the initial value problem is

$$y(t) = \frac{t^2}{6} - \frac{1}{4} + \frac{37/12}{t^4}.$$

Example 1.4.3: Solve the initial value problem

$$y' + 2xy = 3 \qquad y(0) = 4.$$

Solution: This equation is already in standard form. Since the coefficient of y is $p(x) = 2x$, we know that the integrating factor is

$$r(x) = e^{\int p(x)\ dx} = e^{x^2}.$$

We can multiply both sides of the equation by this integrating factor to give

$$y'e^{x^2} + 2xe^{x^2}y = 3e^{x^2}$$

and then want to integrate both sides. The left-hand side of the equation is $\frac{d}{dx}[e^{x^2}y]$, so the antiderivative of that side is just ye^{x^2}. For the right-hand side, we would need to compute

$$\int 3e^{x^2}\ dx,$$

which does not have a closed-form expression. Therefore, we need to represent this as a definite integral. Since our initial condition gives the value of y at zero, we want to use zero as the bottom limit of the integral. Therefore, we can write the solution as

$$ye^{x^2} = \int_0^x 3e^{s^2}\ ds + C$$

and so can solve for y as

$$y(x) = e^{-x^2} \int_0^x 3e^{s^2}\, ds + Ce^{-x^2}.$$

Plugging in the initial condition gives that

$$y(0) = 4 = e^{-0} \int_0^0 3e^{s^2}\, ds + Ce^{-0} = C.$$

Therefore, the solution to the initial value problem is

$$y(x) = e^{-x^2} \int_0^x 3e^{s^2}\, ds + 4e^{-x^2}.$$

□

***Exercise* 1.4.3:** *Write the solution of the following problem as a definite integral, but try to simplify as far as you can. You will not be able to find the solution in closed form.*

$$y' + y = e^{x^2 - x}, \qquad y(0) = 10.$$

1.4.1 Exercises

In the exercises, feel free to leave answer as a definite integral if a closed form solution cannot be found. If you can find a closed form solution, you should give that.

***Exercise* 1.4.4:** *Solve* $y' + xy = x$.

***Exercise* 1.4.5:** *Solve* $y' + 6y = e^x$.

***Exercise* 1.4.6:** *Solve* $y' + 4y = x^2 e^{-4x}$.

***Exercise* 1.4.7:** *Solve* $y' - 3y = xe^x$.

***Exercise* 1.4.8:** *Solve* $y' + 3y = e^{4x} - e^{-2x}$ *with* $y(0) = -3$.

***Exercise* 1.4.9:** *Solve* $y' - 2y = x + 4$.

***Exercise* 1.4.10:** *Solve* $xy' + 4y = x^2 - \frac{1}{x^2}$.

***Exercise* 1.4.11:** *Solve* $xy' - 3y = x - 2$ *with* $y(1) = 3$.

***Exercise* 1.4.12:** *Solve* $y' - 4y = \cos(3t)$.

Exercise* 1.4.13: *Solve* $y' + 3x^2 y = x^2$.

***Exercise* 1.4.14:** *Solve* $y' + 3x^2 y = \sin(x)\, e^{-x^3}$, *with* $y(0) = 1$.

***Exercise* 1.4.15:** *Solve* $y' + \cos(x)y = \cos(x)$.

***Exercise* 1.4.16:** *Solve the IVP* $4ty' + y = 24\sqrt{t};\ y(10000) = 100$.

***Exercise* 1.4.17:** *Solve the IVP* $(t^2 + 1)y' - 2ty = t^2 + 1;\ y(1) = 0$.

***Exercise* 1.4.18:** *Solve* $\frac{1}{x^2+1}\,y' + xy = 3$, *with* $y(0) = 0$.

Exercise* 1.4.19: *Solve* $y' + 2\sin(2x)y = 2\sin(2x)$, $y(\pi/2) = 3$.

***Exercise* 1.4.20:** *Consider the initial value problem*

$$5y' - 3y = e^{-2t} \qquad y(0) = a$$

for an undetermined value a. *Solve the problem and determine the dependence on the the value of* a. *How does the value of the solution as* $t \to \infty$ *depend on the value of* a*?*

***Exercise* 1.4.21:** *Find an expression for the general solution to* $y' + 3y = \sin(t^2)$ *with* $y(0) = 2$. *Simplfy your answer as much as possible.*

1.5 Existence and Uniqueness of Solutions

Attribution: [JL], §1.2.

Learning Objectives

After this section, you will be able to:

- Understand the terms existence and uniqueness as they apply to differential equations and
- Find the maximum guaranteed interval of existence for the solution to an initial value problem.

If we take the differential equation

$$y' = f(x, y) \qquad y(x_0) = y_0,$$

there are two main questions we want to answer about this equation.

(a) Does a solution exist to the differential equation?

(b) Is there only one solution to the differential equation?

These are more commonly referred to as (a) existence of the solution and (b) uniqueness of the solution. These are especially crucial for equations that we are using to model a physical situation. For physical situations, the solution definitely exists (because the system does something and continues to exist) and the solution is unique, because a given system will always do the same thing given the same setup. Since we know that physical systems obey these properties, the equations we use to model them should have these properties as well. These properties do not necessarily hold for all differential equations, as shown in the examples below.

Example 1.5.1: Attempt to solve:

$$y' = \frac{1}{x}, \qquad y(0) = 0.$$

Integrate to find the general solution $y = \ln|x| + C$. The solution does not exist at $x = 0$. See Figure 1.10 on the next page. The equation may have been written as the seemingly harmless $xy' = 1$.

Example 1.5.2: Solve:

$$y' = 2\sqrt{|y|}, \qquad y(0) = 0.$$

See Figure 1.11 on the following page. Note that $y = 0$ is a solution. But another solution is the function

$$y(x) = \begin{cases} x^2 & \text{if } x \geq 0, \\ -x^2 & \text{if } x < 0. \end{cases}$$

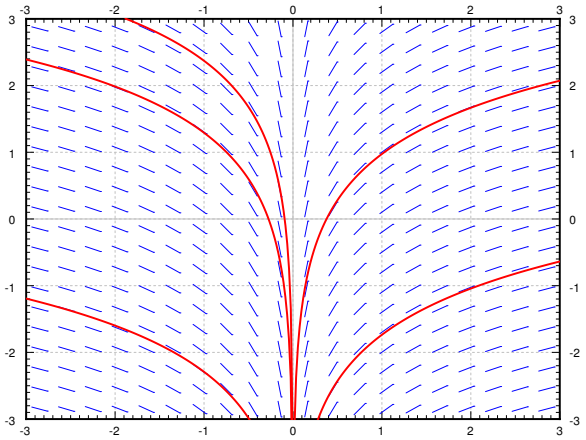


Figure 1.10: Slope field of $y' = 1/x$.

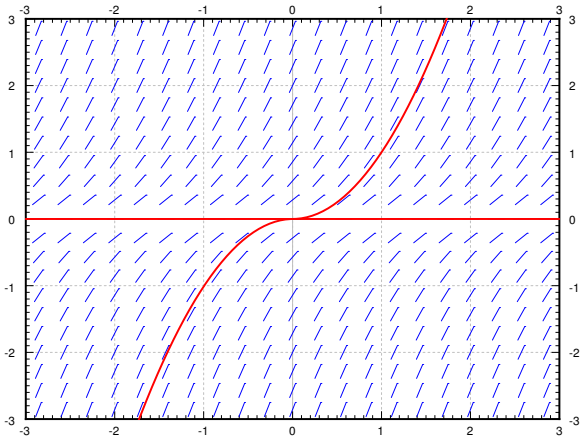


Figure 1.11: Slope field of $y' = 2\sqrt{|y|}$ with two solutions satisfying $y(0) = 0$.

What we see here is a significant problem for trying to represent physical situations. In the first there is no solution at $x = 0$, so if our physical scenario had wanted one, that would be an issue. Similarly, for the second, we do have solutions, but we have two of them, so we can't use this to predict what is going to happen to a physical situation modeled by this equation over time. So, we need both existence and uniqueness to hold for our modeling equation in order to use differential equations to accurately model situations. Thankfully, these properties do apply to most equations, and we have fairly straight-forward criteria that can be used to determine if these properties are true for a given differential equation. For a first-order linear differential equation, the theorem is fairly straight-forward.

Theorem 1.5.1

Assume that we have the first-order linear differential equation given by

$$y' + p(x)y = g(x).$$

If $p(x)$ and $g(x)$ are continuous functions on an interval I that contains a point x_0, then for any y-value y_0, the initial value problem

$$y' + p(x)y = g(x) \qquad y(x_0) = y_0$$

has a unique solution. This solution exists and is unique on the entire interval I.

The idea and proof of this theorem comes from the fact that we have an explicit method for solving these equations no matter what p and g are. We can always find an integrating factor for the equation, convert the left-hand side into a product rule term, take a definite integral of both sides, and then solve for y. Since we have this explicit formula, the solution will exist and be defined on the entire interval where the functions p and g are continuous. This also means that we can answer questions about where and for what values of x the solution to a differential equation exists.

Example 1.5.3: Consider the differential equation

$$(x-1)y' + \frac{1}{x-5}y = e^x$$

What do the existence and uniqueness theorems say about the solution to this differential equation with the initial condition $y(2) = 6$? What about the solution with initial condition $y(-3) = 1$?

Solution: To apply the existence and uniqueness theorem, we need to get the y' term by itself. This results in the differential equation

$$y' + \frac{1}{(x-1)(x-5)}y = \frac{e^x}{x-1}.$$

In order to figure out where this solution exists and is unique, we need to determine where the coefficient functions $p(x)$ and $g(x)$ are continuous. The only two points that we have discontinuities are at $x = 1$ and $x = 5$. Therefore, if we have the initial condition $y(2) = 6$, we start at the x value of 2. Because this equation is linear, it will exist everywhere that these two functions are both continuous containing the point $x = 2$, and since the only discontinuities are at 1 and 5, we know that they are both continuous on $(1, 5)$. This means that we can take $(1, 5)$ as the interval I in the theorem, and know that this solution will exist and be unique on the interval $(1, 5)$.

For the other initial condition, $y(-3) = 1$, we now want an interval where these functions are continuous that contains -3. Again, we only have to avoid $x = 1$ and $x = 5$, so we can take the interval $(-\infty, 1)$ as the interval I in the theorem, and so we know the solution with this initial condition will exist and be unique on $(-\infty, 1)$.

A convenient way to represent this situation is with a number line like that presented in Figure 1.12. On this number line, we mark the places where the functions $p(x)$ or $g(x)$ are discontinuous.

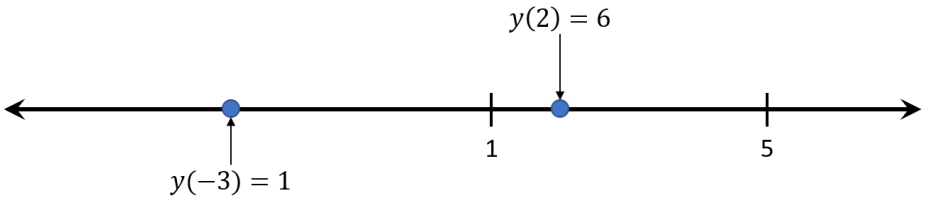


Figure 1.12: Number line representation of the existence intervals for a differential equation.

To interpret this image, we can mark the initial point on the number line, where the point that we mark is the x coordinate of the initial condition. All of the intervals are in terms of x. Then, the existence and uniqueness theorem says that the solution will exist on the entire interval between any marked points on this number line. From that, we can see that the interval of existence for the initial condition $y(2) = 6$ is $(1, 5)$, and the interval for $y(-3) = 1$ is $(-\infty, 1)$.

For non-linear equations, we don't have an explicit method of getting a solution that works for all equations. This means that we can't fall back on this formula to guarantee existence or uniqueness of solutions. For this reason, we expect to get a result that is not as strong for non-linear equations. Thankfully, we do still get a result, which is known as Picard's theorem*.

Theorem 1.5.2 (Picard's theorem on existence and uniqueness)

If $f(x,y)$ is continuous (as a function of two variables) and $\frac{\partial f}{\partial y}$ exists and is continuous near some (x_0, y_0), then a solution to

$$y' = f(x,y), \qquad y(x_0) = y_0,$$

exists (at least for some small interval of x's) and is unique.

The main fact that is "not as strong" about this result is the interval that we get from the theorem. For the linear theorem, we got existence and uniqueness on the entire interval I where p and g are continuous. For the non-linear theorem, we only get existence on *some* interval around the point x_0. Even if $f(x,y)$ and $\frac{\partial f}{\partial y}$ are really nice functions that are continuous everywhere, we can still only guarantee existence on a small interval (that can depend on the initial condition) around the point x_0.

Example 1.5.4: For some constant A, solve:

$$y' = y^2, \qquad y(0) = A.$$

Solution: We know how to solve this equation. First assume that $A \neq 0$, so y is not equal to zero at least for some x near 0. So $x' = 1/y^2$, so $x = -1/y + C$, so $y = \frac{1}{C-x}$. If $y(0) = A$, then $C = 1/A$ so

$$y = \frac{1}{1/A - x}.$$

If $A = 0$, then $y = 0$ is a solution.

For example, when $A = 1$ the solution is

$$y = \frac{1}{1-x}$$

which goes to infinity, and so "blows up", at $x = 1$. This solution here exists only on the interval $(-\infty, 1)$, and hence, the solution does not exist for all x even if the equation is nice everywhere. The equation $y' = y^2$ certainly looks nice.

However, this fact does not contradict our existence and uniqueness theorem for non-linear equations. The theorem only guarantees that the solution to

$$y' = y^2$$

exists and is unique on *some* interval containing 0. It does not guarantee that the solution exists everywhere that y^2 and its derivative are continuous, only that at each point where this

*Named after the French mathematician Charles Émile Picard (1856–1941)

happens, the solution will exist for some interval around that point. The interval $(-\infty, 1)$ is "some interval containing 0", so the theorem still applies and holds here. See the exercises for more detail on how this process works and how we can illustrate the fact that the interval of existence ise "some interval containing 0". ⌟

The other main conclusion that we can draw from these theorems is the fact that two different solution curves to a first-order differential equation can not cross, provided the existence and uniqueness theorems hold. If y_1 and y_2 are two different solutions to $y' = f(x, y)$ and the solution curves for $y_1(x)$ and $y_2(x)$ cross, then this means that for some particular value of x_0 and y_0, we have that

$$y_1(x_0) = y_0 \qquad y_2(x_0) = y_0.$$

If we pick x_0 as a starting point, then the fact that the existence and uniqueness theorems hold imply that, at least for some interval around x_0, there is exactly one solution to

$$y' = f(x, y) \qquad y(x_0) = y_0.$$

However, both y_1 and y_2 satisfy these two properties. Therefore, y_1 and y_2 must be the same, which doesn't make sense because we assumed they were different. So it is impossible for two different solution curves to cross, provided the existence and uniqueness theorem holds. For a comparison, refer back to Example 1.5.2 earlier to see what non-uniqueness looks like, where we do have two solution curves that cross at the point $(0, 0)$.

This fact is useful for analyzing differential equations in general, but will be particularly useful in § 1.7 in dealing with autonomous equations, where we can use simple solutions to provide boundaries over which other solutions can not cross. This fact will come up again in Chapters 4 and 5 in sketching trajectories for these solutions as well.

Example 1.5.5: Consider the differential equation

$$\frac{dy}{dt} = (y-3)^2(y+4).$$

1. Verify that $y = 3$ is a solution to this differential equation.
2. Assume that we solve this problem with initial condition $y(0) = 1$. Is it possible for this solution to ever reach $y = 4$? Why or why not?

Solution:

1. If we take the function $y(t) = 3$, then $y' = 0$, and plugging this into the right hand side also gives 0. Therefore, this function solves the differential equation.
2. If the solutions starts with $y(0) = 1$, this means that it starts below the line $y = 3$. In order to get up to $y = 4$, the solution would need to cross over the line $y = 3$, which would mean that we have solution curves that cross. However, the function $f(t, y) = (y-3)^2(y+4)$ is continuous everywhere, as is the first derivative $\frac{\partial f}{\partial y} = 2(y-3)(y+4) + (y-3)^2$. Therefore, the existence and uniqueness theorem applies everywhere, and so solution curves can not cross. So, it is not possible for the solution to reach $y = 4$, because this would force solution curves to cross, which we know can not happen.

⌟

1.5.1 Exercises

***Exercise* 1.5.1:** *Is it possible to solve the equation* $y' = \frac{xy}{\cos x}$ *for* $y(0) = 1$*? Justify.*

***Exercise* 1.5.2:** *Is it possible to solve the equation* $y' = y\sqrt{|x|}$ *for* $y(0) = 0$*? Is the solution unique? Justify.*

***Exercise* 1.5.3:** *Consider the differential equation* $y' + \frac{1}{t-2}y = \frac{1}{t+3}$.

a) *Is this equation linear or non-linear?*

b) *What is the maximum guaranteed interval of existence for the solution to this equation with initial condition* $y(0) = 3$*?*

c) *What if we start with the initial condition* $y(4) = 0$*?*

***Exercise* 1.5.4:** *Consider the differential equation* $y' + \frac{1}{t+2}y = \frac{\ln(|t|)}{t-4}$.

a) *Is this equation linear or non-linear?*

b) *What is the maximum guaranteed interval of existence for the solution to this equation with initial condition* $y(-3) = 1$*?*

c) *What if we start with the initial condition* $y(2) = 5$*?*

d) *What happens if we want to start with* $y(4) = 3$*?*

***Exercise* 1.5.5:** *Consider the differential equation* $(t+3)y' + t^2y = \frac{1}{t-2}$.

a) *Is this equation linear or non-linear?*

b) *What is the maximum guaranteed interval of existence for the solution to this equation with initial condition* $y(-2) = 1$*?*

c) *What if we start with the initial condition* $y(-4) = 5$*?*

d) *What happens if we want to start with* $y(4) = 2$*?*

***Exercise* 1.5.6:** *Consider the differential equation* $y' = y^2$.

a) *Is this equation linear or non-linear?*

b) *What is the most we can say about the interval of existence for the solution to this equation with initial condition* $y(0) = 1$*?*

c) *Find the solution to this differential equation with* $y(0) = 1$*. Over what values of* x *does this solution exist?*

d) *Find the solution to this differential equation with* $y(0) = 4$*. Over what values of* x *does this solution exist?*

e) *Find the solution to this differential equation with* $y(0) = -2$*. Over what values of* x *does this solution exist?*

f) *Do any of these contradict your answer in (b)?*

***Exercise* 1.5.7:** *Consider the differential equation $y' = y^2 + 4$.*

a) Is this equation linear or non-linear?

b) What is the most we can say about the interval of existence for the solution to this equation with initial condition $y(0) = 0$?

c) Find the solution to this differential equation with $y(0) = 0$. Over what values of x does this solution exist?

***Exercise* 1.5.8:** *Consider the differential equation $y' = x(y + 1)^2$.*

a) Is this equation linear or non-linear?

b) If we set $f(x, y) = x(y + 1)^2$, for what values of x and y are f and $\frac{\partial f}{\partial y}$ continuous?

c) What is the most we can say about the interval of existence for the solution to this equation with initial condition $y(0) = 1$?

d) Find the solution to this differential equation with $y(0) = 1$. Over what values of x does this solution exist?

***Exercise* 1.5.9** (challenging)**:** *Take $(y - x)y' = 0$, $y(0) = 0$.*

a) Find two distinct solutions.

b) Explain why this does not violate Picard's theorem.

***Exercise* 1.5.10:** *Find a solution to $y' = |y|$, $y(0) = 0$. Does Picard's theorem apply?*

***Exercise* 1.5.11:** *Consider the IVP $y' \cos t + y \sin t = 1$; $y(\pi/6) = 1$.*

a) The Existence and Uniqueness Theorem guarantees a unique solution to this IVP on what interval?

b) Find this solution explicitly.

***Exercise* 1.5.12:** *Take an equation $y' = (y - 2x)g(x, y) + 2$ for some function $g(x, y)$. Can you solve the problem for the initial condition $y(0) = 0$, and if so what is the solution?*

***Exercise* 1.5.13:** *Consider the differential equation $y' = e^x(y - 2)$.*

a) Verify that $y = 2$ is a solution to this differential equation.

b) Assume that we look for the solution with $y(0) = 0$. Is it possible that $y(x) = 3$ for some later time x? Why or why not?

c) Based on this, what do we know about the solution with $y(0) = 5$?

Exercise* 1.5.14** (challenging):*** *Suppose $y' = f(x, y)$ is such that $f(x, 1) = 0$ for every x, f is continuous and $\frac{\partial f}{\partial y}$ exists and is continuous for every x and y.*

a) *Guess a solution given the initial condition $y(0) = 1$.*

b) *Can graphs of two solutions of the equation for different initial conditions ever intersect?*

c) *Given $y(0) = 0$, what can you say about the solution. In particular, can $y(x) > 1$ for any x? Can $y(x) = 1$ for any x? Why or why not?*

Exercise* 1.5.15:*** *Consider the differential equation $y' = y^2 - 4$.*

a) *Verify that $y = 2$ and $y = -2$ are both solutions to this differential equation.*

b) *Verify that the hypotheses of Picard's theorem are satisfies for this equation.*

c) *Assume that we solve this differential equation with $y(0) = 1$. Is it possible for the solution to reach $y = 3$ at any point? Why or why not?*

d) *Assume that we solve this differential equation with $y(0) = -1$. Is it possible for the solution to reach $y = -4$ at any point? Why or why not?*

Exercise* 1.5.16:** *Is it possible to solve $y' = xy$ for $y(0) = 0$? Is the solution unique?*

Exercise* 1.5.17:*** *Is it possible to solve $y' = \frac{x}{x^2-1}$ for $y(1) = 0$?*

Exercise* 1.5.18** (tricky):*** *Suppose*

$$f(y) = \begin{cases} 0 & \text{if } y > 0, \\ 1 & \text{if } y \leq 0. \end{cases}$$

Does $y' = f(y)$, $y(0) = 0$ have a continuously differentiable solution? Does Picard apply? Why, or why not?

Exercise* 1.5.19:** *Consider an equation of the form $y' = f(x)$ for some continuous function f, and an initial condition $y(x_0) = y_0$. Does a solution exist for all x? Why or why not?*

1.6 Numerical methods: Euler's method

Attribution: [JL], §1.7.

Learning Objectives

After this section, you will be able to:

- Use Euler's method to numerically approximate solutions to first order differential equations,
- Compute the error in a numerical method using the true solution, and
- Compare a variety of numerical methods, including built-in Matlab methods.

Unless $f(x, y)$ is of a special form, it is generally very hard if not impossible to get a nice formula for the solution of the problem

$$y' = f(x, y), \qquad y(x_0) = y_0.$$

If the equation can be solved in closed form, we should do that. But what if we have an equation that cannot be solved in closed form? What if we want to find the value of the solution at some particular x? Or perhaps we want to produce a graph of the solution to inspect the behavior. In this section we will learn about the basics of numerical approximation of solutions.

The simplest method for approximating a solution is *Euler's method**. It works as follows: Take x_0 and compute the slope $k = f(x_0, y_0)$. The slope is the change in y per unit change in x. Follow the line for an interval of length h on the x-axis. Hence if $y = y_0$ at x_0, then we say that y_1 (the approximate value of y at $x_1 = x_0 + h$) is $y_1 = y_0 + hk$. Rinse, repeat! Let $k = f(x_1, y_1)$, and then compute $x_2 = x_1 + h$, and $y_2 = y_1 + hk$. Now compute x_3 and y_3 using x_2 and y_2, etc. Consider the equation $y' = y^2/3$, $y(0) = 1$, and $h = 1$. Then $x_0 = 0$ and $y_0 = 1$. We compute

$$x_1 = x_0 + h = 0 + 1 = 1, \qquad y_1 = y_0 + h\,f(x_0, y_0) = 1 + 1 \cdot 1^2/3 = 4/3 \approx 1.333,$$

$$x_2 = x_1 + h = 1 + 1 = 2, \qquad y_2 = y_1 + h\,f(x_1, y_1) = 4/3 + 1 \cdot \frac{(4/3)^2}{3} = 52/27 \approx 1.926.$$

We then draw an approximate graph of the solution by connecting the points (x_0, y_0), (x_1, y_1), (x_2, y_2),.... For the first two steps of the method see Figure 1.13 on the following page.

More abstractly, for any $i = 0, 1, 2, 3, \ldots$, we compute

$$x_{i+1} = x_i + h, \qquad y_{i+1} = y_i + h\,f(x_i, y_i).$$

This can be worked out by hand for a few steps, but the formulas here lend themselves very well to being coded into a looping structure for more involved processes. The line segments we get are an approximate graph of the solution. Generally it is not exactly the solution. See Figure 1.14 on the next page for the plot of the real solution and the approximation.

*Named after the Swiss mathematician Leonhard Paul Euler (1707–1783). The correct pronunciation of the name sounds more like "oiler."

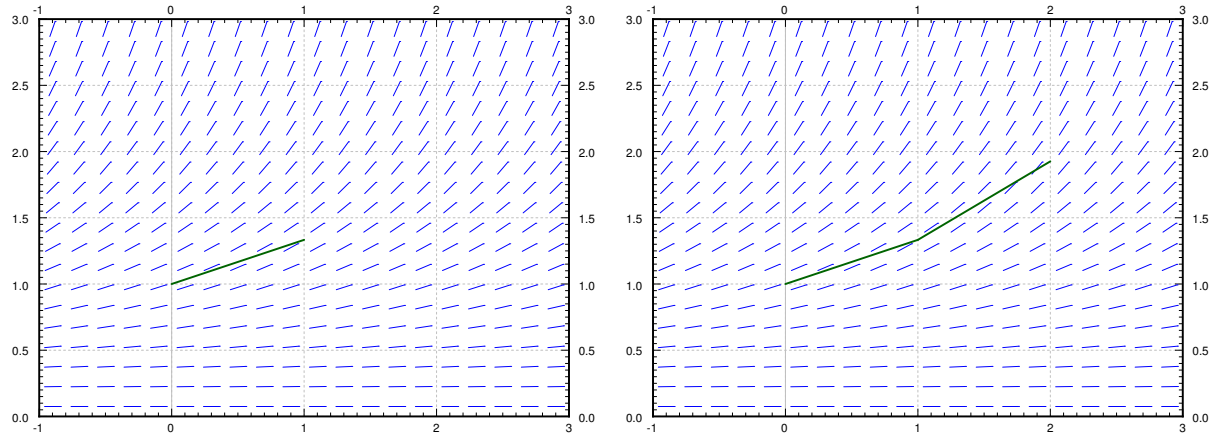


Figure 1.13: First two steps of Euler's method with $h = 1$ for the equation $y' = \frac{y^2}{3}$ with initial conditions $y(0) = 1$.

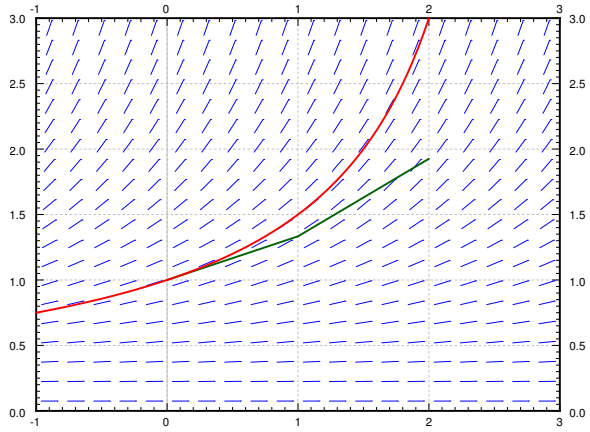


Figure 1.14: Two steps of Euler's method (step size 1) and the exact solution for the equation $y' = \frac{y^2}{3}$ with initial conditions $y(0) = 1$.

We continue with the equation $y' = y^2/3$, $y(0) = 1$. Let us try to approximate $y(2)$ using Euler's method. In Figures 1.13 and 1.14 we have graphically approximated $y(2)$ with step size 1. With step size 1, we have $y(2) \approx 1.926$. The real answer is 3. We are approximately 1.074 off. Let us halve the step size. Computing y_4 with $h = 0.5$, we find that $y(2) \approx 2.209$, so an error of about 0.791. Table 1.1 on the facing page gives the values computed for various parameters.

***Exercise* 1.6.1:** *Solve this equation exactly and show that $y(2) = 3$.*

The difference between the actual solution and the approximate solution is called the error. We usually talk about just the size of the error and we do not care much about its sign. The point is, we usually do not know the real solution, so we only have a vague understanding of the error. If we knew the error exactly ... what is the point of doing the approximation?

h	Approximate $y(2)$	Error	$\frac{\text{Error}}{\text{Previous error}}$
1	1.92593	1.07407	
0.5	2.20861	0.79139	0.73681
0.25	2.47250	0.52751	0.66656
0.125	2.68034	0.31966	0.60599
0.0625	2.82040	0.17960	0.56184
0.03125	2.90412	0.09588	0.53385
0.015625	2.95035	0.04965	0.51779
0.0078125	2.97472	0.02528	0.50913

Table 1.1: Euler's method approximation of $y(2)$ where of $y' = y^2/3$, $y(0) = 1$.

Notice that except for the first few times, every time we halved the step size the error approximately halved. This halving of the error is a general feature of Euler's method as it is a *first order method*. There exists an improved Euler method, see the exercises, which is a second order method. A second order method reduces the error to approximately one quarter every time we halve the interval. The meaning of "second" order is the squaring in $1/4 = 1/2 \times 1/2 = (1/2)^2$.

To get the error to be within 0.1 of the answer we had to already do 64 steps. To get it to within 0.01 we would have to halve another three or four times, meaning doing 512 to 1024 steps. That is quite a bit to do by hand. The improved Euler method from the exercises should quarter the error every time we halve the interval, so we would have to approximately do half as many "halvings" to get the same error. This reduction can be a big deal. With 10 halvings (starting at $h = 1$) we have 1024 steps, whereas with 5 halvings we only have to do 32 steps, assuming that the error was comparable to start with. A computer may not care about this difference for a problem this simple, but suppose each step would take a second to compute (the function may be substantially more difficult to compute than $y^2/3$). Then the difference is 32 seconds versus about 17 minutes. We are not being altogether fair, a second order method would probably double the time to do each step. Even so, it is 1 minute versus 17 minutes. Next, suppose that we have to repeat such a calculation for different parameters a thousand times. You get the idea.

Note that in practice we do not know how large the error is! How do we know what is the right step size? Well, essentially we keep halving the interval, and if we are lucky, we can estimate the error from a few of these calculations and the assumption that the error goes down by a factor of one half each time (if we are using standard Euler).

***Exercise* 1.6.2:** *In the table above, suppose you do not know the error. Take the approximate values of the function in the last two lines, assume that the error goes down by a factor of 2. Can you estimate the error in the last time from this? Does it (approximately) agree with the table? Now do it for the first two rows. Does this agree with the table?*

Let us talk a little bit more about the example $y' = \frac{y^2}{3}$, $y(0) = 1$. Suppose that instead

of the value $y(2)$ we wish to find $y(3)$. The results of this effort are listed in Table 1.2 for successive halvings of h. What is going on here? Well, you should solve the equation exactly and you will notice that the solution does not exist at $x = 3$. In fact, the solution goes to infinity when you approach $x = 3$.

h	Approximate $y(3)$
1	3.16232
0.5	4.54329
0.25	6.86079
0.125	10.80321
0.0625	17.59893
0.03125	29.46004
0.015625	50.40121
0.0078125	87.75769

Table 1.2: Attempts to use Euler's to approximate $y(3)$ where of $y' = y^2/3$, $y(0) = 1$.

Another case where things go bad is if the solution oscillates wildly near some point. The solution may exist at all points, but even a much better numerical method than Euler would need an insanely small step size to approximate the solution with reasonable precision. And computers might not be able to easily handle such a small step size.

In real applications we would not use a simple method such as Euler's. The simplest method that would probably be used in a real application is the standard Runge–Kutta method (see exercises). That is a fourth order method, meaning that if we halve the interval, the error generally goes down by a factor of 16 (it is fourth order as $1/16 = 1/2 \times 1/2 \times 1/2 \times 1/2$).

Choosing the right method to use and the right step size can be very tricky. There are several competing factors to consider.

- Computational time: Each step takes computer time. Even if the function f is simple to compute, we do it many times over. Large step size means faster computation, but perhaps not the right precision.

- Roundoff errors: Computers only compute with a certain number of significant digits. Errors introduced by rounding numbers off during our computations become noticeable when the step size becomes too small relative to the quantities we are working with. So reducing step size may in fact make errors worse. There is a certain optimum step size such that the precision increases as we approach it, but then starts getting worse as we make our step size smaller still. Trouble is: this optimum may be hard to find.

- Stability: Certain equations may be numerically unstable. What may happen is that the numbers never seem to stabilize no matter how many times we halve the interval. We may need a ridiculously small interval size, which may not be practical due to roundoff

errors or computational time considerations. Such problems are sometimes called *stiff*. In the worst case, the numerical computations might be giving us bogus numbers that look like a correct answer. Just because the numbers seem to have stabilized after successive halving, does not mean that we must have the right answer.

We have seen just the beginnings of the challenges that appear in real applications. Numerical approximation of solutions to differential equations is an active research area for engineers and mathematicians. For example, the general purpose method used for the ODE solver in Matlab and Octave (as of this writing) is a method that appeared in the literature only in the 1980s.

The method used in Matlab and Octave is a fair bit different from the methods discussed previously. We don't need to go too much in detail about it, but some information will be helpful. The main difference that will be visible when running these methods is that they are *adaptive* method. This means that they adjust the step-size based on what the differential equation looks like at a given point. Euler's method, along with the improved Euler and Runge-Kutta methods, is a fixed step-size method, where the steps are always the same no matter what. Adaptive methods are harder to write and optimize, but can solve many problems faster because the adaptive nature of the method allows them to get similar accuracy to fixed step methods, but at many fewer steps. In the example below, the initial value problem

$$\frac{dy}{dt} = y \qquad y(0) = 1$$

is solved with an Euler's method and Matlab's built-in `ode45` method. Both of the solutions are plotted along with the actual solution $y = e^t$

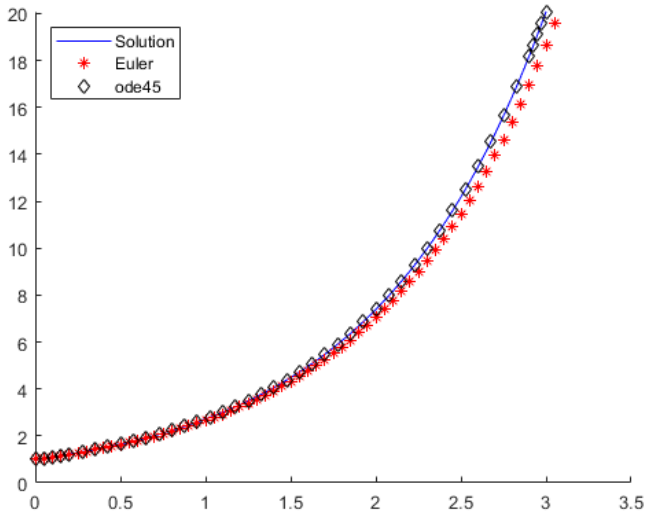


Figure 1.15: Comparison of the solution from Euler's Method and ode45 to the actual solution of $\frac{dy}{dt} = y$.

The Euler's method takes 60 steps in this computation, but is still not as accurate as the `ode45` method, which only takes 45 steps. In addition, the black diamonds, representing the different values computed by `ode45` are not evenly spaced, illustrating the adaptive nature of this solver, while the red stars are all evenly spaced in the t-direction, as is expected from Euler's method.

1.6.1 Exercises

***Exercise* 1.6.3:** *Consider* $\frac{dx}{dt} = (2t - x)^2$, $x(0) = 2$. *Use Euler's method with step size* $h = 0.5$ *to approximate* $x(1)$.

***Exercise* 1.6.4:** *Consider the differential equation* $\frac{dy}{dt} = t^2 - 3y + 1$ *with* $y(1) = 4$. *Approximate the solution at* $t = 3$ *using Euler's method with a step size of* $h = 1$ *and* $h = 0.5$. *Compare these values with the actual solution at* $t = 3$.

***Exercise* 1.6.5:** *Consider the differential equation* $\frac{dy}{dt} = 2ty + y^2$ *with* $y(0) = 1$. *Approximate the solution at* $t = 2$ *using Euler's method with a step size of* $h = 1$ *and* $h = 0.5$.

***Exercise* 1.6.6:** *Consider* $\frac{dx}{dt} = t - x$, $x(0) = 1$.

a) *Use Euler's method with step sizes* $h = 1, 1/2, 1/4, 1/8$ *to approximate* $x(1)$.

b) *Solve the equation exactly.*

c) *Describe what happens to the errors for each* h *you used. That is, find the factor by which the error changed each time you halved the interval.*

Exercise* 1.6.7: *Let* $x' = \sin(xt)$, *and* $x(0) = 1$. *Approximate* $x(1)$ *using Euler's method with step sizes 1, 0.5, 0.25. Use a calculator and compute up to 4 decimal digits.*

***Exercise* 1.6.8:** *Approximate the value of* e *by looking at the initial value problem* $y' = y$ *with* $y(0) = 1$ *and approximating* $y(1)$ *using Euler's method with a step size of* 0.2.

Exercise* 1.6.9: *Let* $x' = 2t$, *and* $x(0) = 0$.

a) *Approximate* $x(4)$ *using Euler's method with step sizes 4, 2, and 1.*

b) *Solve exactly, and compute the errors.*

c) *Compute the factor by which the errors changed.*

Exercise* 1.6.10: *Let* $x' = xe^{xt+1}$, *and* $x(0) = 0$.

a) *Approximate* $x(4)$ *using Euler's method with step sizes 4, 2, and 1.*

b) *Guess an exact solution based on part a) and compute the errors.*

***Exercise* 1.6.11:** *Example of numerical instability: Take $y' = -5y$, $y(0) = 1$. We know that the solution should decay to zero as x grows. Using Euler's method, start with $h = 1$ and compute y_1, y_2, y_3, y_4 to try to approximate $y(4)$. What happened? Now halve the interval. Keep halving the interval and approximating $y(4)$ until the numbers you are getting start to stabilize (that is, until they start going towards zero). Note: You might want to use a calculator.*

There is a simple way to improve Euler's method to make it a second order method by doing just one extra step. Consider $\frac{dy}{dx} = f(x, y)$, $y(x_0) = y_0$, and a step size h. What we do is to pretend we compute the next step as in Euler, that is, we start with (x_i, y_i), we compute a slope $k_1 = f(x_i, y_i)$, and then look at the point $(x_i + h, y_i + k_1 h)$. Instead of letting our new point be $(x_i + h, y_i + k_1 h)$, we compute the slope at that point, call it k_2, and then take the average of k_1 and k_2, hoping that the average is going to be closer to the actual slope on the interval from x_i to $x_i + h$. And we are correct, if we halve the step, the error should go down by a factor of $2^2 = 4$. To summarize, the setup is the same as for regular Euler, except the computation of y_{i+1} and x_{i+1}.

$$\begin{aligned} k_1 &= f(x_i, y_i), & x_{i+1} &= x_i + h, \\ k_2 &= f(x_i + h, y_i + k_1 h), & y_{i+1} &= y_i + \frac{k_1 + k_2}{2}\, h. \end{aligned}$$

Exercise* 1.6.12: *Consider $\dfrac{dy}{dx} = x + y$, $y(0) = 1$.*

a) Use the improved Euler's method (see above) with step sizes $h = 1/4$ and $h = 1/8$ to approximate $y(1)$.

b) Use Euler's method with $h = 1/4$ and $h = 1/8$.

c) Solve exactly, find the exact value of $y(1)$.

d) Compute the errors, and the factors by which the errors changed.

The simplest method used in practice is the *Runge–Kutta method*. Consider $\frac{dy}{dx} = f(x, y)$, $y(x_0) = y_0$, and a step size h. Everything is the same as in Euler's method, except the computation of y_{i+1} and x_{i+1}.

$$\begin{aligned} k_1 &= f(x_i, y_i), \\ k_2 &= f\big(x_i + h/2, y_i + k_1(h/2)\big), & x_{i+1} &= x_i + h, \\ k_3 &= f\big(x_i + h/2, y_i + k_2(h/2)\big), & y_{i+1} &= y_i + \frac{k_1 + 2k_2 + 2k_3 + k_4}{6}\, h, \\ k_4 &= f(x_i + h, y_i + k_3 h). \end{aligned}$$

***Exercise* 1.6.13:** *Consider $\dfrac{dy}{dx} = yx^2$, $y(0) = 1$.*

a) Use Runge–Kutta (see above) with step sizes $h = 1$ and $h = 1/2$ to approximate $y(1)$.

b) Use Euler's method with $h = 1$ and $h = 1/2$.

c) Solve exactly, find the exact value of $y(1)$, and compare.

1.7 Autonomous equations

Attribution: [JL], §1.6.

Learning Objectives

After this section, you will be able to:

- Identify autonomous first order differential equations,
- Find critical points or equilibrium solutions for autonomous equations, and
- Sketch a phase line for these equations.

Definition 1.7.1

An equation of the form

$$\frac{dx}{dt} = f(x), \tag{1.5}$$

where the derivative of solutions depends only on x (the dependent variable) is called an *autonomous equation*. If we think of t as time, the naming comes from the fact that the equation is independent of time.

We return to the cooling coffee problem (Example 1.3.3). Newton's law of cooling says

$$\frac{dx}{dt} = k(A - x),$$

where x is the temperature, t is time, k is some positive constant, and A is the ambient temperature. See Figure 1.16 on the facing page for an example with $k = 0.3$ and $A = 5$.

Note the solution $x(t) = A$ (in the figure $x = 5$). We call these constant solutions the *equilibrium solutions*. The points on the x-axis where $f(x) = 0$ are called *critical points* of the differential equation (1.5). The point $x = A$ is a critical point. In fact, each critical point corresponds to an equilibrium solution.

Now, we want to determine what happens for other values of x that are not A. Based on the existence and uniqueness theorem in § 1.5 for first order differential equations, the fact that $k(A - x)$ and its partial derivative in x, $-k$, are continuous everywhere gives that solution curves can not cross. This means that since we know $x(t) = A$ is a solution, if a solution starts below $x(t) = A$, it must always stay there, and solutions that start above $x(t) = A$ will also stay there. For more information about what the solutions do, we'll need to look back at the equation and some sample solution curves.

Note also, by looking at the graph, that the solution $x = A$ is "stable" in that small perturbations in x do not lead to substantially different solutions as t grows. If we change the initial condition a little bit, then as $t \to \infty$ we get $x(t) \to A$. We call such a critical point *asymptotically stable*. In this simple example, it turns out that all solutions in fact go to A as $t \to \infty$. If there is a critical point where all nearby solutions move away from the critical point, we say it is *unstable*. If some nearby solutions go towards the critical point, and some others move away, then we say it is *semistable*. The final option is that solutions

nearby neither move towards nor away from the critical point, and these critical points are called *stable*.

The last of these options may seem strange at first, and that is because stable critical points are not possible for autonomous equations with one unknown function. If a solution does not move towards or away from a critical point, that means it doesn't move anywhere, and so is a critical point on its own. However, when we get to autonomous systems in § 4.7 and § 5.1, we will see that in two dimensions, this is possible (think of a circle that does not spiral into or away from the center point).

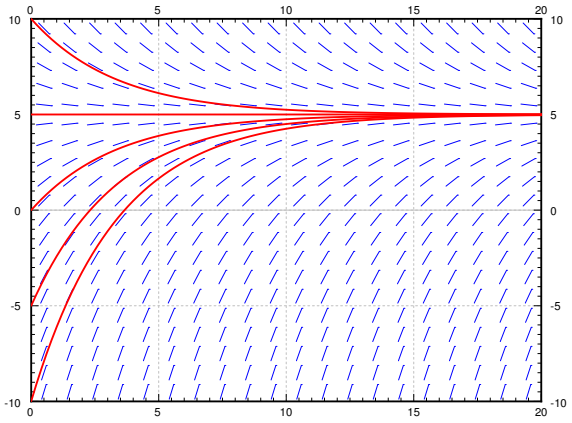


Figure 1.16: The slope field and some solutions of $x' = 0.3\,(5 - x)$.

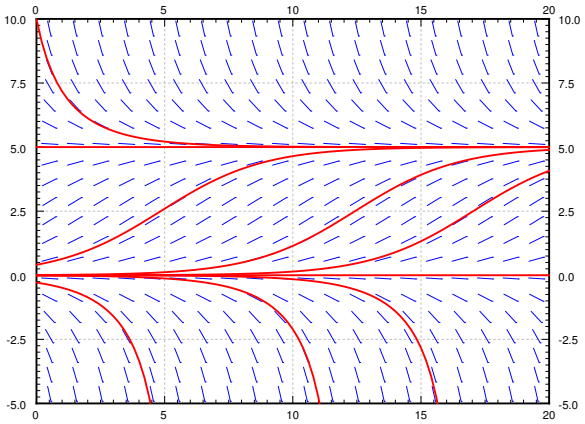


Figure 1.17: The slope field and some solutions of $x' = 0.1\,x\,(5 - x)$.

Consider now the *logistic equation*

$$\frac{dx}{dt} = kx(M - x),$$

for some positive k and M. This equation is commonly used to model population if we know the limiting population M, that is the maximum sustainable population. The logistic equation leads to less catastrophic predictions on world population than $x' = kx$. In the real world there is no such thing as negative population, but we will still consider negative x for the purposes of the math.

See Figure 1.17 for an example, $x' = 0.1x(5 - x)$. There are two critical points, $x = 0$ and $x = 5$. The critical point at $x = 5$ is asymptotically stable, while the critical point at $x = 0$ is unstable.

It is not necessary to find the exact solutions to talk about the long term behavior of the solutions. From the slope field above of $x' = 0.1x(5 - x)$, we see that

$$\lim_{t\to\infty} x(t) = \begin{cases} 5 & \text{if } x(0) > 0, \\ 0 & \text{if } x(0) = 0, \\ \text{DNE or } -\infty & \text{if } x(0) < 0. \end{cases}$$

Here DNE means "does not exist." From just looking at the slope field we cannot quite decide what happens if $x(0) < 0$. It could be that the solution does not exist for t all the way to ∞. Think of the equation $x' = x^2$; we have seen that solutions only exist for some finite period of time. Same can happen here. In our example equation above it turns out that the solution does not exist for all time, but to see that we would have to solve the equation. In any case, the solution does go to $-\infty$, but it may get there rather quickly.

If we are interested only in the long term behavior of the solution, we would be doing unnecessary work if we solved the equation exactly. We could draw the slope field, but it is easier to just look at the *phase diagram* or *phase line*, which is a simple way to visualize the behavior of autonomous equations. The phase line for this equation is visible in Figure 1.18. In this case there is one dependent variable x. We draw the x-axis, we mark all the critical points, and then we draw arrows in between. Since x is the dependent variable we draw the axis vertically, as it appears in the slope field diagrams above. If $f(x) > 0$, we draw an up arrow. If $f(x) < 0$, we draw a down arrow. To figure this out, we could just plug in some x between the critical points, $f(x)$ will have the same sign at all x between two critical points as long $f(x)$ is continuous. For example, $f(6) = -0.6 < 0$, so $f(x) < 0$ for $x > 5$, and the arrow above $x = 5$ is a down arrow. Next, $f(1) = 0.4 > 0$, so $f(x) > 0$ whenever $0 < x < 5$, and the arrow points up. Finally, $f(-1) = -0.6 < 0$ so $f(x) < 0$ when $x < 0$, and the arrow points down.

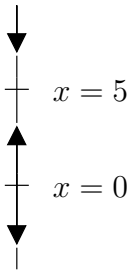


Figure 1.18: Phase line for the differential equation $x' = 0.1x(5 - x)$.

Armed with the phase diagram, it is easy to sketch the solutions approximately: As time t moves from left to right, the graph of a solution goes up if the arrow is up, and it goes down if the arrow is down.

***Exercise* 1.7.1:** *Try sketching a few solutions simply from looking at the phase diagram. Check with the preceding graphs to see if you are getting the types of curves that match the solutions.*

Once we draw the phase diagram, we can use it to classify critical points as asymptotically stable, semistable, or unstable based on whether the "arrows" point into or away from the critical point on each side. Two arrows in means that the critical point is asymptotically stable, two arrows away means unstable, and one in one out means semistable.

Example 1.7.1: Consider the autonomous differential equation

$$\frac{dx}{dt} = x(x-2)^2(x+3)(x-4) \tag{1.6}$$

Find all equilibrium solutions for this equation, and determine their stability. Draw a phase line and use this information to sketch some approximate solution curves.

Solution: This equation is already in factored form. This makes it simple to determine the equilibrium solutions as $x = 0$, $x = 2$, $x = -3$ and $x = 4$. In order to determine the stability of each critical point and draw the phase line, we need to plug in values surrounding these points to $f(x) = x(x-2)^2(x+3)(x-4)$. We can see that

$$\begin{aligned} f(-4) &= (-4)(-6)^2(-1)(-8) < 0, \\ f(-1) &= (-1)(-3)^2(2)(-5) > 0, \\ f(1) &= (1)(-1)^2(4)(-3) < 0, \\ f(3) &= (3)(1)^2(6)(-1) < 0, \\ f(5) &= (5)(3)^2(8)(1) > 0. \end{aligned}$$

This lets us draw the phase line and determine the stability of each critical point. Thus, we see that $x = -3$ is an unstable critical point, $x = 0$ is asymptotically stable, $x = 2$ is semistable, and $x = 4$ is unstable. A set of sample solution curves also validates these conclusions.

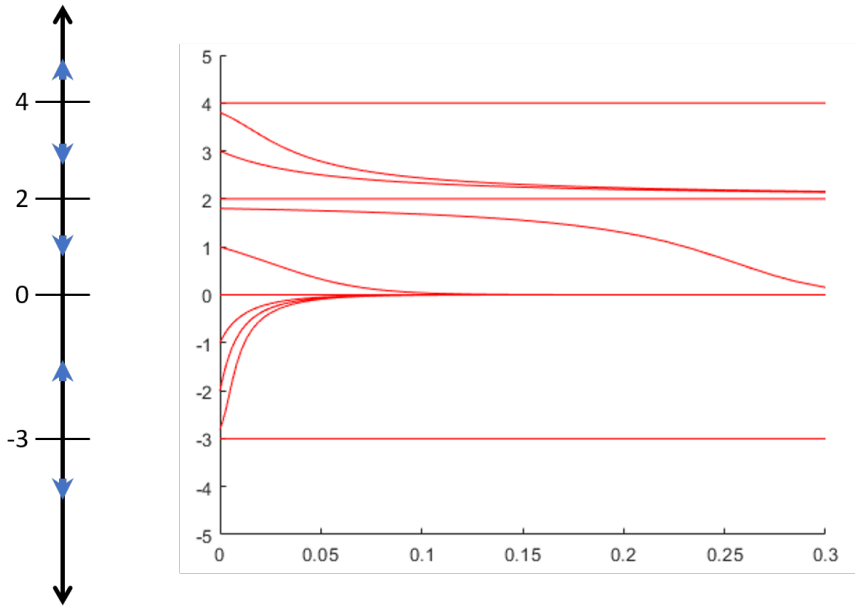


Figure 1.19: Phase line for the differential equation $\frac{dx}{dt} = x(x-2)^2(x+3)(x-4)$ *and a plot of some solutions to this equation.*

1.7.1 Concavity of Solutions

We can tell from the phase line for an autonomous equation when the solution will be increasing or decreasing. Is there any more we can learn about the shape of these graphs?

There is, and it comes from looking for the concavity, which is determined by the second derivative.

We can compute the second derivative

$$\frac{d^2x}{dt^2} = \frac{d}{dx}\left[\frac{dx}{dt}\right]$$

of our solution by noticing that $\frac{dx}{dt} = f(x)$. This function can be differentiated by the chain rule

$$\frac{d}{dt}f(x) = f'(x)\frac{dx}{dt} = f'(x)f(x).$$

So, the solution is concave up if $f'(x)f(x)$ is positive, and concave down if that is negative. Phrased another way, the solution is concave up if f and f' have the same sign, and it is concave down if f and f' have opposite signs.

Let's see what this looks like in action. Take the logistic equation $x' = 0.1x(5-x)$, whose solutions are plotted in Figure 1.17. Figure 1.20 shows the graph of $f(x)$ as a function of x for this scenario. When do f and f' have the same sign? Well, this happens when f is both positive and increasing, or negative and decreasing. This happens between 0 and the vertex, as well as for $x > 5$. The vertex here is at $x = 2.5$, and so we conlude that the solution should be concave up when x is on the intervals $(0, 2.5)$ and $(5, \infty)$, and be concave down otherwise. Looking back at Figure 1.17, this is exactly what we observe. All of the solutions between 0 and 5 seem to "flip over" to be concave down when x crosses 2.5.

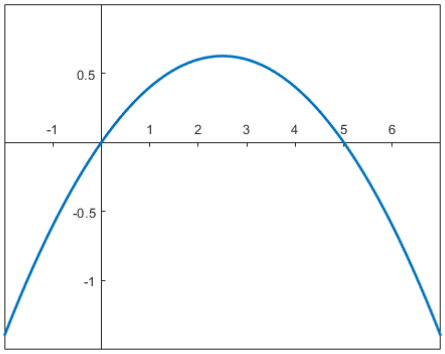


Figure 1.20: Plot of x vs. $f(x)$ for the differential equation $\frac{dx}{dt} = 0.1x(5-x)$.

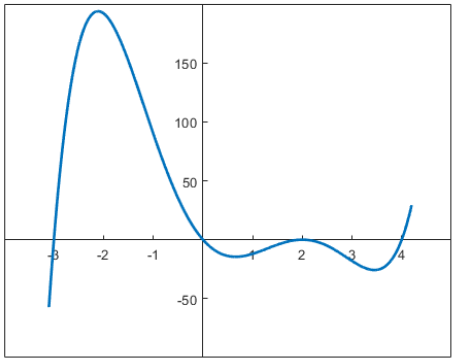


Figure 1.21: Plot of x vs. $f(x)$ for the differential equation $\frac{dx}{dt} = x(x-2)^2(x+3)(x-4)$.

The same can be seen for solutions to (1.6), even though we can't compute the extreme values explicitly. Figure 1.21 shows the graph of $f(x)$ vs. x for this situation. Between each pair of equilibrium solutions there is a critical point of f (in the Calculus 1 sense) where the derivative is zero, and at this point, the derivative changes sign, and since the function value does not change sign, the concavity of the solution to the differential equation flips at this point. Comparing this graph and these points where concavity shifts with the solutions drawn in Figure 1.19 again validates these results.

1.7.2 Exercises

***Exercise* 1.7.2:** *Consider* $x' = x^2$.

a) *Draw the phase diagram, find the critical points, and mark them asymptotically stable, semistable, or unstable.*

b) *Sketch typical solutions of the equation.*

c) *Find* $\lim_{t\to\infty} x(t)$ *for the solution with the initial condition* $x(0) = -1$.

***Exercise* 1.7.3:** *Consider* $x' = \sin x$.

a) *Draw the phase diagram for* $-4\pi \le x \le 4\pi$. *On this interval mark the critical points asymptotically stable, semistable, or unstable.*

b) *Sketch typical solutions of the equation.*

c) *Find* $\lim_{t\to\infty} x(t)$ *for the solution with the initial condition* $x(0) = 1$.

Exercise* 1.7.4: *Let* $x' = (x-1)(x-2)x^2$.

a) *Sketch the phase diagram and find critical points.*

b) *Classify the critical points.*

c) *If* $x(0) = 0.5$, *then find* $\lim_{t\to\infty} x(t)$.

***Exercise* 1.7.5:** *Let* $y' = (y-2)(y^2+1)(y+3)$. *Sketch a phase diagram for this differential equation. Find and classify all critical points. If* $y(0) = 0$, *what will happen to the solution as* $t \to \infty$*?*

***Exercise* 1.7.6:** *Find and classify all equilibrium solutions for the differential equation* $x' = (x-2)^2(x+1)(x+3)^3(x+2)$.

***Exercise* 1.7.7:** *Let* $y' = (y-3)(y+2)^2e^y$. *Sketch a phase diagram for this differential equation. Find and classify all critical points. If* $y(0) = 0$, *what will happen to the solution as* $t \to \infty$*?*

***Exercise* 1.7.8:** *Consider the DE* $\dfrac{dy}{dt} = y^5 - 3y^4 + 3y^3 - y^2$. *Find and classify all equilibrium solutions of this DE. Then sketch a representative selection of solution curves.*

Exercise* 1.7.9: *Let* $x' = e^{-x}$.

a) *Find and classify all critical points.* b) *Find* $\lim_{t\to\infty} x(t)$ *given any initial condition.*

***Exercise* 1.7.10:** *Suppose* $f(x)$ *is positive for* $0 < x < 1$, *it is zero when* $x = 0$ *and* $x = 1$, *and it is negative for all other* x.

a) *Draw the phase diagram for $x' = f(x)$, find the critical points, and mark them asymptotically stable, semistable, or unstable.*

b) *Sketch typical solutions of the equation.*

c) *Find $\lim_{t\to\infty} x(t)$ for the solution with the initial condition $x(0) = 0.5$.*

Exercise* 1.7.11*: *Suppose $\frac{dx}{dt} = (x-\alpha)(x-\beta)$ for two numbers $\alpha < \beta$.*

a) *Find the critical points, and classify them.*

For b), c), d), find $\lim_{t\to\infty} x(t)$ based on the phase diagram.

b) $x(0) < \alpha$, c) $\alpha < x(0) < \beta$, d) $\beta < x(0)$.

***Exercise* 1.7.12:** *A disease is spreading through the country. Let x be the number of people infected. Let the constant S be the number of people susceptible to infection. The infection rate $\frac{dx}{dt}$ is proportional to the product of already infected people, x, and the number of susceptible but uninfected people, $S - x$.*

a) *Write down the differential equation.*

b) *Supposing $x(0) > 0$, that is, some people are infected at time $t = 0$, what is $\lim_{t\to\infty} x(t)$.*

c) *Does the solution to part b) agree with your intuition? Why or why not?*

1.8 Bifurcation diagrams

Attribution: [JL], §1.6.

Learning Objectives

After this section, you will be able to:

- Draw and analyze bifurcation diagrams for autonomous equations with parameter.

An extension of the topic of autonomous equation is *autonomous equations with parameter*. The idea is that we have a differential equation that has no explicit dependence on time, but does have a dependence on an outside parameter, which is a constant set by the physical situation. In terms of physical problems, this parameter will tend to be something that we can adjust to change how the differential equation behaves. For example, in a logistic differential equation

$$\frac{dx}{dt} = ax(K - x)$$

either the a or the K (or both) could be adjustable parameters. For a given value of the parameter, the differential equation behaves like a standard autonomous differential equation, but we can get different properties of this equation for different values of the parameter.

Definition 1.8.1

An *autonomous equation with parameter* α is a differential equation of the form

$$\frac{dx}{dt} = f_\alpha(x)$$

where, for every value of α, $f_\alpha(x)$ is a function of the single variable x.

Later, we will want to view $f_\alpha(x)$ as a two-variable function of x and α, but for now, we want to think about it as a function of just x for a fixed value of α. We want to be able to analyze what happens to this equation for different values of α. Since it is an autonomous equation, we can do this using phase lines. This will be easiest to see through an example.

Example 1.8.1: Consider the differential equation

$$\frac{dx}{dt} = x(x^2 - \alpha),$$

which fits the description of an autonomous equation with parameter α. Describe what happens in this differential equation for $\alpha = -4$, $\alpha = 0$, and $\alpha = 1$.

Solution: We can draw a phase line for $\alpha = -4$, $\alpha = 0$ and $\alpha = 1$. It is clear that something happens with this equation between $\alpha = -4$ and $\alpha = 1$. We go from having only one equilibrium solution at $\alpha = -4$ to having three equilibrium solutions at $\alpha = 1$. In addition, the solution at $y = 0$ is unstable for $\alpha = -4$, while it is asymptotically stable for $\alpha = 1$. If we want to figure out when this change happens, we'll need a better way to analyze this problem.

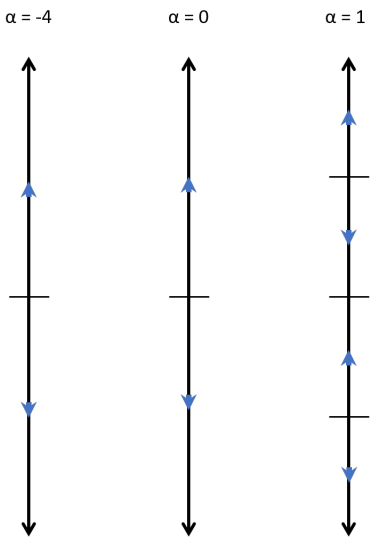


Figure 1.22: Phase lines for the differential equation $\frac{dx}{dt} = x(x^2 - \alpha)$ *for* $\alpha = -4,\ 0,\ 1$.

How can we better approach this problem? The idea is to think about when the solution to the differential equation will be increasing or decreasing as a function of the two variables α and x. Based on the structure of the differential equation, the solution will be increasing when the function $f_\alpha(x)$ is positive and will be decreasing when $f_\alpha(x)$ is negative. Since a phase line is a plot of this information for a given value of α, we essentially want to plot all of these phase lines on a two-dimensional graph. This graph is called a *bifurcation diagram*. Figure 1.23 shows a bifurcation diagram for the example $\frac{dx}{dt} = x(x^2 - \alpha)$.

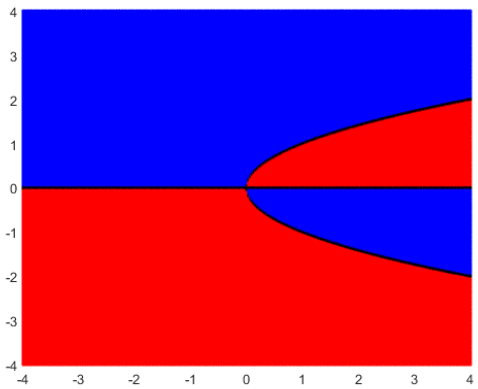


Figure 1.23: Bifurcation Diagram for the differential equation $\frac{dx}{dt} = x(x^2 - \alpha)$. *In this figure, a blue region means the solution will be increasing and red indicates decreasing.*

Within this picture, we can see all of our phase lines from before, because at any value of α, taking the vertical slice of this graph at that value, we get the phase line. If we want to consider $\alpha = -4$, then we can look above the horizontal coordinate -4, and that will give us the phase line for $\alpha = -4$. The same goes for any other value of α we want to consider. For instance, we can also see that for any $\alpha \leq 0$, there will be one equilibrium solution, and for $\alpha > 0$ there are three equilibrium solutions, indicated by the three black curves above each of those α values.

From this, we can see that the point at which the behavior changes is $\alpha = 0$. Thus, for this problem $\alpha = 0$ is called the *bifurcation point*. This is defined to be the value of the parameter for which the overall behavior of the equation changes. This can be a change in the number of equilibrium solutions, the stability of these equilibrium solutions, or both. For this particular example, we have both of these. We go from 1 equilibrium solution to 3, and the solution at $y = 0$ changes in stability. This type of bifurcation is called a "pitchfork bifurcation" based on the shape of the equilibrium solutions near the bifurcation point.

Another example of a bifurcation of a different form can be seen in the example of the logistic equation with harvesting. Suppose an alien race really likes to eat humans. They keep a planet with humans on it and harvest the humans at a rate of h million humans per year. Suppose x is the number of humans in millions on the planet and t is time in years. Let M be the limiting population when no harvesting is done. The number $k > 0$ is a constant depending on how fast humans multiply. Our equation becomes

$$\frac{dx}{dt} = kx(M - x) - h.$$

In this setup, M and k are fixed values, and the parameter that is being adjusted for this equation is h. We expand the right-hand side and set it to zero.

$$kx(M - x) - h = -kx^2 + kMx - h = 0.$$

Solving for the critical points using the quadratic formula, let us call them A and B, we get

$$A = \frac{kM + \sqrt{(kM)^2 - 4hk}}{2k}, \qquad B = \frac{kM - \sqrt{(kM)^2 - 4hk}}{2k}.$$

***Exercise* 1.8.1:** *Sketch a phase diagram for different possibilities. Note that these possibilities are $A > B$, or $A = B$, or A and B both complex (i.e. no real solutions). Hint: Fix some simple k and M and then vary h.*

Example 1.8.2: For example, let $M = 8$ and $k = 0.1$. What happens for different values of h in this situation?

Solution: When $h = 1$, then A and B are distinct and positive. The slope field we get is in Figure 1.24 on the following page. As long as the population starts above B, which is approximately 1.55 million, then the population will not die out. It will in fact tend towards $A \approx 6.45$ million. If ever some catastrophe happens and the population drops below B, humans will die out, and the fast food restaurant serving them will go out of business.

When $h = 1.6$, then $A = B = 4$. There is only one critical point and it is semistable. When the population starts above 4 million it will tend towards 4 million. If it ever drops

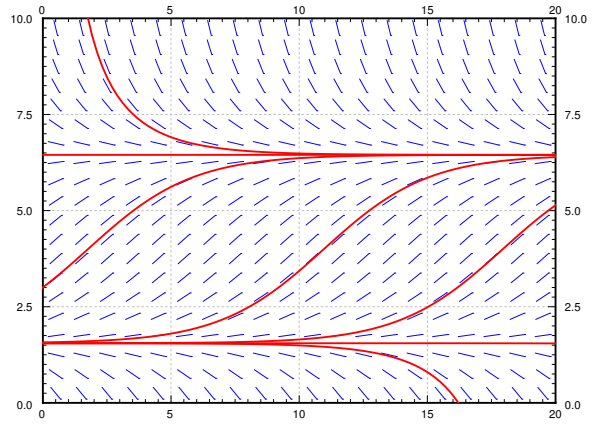


Figure 1.24: The slope field and some solutions of $x' = 0.1\,x\,(8 - x) - 1$.

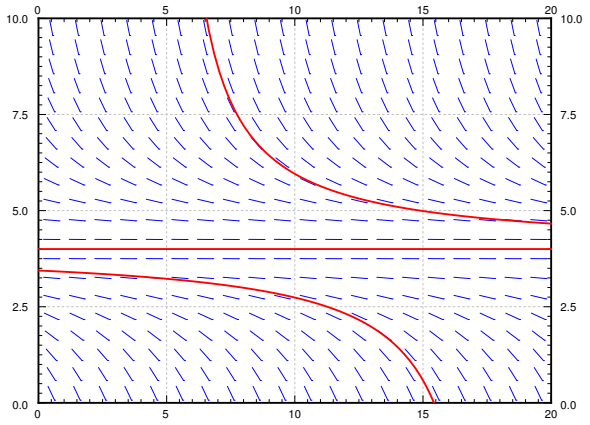


Figure 1.25: The slope field and some solutions of $x' = 0.1\,x\,(8 - x) - 1.6$.

below 4 million, humans will die out on the planet. This scenario is not one that we (as the human fast food proprietor) want to be in. A small perturbation of the equilibrium state and we are out of business. There is no room for error. See Figure 1.25.

Finally if we are harvesting at 2 million humans per year, there are no critical points. The population will always plummet towards zero, no matter how well stocked the planet starts. See Figure 1.26.

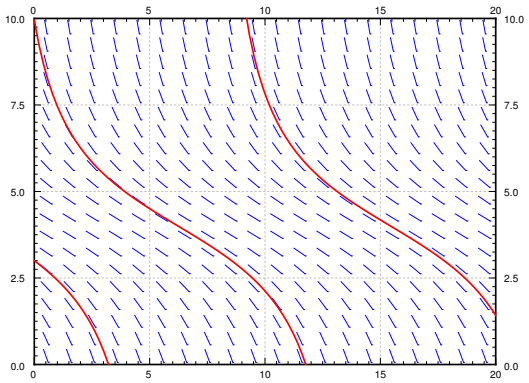


Figure 1.26: The slope field and some solutions of $x' = 0.1\,x\,(8 - x) - 2$.

All of these can also be seen from the bifurcation diagram, which is drawn in Figure 1.27 on the next page. The values A and B discussed above represent the upper and lower branches of the parabola in the figure. For any $h > 1.6$, there are no equilibrium solutions and the phase line is entirely decreasing, meaning the solution will converge to zero no matter what. For $h < 1.6$, there are two equilibrium solutions, with the top one asymptotically stable and the bottom one unstable. At $h = 1.6$ is where the bifurcation point occurs for this example. This is an example of a "saddle-node" bifurcation, as the two equilibrium solutions collide with each other at the bifurcation point and disappear.

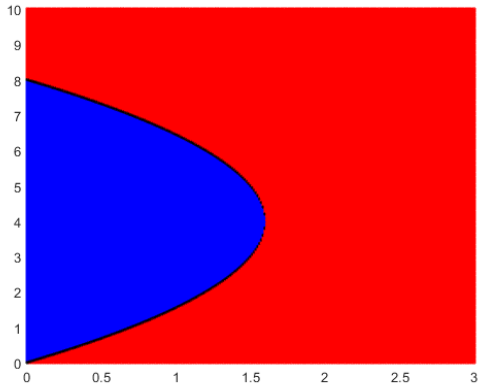


Figure 1.27: Bifurcation diagram for the differential equation $x' = 0.1\,x\,(8 - x) - h$.

Another way to visualize this situation is by plotting the function $f_\alpha(x)$ for the different values of α. The places where this function is zero give the equilibrium solutions, and we can determine *bifurcation values* by looking for where the zeros of this function change behavior. For this particular example, the graphs of $f_\alpha(x)$ are drawn in Figure 1.28.

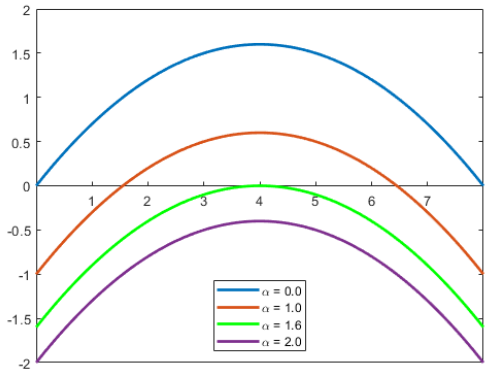


Figure 1.28: Graph of $f_\alpha(x) = 0.1x(8 - x) - \alpha$ *for* $\alpha = 0,\ 1.0,\ 1.6,\ 2.0$.

The values of α we are looking for are those where the number and types of zeros change for the function $f_\alpha(x)$. In this figure, we see that for $\alpha < 1.6$, the parabola crosses the x axis twice, resulting in two zeros and two equilibrium solutions. For $\alpha = 1.6$, there is one (double) root, and for $\alpha > 1.6$, there are no equilibrium solutions, and the function $f_\alpha(x)$ is always negative. Since the number of roots/zeros changes at $\alpha = 1.6$, that means that 1.6 is the bifurcation point for this equation. We can also see this from the equation, since the equilibrium solutions are determined by the values of x where

$$0.1x(8 - x) - \alpha = 0 \qquad \text{or} \qquad -0.1x^2 + 0.8x - \alpha = 0$$

which can be found by the quadratic formula

$$x = \frac{0.8 \pm \sqrt{0.64 - 4(0.1)(\alpha)}}{0.2}.$$

Roots to this equation do not exist (because they are complex) if $0.64 - 0.4\alpha < 0$, or $\alpha > 1.6$.

1.8.1 Exercises

***Exercise* 1.8.2:** *Start with the logistic equation $\frac{dx}{dt} = kx(M - x)$. Suppose we modify our harvesting. That is we will only harvest an amount proportional to current population. In other words, we harvest hx per unit of time for some $h > 0$ (Similar to earlier example with h replaced with hx).*

a) Construct the differential equation.

b) Show that if $kM > h$, then the equation is still logistic.

c) What happens when $kM < h$?

Exercise* 1.8.3: *Assume that a population of fish in a lake satisfies $\frac{dx}{dt} = kx(M - x)$. Now suppose that fish are continually added at A fish per unit of time.*

a) Find the differential equation for x. *b) What is the new limiting population?*

***Exercise* 1.8.4:** *Consider the differential equation with parameter α given by $y' = y(y - \alpha + 1)$.*

a) Sketch a phase diagram for this differential equation with $\alpha = -3$, $\alpha = 1$, and $\alpha = 3$.

b) Draw a bifurcation diagram for this differential equation with parameter.

c) What is the bifurcation point for this equation? What changes when α passes over the bifurcation point?

***Exercise* 1.8.5:** *Consider the differential equation with parameter α given by $y' = y^2(y^2 - \alpha)$.*

a) Sketch a phase diagram for this differential equation with $\alpha = -3$, $\alpha = 0$, and $\alpha = 3$.

b) Draw a bifurcation diagram for this differential equation with parameter.

c) What is the bifurcation point for this equation? What changes when α passes over the bifurcation point?

***Exercise* 1.8.6:** *Consider the differential equation with parameter α given by $y' = y(\alpha - y)$.*

a) Sketch a phase diagram for this differential equation with $\alpha = -3$, $\alpha = 0$, and $\alpha = 3$.

b) Draw a bifurcation diagram for this differential equation with parameter.

c) What is the bifurcation point for this equation? What changes when α passes over the bifurcation point?

1.9 Exact equations

Attribution: [JL], §1.8.

Learning Objectives

After this section, you will be able to:

- Determine if a first order differential equation is exact,
- Find the general solution to an exact equation,
- Solve initial value problems for exact equations, and
- Use integrating factors to make some non-exact equations exact in order to solve them.

Another type of equation that comes up quite often in physics and engineering is an *exact equation*. Suppose $F(x, y)$ is a function of two variables, which we call the *potential function*. The naming should suggest potential energy, or electric potential. Exact equations and potential functions appear when there is a conservation law at play, such as conservation of energy. Let us make up a simple example. Let

$$F(x, y) = x^2 + y^2.$$

We are interested in the lines of constant energy, that is lines where the energy is conserved; we want curves where $F(x, y) = C$, for some constant C, since F represents the energy of the system. In our example, the curves $x^2 + y^2 = C$ are circles. See Figure 1.29.

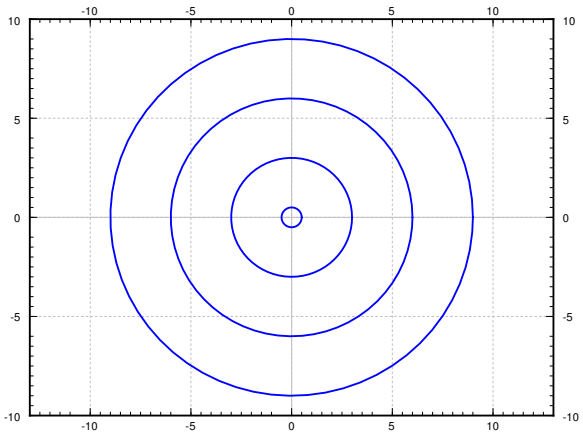


Figure 1.29: Solutions to $F(x, y) = x^2 + y^2 = C$ for various C.

We take the *total derivative* of F:

$$dF = \frac{\partial F}{\partial x}dx + \frac{\partial F}{\partial y}dy.$$

For convenience, we will make use of the notation of $F_x = \frac{\partial F}{\partial x}$ and $F_y = \frac{\partial F}{\partial y}$. In our example,

$$dF = 2x\,dx + 2y\,dy.$$

We apply the total derivative to $F(x, y) = C$, to find the differential equation $dF = 0$. The differential equation we obtain in such a way has the form

$$M\,dx + N\,dy = 0, \qquad \text{or} \qquad M + N\,\frac{dy}{dx} = 0.$$

Definition 1.9.1

An equation of the form

$$M(x,y) + N(x,y)\,\frac{dy}{dx} = 0$$

is called *exact* if it was obtained as $dF = 0$ for some potential function F.

In our simple example, we obtain the equation

$$2x\,dx + 2y\,dy = 0, \qquad \text{or} \qquad 2x + 2y\,\frac{dy}{dx} = 0.$$

Since we obtained this equation by differentiating $x^2 + y^2 = C$, the equation is exact. We often wish to solve for y in terms of x. In our example,

$$y = \pm\sqrt{C^2 - x^2}.$$

An interpretation of the setup is that at each point in the plane $\vec{v} = (M, N)$ is a vector, that is, a direction and a magnitude. As M and N are functions of (x, y), we have a *vector field*. The particular vector field $\vec{v}$ that comes from an exact equation is a so-called *conservative vector field*, that is, a vector field that comes with a potential function $F(x, y)$, such that

$$\vec{v} = \left(\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}\right).$$

This is something that you may have seen in your Calculus 3 course, and if so, the process for solving exact equations is basically identical to the process of finding a potential function for a conservative vector field. The physical interpretation of conservative vector fields is as follows. Let γ be a path in the plane starting at (x_1, y_1) and ending at (x_2, y_2). If we think of $\vec{v}$ as force, then the work required to move along γ is

$$\int_\gamma \vec{v}(\vec{r}) \cdot d\vec{r} = \int_\gamma M\,dx + N\,dy = F(x_2, y_2) - F(x_1, y_1).$$

That is, the work done only depends on endpoints, that is where we start and where we end. For example, suppose F is gravitational potential. The derivative of F given by $\vec{v}$ is the gravitational force. What we are saying is that the work required to move a heavy box from the ground floor to the roof only depends on the change in potential energy. That is, the work done is the same no matter what path we took; if we took the stairs or the elevator. Although if we took the elevator, the elevator is doing the work for us. The curves $F(x, y) = C$ are those where no work need be done, such as the heavy box sliding along without accelerating or breaking on a perfectly flat roof, on a cart with incredibly well oiled wheels. Effectively, an exact equation is a conservative vector field, and the implicit solution of this equation is the potential function.

1.9.1 Solving exact equations

Now you, the reader, should ask: Where did we solve a differential equation? Well, in applications we generally know M and N, but we do not know F. That is, we may have just

started with $2x + 2y\frac{dy}{dx} = 0$, or perhaps even

$$x + y\frac{dy}{dx} = 0.$$

It is up to us to find some potential F that works. Many different F will work; adding a constant to F does not change the equation. Once we have a potential function F, the equation $F\big(x, y(x)\big) = C$ gives an implicit solution of the ODE.

Example 1.9.1: Let us find the general solution to $2x + 2y\frac{dy}{dx} = 0$. Forget we knew what F was.

Solution: If we know that this is an exact equation, we start looking for a potential function F. We have $M = 2x$ and $N = 2y$. If F exists, it must be such that $F_x(x, y) = 2x$. Integrate in the x variable to find

$$F(x, y) = x^2 + A(y), \tag{1.7}$$

for some function $A(y)$. The function A is the "constant of integration", though it is only constant as far as x is concerned, and may still depend on y. Now differentiate (1.7) in y and set it equal to N, which is what F_y is supposed to be:

$$2y = F_y(x, y) = A'(y).$$

Integrating, we find $A(y) = y^2$. We could add a constant of integration if we wanted to, but there is no need. We found $F(x, y) = x^2 + y^2$. Next for a constant C, we solve

$$F\big(x, y(x)\big) = C.$$

for y in terms of x. In this case, we obtain $y = \pm\sqrt{C^2 - x^2}$ as we did before. □

***Exercise* 1.9.1:** *Why did we not need to add a constant of integration when integrating $A'(y) = 2y$? Add a constant of integration, say 3, and see what F you get. What is the difference from what we got above, and why does it not matter?*

In the previous example, you may have also noticed that the equation $2x + 2y\frac{dy}{dx} = 0$ is separable, and we could have solved it via that method as well. This is not a coincidence, as every separable equation is exact (see Exercise 1.9.15 for the details) but there are many exact equations that are not separable, which we will see throughout the examples here.

The procedure, once we know that the equation is exact, is:

(i) Integrate $F_x = M$ in x resulting in $F(x, y) = \text{something} + A(y)$.

(ii) Differentiate this F in y, and set that equal to N, so that we may find $A(y)$ by integration.

The procedure can also be done by first integrating in y and then differentiating in x. Pretty easy huh? Let's try this again.

Example 1.9.2: Consider now $2x + y + xy\frac{dy}{dx} = 0$.

Solution: OK, so $M = 2x + y$ and $N = xy$. We try to proceed as before. Suppose F exists. Then $F_x(x, y) = 2x + y$. We integrate:

$$F(x, y) = x^2 + xy + A(y)$$

for some function $A(y)$. Differentiate in y and set equal to N:

$$N = xy = F_y(x, y) = x + A'(y).$$

But there is no way to satisfy this requirement! The function xy cannot be written as x plus a function of y. The equation is not exact; no potential function F exists. □

Is there an easier way to check for the existence of F, other than failing in trying to find it? Turns out there is. Suppose $M = F_x$ and $N = F_y$. Then as long as the second derivatives are continuous,

$$\frac{\partial M}{\partial y} = \frac{\partial^2 F}{\partial y \partial x} = \frac{\partial^2 F}{\partial x \partial y} = \frac{\partial N}{\partial x}.$$

Let us state it as a theorem. Usually this is called the Poincaré Lemma*.

Theorem 1.9.1 (Poincaré)

If M and N are continuously differentiable functions of (x, y), and $\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}$, then near any point there is a function $F(x, y)$ such that $M = \frac{\partial F}{\partial x}$ and $N = \frac{\partial F}{\partial y}$.

The theorem doesn't give us a global F defined everywhere. In general, we can only find the potential locally, near some initial point. By this time, we have come to expect this from differential equations.

Let us return to the example above where $M = 2x + y$ and $N = xy$. Notice $M_y = 1$ and $N_x = y$, which are clearly not equal. The equation is not exact.

Example 1.9.3: Solve

$$\frac{dy}{dx} = \frac{-2x - y}{x - 1}, \qquad y(0) = 1.$$

Solution: We write the equation as

$$(2x + y) + (x - 1)\frac{dy}{dx} = 0,$$

so $M = 2x + y$ and $N = x - 1$. Then

$$M_y = 1 = N_x.$$

The equation is exact. Integrating M in x, we find

$$F(x, y) = x^2 + xy + A(y).$$

*Named for the French polymath Jules Henri Poincaré (1854–1912).

Differentiating in y and setting to N, we find

$$x - 1 = x + A'(y).$$

So $A'(y) = -1$, and $A(y) = -y$ will work. Take $F(x, y) = x^2 + xy - y$. We wish to solve $x^2 + xy - y = C$. First let us find C. As $y(0) = 1$ then $F(0, 1) = C$. Therefore $0^2 + 0 \times 1 - 1 = C$, so $C = -1$. Now we solve $x^2 + xy - y = -1$ for y to get

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

Example 1.9.4: Solve

$$-\frac{y}{x^2 + y^2}dx + \frac{x}{x^2 + y^2}dy = 0, \qquad y(1) = 2.$$

Solution: We leave to the reader to check that $M_y = N_x$.

This vector field (M, N) is not conservative if considered as a vector field of the entire plane minus the origin. The problem is that if the curve γ is a circle around the origin, say starting at $(1, 0)$ and ending at $(1, 0)$ going counterclockwise, then if F existed we would expect

$$0 = F(1, 0) - F(1, 0) = \int_\gamma F_x\, dx + F_y\, dy = \int_\gamma \frac{-y}{x^2 + y^2}\, dx + \frac{x}{x^2 + y^2}\, dy = 2\pi.$$

That is nonsense! We leave the computation of the path integral to the interested reader, or you can consult your multivariable calculus textbook. So there is no potential function F defined everywhere outside the origin $(0, 0)$.

If we think back to the theorem, it does not guarantee such a function anyway. It only guarantees a potential function locally, that is only in some region near the initial point. As $y(1) = 2$ we start at the point $(1, 2)$. Considering $x > 0$ and integrating M in x or N in y, we find

$$F(x, y) = \arctan\left(y/x\right).$$

The implicit solution is $\arctan\left(y/x\right) = C$. Solving, $y = \tan(C)x$. That is, the solution is a straight line. Solving $y(1) = 2$ gives us that $\tan(C) = 2$, and so $y = 2x$ is the desired solution. See Figure 1.30 on the following page, and note that the solution only exists for $x > 0$.

Example 1.9.5: Solve

$$x^2 + y^2 + 2y(x + 1)\frac{dy}{dx} = 0.$$

Solution: The reader should check that this equation is exact. Let $M = x^2 + y^2$ and $N = 2y(x + 1)$. We follow the procedure for exact equations

$$F(x, y) = \frac{1}{3}x^3 + xy^2 + A(y),$$

and

$$2y(x + 1) = 2xy + A'(y).$$

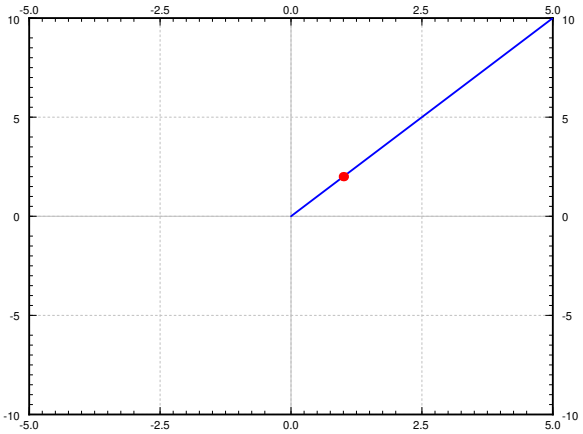


Figure 1.30: Solution to $-\frac{y}{x^2+y^2}dx + \frac{x}{x^2+y^2}dy = 0$, $y(1) = 2$, *with initial point marked.*

Therefore $A'(y) = 2y$ or $A(y) = y^2$ and $F(x, y) = \frac{1}{3}x^3 + xy^2 + y^2$. We try to solve $F(x, y) = C$. We easily solve for y^2 and then just take the square root:

$$y^2 = \frac{C - (1/3)x^3}{x + 1}, \qquad \text{so} \qquad y = \pm\sqrt{\frac{C - (1/3)x^3}{x + 1}}.$$

When $x = -1$, the term in front of $\frac{dy}{dx}$ vanishes. You can also see that our solution is not valid in that case. However, one could in that case try to solve for x in terms of y starting from the implicit solution $\frac{1}{3}x^3 + xy^2 + y^2 = C$. The solution is somewhat messy and we leave it as implicit. ⌟

1.9.2 Integrating factors

Sometimes an equation $M\,dx + N\,dy = 0$ is not exact, but it can be made exact by multiplying with a function $u(x, y)$. That is, perhaps for some nonzero function $u(x, y)$,

$$u(x, y)M(x, y)\,dx + u(x, y)N(x, y)\,dy = 0$$

is exact. Any solution to this new equation is also a solution to $M\,dx + N\,dy = 0$.

In fact, a linear equation

$$\frac{dy}{dx} + p(x)y = f(x), \qquad \text{or} \qquad \big(p(x)y - f(x)\big)\,dx + dy = 0$$

is always such an equation. Let $r(x) = e^{\int p(x)\,dx}$ be the integrating factor for a linear equation. Multiply the equation by $r(x)$ and write it in the form of $M + N\frac{dy}{dx} = 0$.

$$r(x)p(x)y - r(x)f(x) + r(x)\frac{dy}{dx} = 0.$$

Then $M = r(x)p(x)y - r(x)f(x)$, so $M_y = r(x)p(x)$, while $N = r(x)$, so $N_x = r'(x) = r(x)p(x)$. In other words, we have an exact equation. Integrating factors for linear functions are just a special case of integrating factors for exact equations.

But how do we find the integrating factor u? Well, given an equation

$$M\,dx + N\,dy = 0,$$

u should be a function such that

$$\frac{\partial}{\partial y}\big[uM\big] = u_y M + uM_y = \frac{\partial}{\partial x}\big[uN\big] = u_x N + uN_x.$$

Therefore,

$$(M_y - N_x)u = u_x N - u_y M.$$

At first it may seem we replaced one differential equation by another. True, but all hope is not lost.

A strategy that often works is to look for a u that is a function of x alone, or a function of y alone. If u is a function of x alone, that is $u(x)$, then we write $u'(x)$ instead of u_x, and u_y is just zero. Then

$$\frac{M_y - N_x}{N}u = u'.$$

In particular, $\frac{M_y - N_x}{N}$ ought to be a function of x alone (not depend on y). If so, then we have a linear equation

$$u' - \frac{M_y - N_x}{N}u = 0.$$

Letting $p(x) = \frac{M_y - N_x}{N}$, we solve using the standard integrating factor method, to find $u(x) = Ce^{\int p(x)\,dx}$. The constant in the solution is not relevant, we need any nonzero solution, so we take $C = 1$. Then $u(x) = e^{\int p(x)\,dx}$ is the integrating factor.

Similarly we could try a function of the form $u(y)$. Then

$$\frac{M_y - N_x}{M}u = -u'.$$

In particular, $\frac{M_y - N_x}{M}$ ought to be a function of y alone. If so, then we have a linear equation

$$u' + \frac{M_y - N_x}{M}u = 0.$$

Letting $q(y) = \frac{M_y - N_x}{M}$, we find $u(y) = Ce^{-\int q(y)\,dy}$. We take $C = 1$. So $u(y) = e^{-\int q(y)\,dy}$ is the integrating factor.

Example 1.9.6: Solve

$$\frac{x^2 + y^2}{x + 1} + 2y\frac{dy}{dx} = 0.$$

Solution: Let $M = \frac{x^2+y^2}{x+1}$ and $N = 2y$. Compute

$$M_y - N_x = \frac{2y}{x + 1} - 0 = \frac{2y}{x + 1}.$$

As this is not zero, the equation is not exact. We notice

$$P(x) = \frac{M_y - N_x}{N} = \frac{2y}{x+1}\frac{1}{2y} = \frac{1}{x+1}$$

is a function of x alone. We compute the integrating factor

$$e^{\int P(x)\,dx} = e^{\ln|x+1|} = |x+1|.$$

Assuming that we want to look at $x > -1$, we multiply our given equation by $(x+1)$ to obtain

$$x^2 + y^2 + 2y(x+1)\frac{dy}{dx} = 0,$$

which is an exact equation that we solved in Example 1.9.5. The solution was

$$y = \pm\sqrt{\frac{C - (1/3)x^3}{x+1}}.$$

If, instead, we had wanted a solution with $x < -1$, we would have needed to multiply by $-(x+1)$, which would have given a very similar result. ⌟

Example 1.9.7: Solve

$$y^2 + (xy+1)\frac{dy}{dx} = 0.$$

Solution: First compute

$$M_y - N_x = 2y - y = y.$$

As this is not zero, the equation is not exact. We observe

$$Q(y) = \frac{M_y - N_x}{M} = \frac{y}{y^2} = \frac{1}{y}$$

is a function of y alone. We compute the integrating factor

$$e^{-\int Q(y)\,dy} = e^{-\ln y} = \frac{1}{y}.$$

Therefore we look at the exact equation

$$y + \frac{xy+1}{y}\frac{dy}{dx} = 0.$$

The reader should double check that this equation is exact. We follow the procedure for exact equations

$$F(x,y) = xy + A(y),$$

and

$$\frac{xy+1}{y} = x + \frac{1}{y} = x + A'(y). \tag{1.8}$$

Consequently $A'(y) = \frac{1}{y}$ or $A(y) = \ln y$. Thus $F(x, y) = xy + \ln y$. It is not possible to solve $F(x, y) = C$ for y in terms of elementary functions, so let us be content with the implicit solution:

$$xy + \ln y = C.$$

We are looking for the general solution and we divided by y above. We should check what happens when $y = 0$, as the equation itself makes perfect sense in that case. We plug in $y = 0$ to find the equation is satisfied. So $y(x) = 0$ is also a solution. □

1.9.3 Exercises

Exercise 1.9.2: *Solve the following exact equations, implicit general solutions will suffice:*

a) $(2xy + x^2)\,dx + (x^2 + y^2 + 1)\,dy = 0$

b) $x^5 + y^5 \frac{dy}{dx} = 0$

c) $e^x + y^3 + 3xy^2 \frac{dy}{dx} = 0$

d) $(x + y)\cos(x) + \sin(x) + \sin(x)y' = 0$

Exercise 1.9.3:* *Solve the following exact equations, implicit general solutions will suffice:*

a) $\cos(x) + ye^{xy} + xe^{xy}y' = 0$

b) $(2x + y)\,dx + (x - 4y)\,dy = 0$

c) $e^x + e^y \frac{dy}{dx} = 0$

d) $(3x^2 + 3y)\,dx + (3y^2 + 3x)\,dy = 0$

Exercise 1.9.4: *Solve the differential equation* $(2ye^{2xy} - 2x) + (2xe^{2xy} + \cos(y))y' = 0$

Exercise 1.9.5: *Solve the differential equation* $(-y\sin(xy) - 2xe^{x^2}) + (-x\sin(xy) + 1)y' = 0$

Exercise 1.9.6: *Solve the differential equation* $(2x + 3y\sin(xy)) + (3x\sin(xy) - e^y)y' = 0$ *with* $y(2) = 0$.

Exercise 1.9.7: *Solve the differential equation* $x + yy' = 0$ *with* $y(0) = 8$. *Write this as an explicit function and determine the interval of* x *values where the solution is valid.*

Exercise 1.9.8: *Solve the differential equation* $2x - 2 + (8y + 16)y' = 0$ *with* $y(2) = 0$. *Write this as an explicit function and determine the interval of* x *values where the solution is valid.*

Exercise 1.9.9: *Find the integrating factor for the following equations making them into exact equations:*

a) $e^{xy}\,dx + \frac{y}{x}e^{xy}\,dy = 0$

b) $\frac{e^x + y^3}{y^2}\,dx + 3x\,dy = 0$

c) $4(y^2 + x)\,dx + \frac{2x + 2y^2}{y}\,dy = 0$

d) $2\sin(y)\,dx + x\cos(y)\,dy = 0$

Exercise 1.9.10:* *Find the integrating factor for the following equations making them into exact equations:*

a) $\frac{1}{y}\,dx + 3y\,dy = 0$

b) $dx - e^{-x-y}\,dy = 0$

c) $\left(\frac{\cos(x)}{y^2} + \frac{1}{y}\right)dx + \frac{x}{y^2}\,dy = 0$

d) $\left(2y + \frac{y^2}{x}\right)dx + (2y + x)\,dy = 0$

***Exercise* 1.9.11:** *Suppose you have an equation of the form:* $f(x) + g(y)\frac{dy}{dx} = 0$.

a) *Show it is exact.*

b) *Find the form of the potential function in terms of* f *and* g.

***Exercise* 1.9.12:** *Suppose that we have the equation* $f(x)\,dx - dy = 0$.

a) *Is this equation exact?*

b) *Find the general solution using a definite integral.*

***Exercise* 1.9.13:** *Find the potential function* $F(x, y)$ *of the exact equation* $\frac{1+xy}{x}\,dx + \left(1/y + x\right)dy = 0$ *in two different ways.*

a) *Integrate* M *in terms of* x *and then differentiate in* y *and set to* N.

b) *Integrate* N *in terms of* y *and then differentiate in* x *and set to* M.

***Exercise* 1.9.14:** *A function* $u(x, y)$ *is said to be a* harmonic function *if* $u_{xx} + u_{yy} = 0$.

a) *Show if* u *is harmonic,* $-u_y\,dx + u_x\,dy = 0$ *is an exact equation. So there exists (at least locally) the so-called* harmonic conjugate *function* $v(x, y)$ *such that* $v_x = -u_y$ *and* $v_y = u_x$.

Verify that the following u *are harmonic and find the corresponding harmonic conjugates* v*:*

b) $u = 2xy$ c) $u = e^x \cos y$ d) $u = x^3 - 3xy^2$

Exercise* 1.9.15:

a) *Show that every separable equation* $y' = f(x)g(y)$ *can be written as an exact equation, and verify that it is indeed exact.*

b) *Using this rewrite* $y' = xy$ *as an exact equation, solve it and verify that the solution is the same as it was in Example 1.3.1.*

1.10 Modeling with First Order Equations

Learning Objectives

After this section, you will be able to:

- Write a first-order differential equation to model a physical situation and
- Interpret the solution to a differential equation in the context of a physical problem.

One of the main reasons to study and learn about differential equations, particularly for scientists and engineers, is their application and use in mathematical modeling. Since the derivative of a function represents the rate of change of that quantity, if we can use physical or scientific principles to develop an equation for the rate of change of some quantity in terms of the quantity and time, there's a chance that we can write a differential equation for this quantity and solve it to determine how the quantity will change.

1.10.1 Principles of Mathematical Modeling

The process of mathematical modeling involves three main steps. The first of these is to write the model. This part comes from basic science or engineering principles and involves writing a differential equation that fits the given situation. If we can determine the rate at which a quantity will change based on the surrounding factors, we have a good shot of getting to such an equation. One main principle that can be used to write these equations is the accumuilation equation, which will be discussed in the next subsection.

The second step of this process is to solve the differential equation. This can mean either an analytic solution or a numeric one, and this is where the work of this class comes into play. We are going through a bunch of different techniques for solving differential equations and analyzing the overall behavior of such equations so that we can use them in this way. The end goal is to get an equation or a graph for how the quantity that we made a model for is going to change in time.

The final step of the process is to validate the model by comparing with experimental data. Once we have written the model and solved the corresponding differential equation, we want to make sure that the model works. To do this, we can take a new version of the original scenario, run the model as well as the physical experiment and see how the results compare. If the results are "close" (in whatever sense makes logical sense for the problem), then we have a good model and can keep it. However, if our results differ significantly, then the model we used probably doesn't apply to this problem. We need to go back to step 1 to try to figure out a better model for the physical situation in order to get more accurate results.

Why do we care about mathematical modeling? The biggest thing that it does from an engineering point of view is reduce the need for repeated testing. If we have a mathematical model that works for a given physical system, we can see how the system will be have under slightly different conditions and with different initial conditions without needing to run the physical experiment over and over again. We can do all of this testing on the model, and

since we have validated the model, we can assume that the actual results will be similar. This also allows us to change some aspects of the physical situation to try to optimize it, but do so just by modifying the mathematical model, not the physical setup. This can significantly cut down on costs and allow for more optimal system design at the same time.

1.10.2 The Accumulation Equation

The accumulation equation is one of the simplest general mathematical formulations that can be used to develop mathematical models. This equation comes down to the fact that the rate of change of some quantity should be equal to the rate at which it is being added minus the rate at which it is being removed. If we let x be the quantity in question, this can be written as

$$\frac{dx}{dt} = \text{rate in} - \text{rate out}. \tag{1.9}$$

This may seem fairly simple. However, it shows up in many places in science and engineering. Any mass or energy balance equations are examples of accumulation equations. These types of equations can also be written for the accumulation of momentum, and doing so for fluids gives rise to the Navier-Stokes equations, providing the basis for several fields of engineering. The examples that we see here will be simpler than that, but the idea is still the same.

Example 1.10.1: A tank initially contains 70 gallons of water and 5 lbs of salt. A solution with salt concentration 0.2 lbs per gallon flows into the tank at a rate of 3 gal/min. The tank is well stirred, and water is removed from the tank at a rate of 3 gal/min. Find the amount of salt in the tank at any time t? What happens as $t \to \infty$? Does this make sense?

Solution: To solve this problem, we use the accumulation equation (1.9) on the amount of salt in the tank. In order to compute with this, we recognize that in terms of mass of salt moving into the tank

$$\text{rate in} = \text{flow in} \times \text{concentration in}$$

and similarly for the mass of salt leaving the tank.

If we let x represent the amount of salt in the tank at any time t (which is the goal of the problem), we can write a differential equation for this using the accumulation equation (1.9). This gives us that

$$\frac{dx}{dt} = \text{rate in} - \text{rate out} = \text{flow in} \times \text{concentration in} - \text{flow out} \times \text{concentration out}$$

For this problem, we have that

$$\begin{aligned} \text{flow in} &= 3, \\ \text{concentration in} &= 0.2, \\ \text{flow out} &= 3, \\ \text{concentration out} &= \frac{x}{\text{volume}} = \frac{x}{70}. \end{aligned}$$

The last of these lines comes from the fact that the tank is "well stirred" or "well-mixed." This implies that the concentration of salt in the water leaving the tank is the same as the

concentration in the tank, which we can compute as $\frac{x}{\text{volume}}$. In this case, since the flow rate in and out are both 3 gal/min, the volume of water in the tank is fixed at 70 gallons, so we can put this in the equation.

Therefore, our equation becomes

$$\frac{dx}{dt} = (3 \times 0.2) - \left(3 \times \frac{x}{70}\right).$$

We can rewrite this equation as

$$\frac{dx}{dt} + \frac{3}{70}x = 0.6$$

which we recognize as a first order linear equation. We can then solve this using the method of integrating factors. Our factor $r(t)$ is

$$r(t) = e^{\int p(t)\,dt} = e^{\int \frac{3}{70}\,dt} = e^{\frac{3}{70}t},$$

which we can multiply on both sides of the equation to obtain

$$e^{\frac{3}{70}t}\frac{dx}{dt} + e^{\frac{3}{70}t}\frac{3}{70}x = 0.6e^{\frac{3}{70}t}.$$

The left side of this is a product rule derivative, so we can integrate both sides to obtain

$$e^{\frac{3}{70}t}x = 0.6\frac{70}{3}e^{\frac{3}{70}t} + C.$$

We can then isolate x to get our general solution as

$$x = 14 + Ce^{-\frac{3}{70}t}.$$

Our initial condition tells us that $x(0) = 5$. Plugging this in gives that

$$5 = x(0) = 14 + C \quad \Rightarrow \quad C = -9,$$

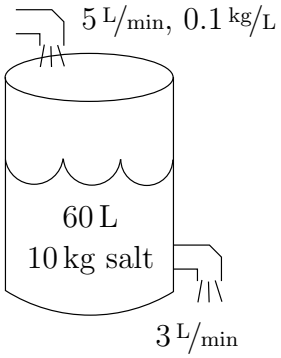
so the solution to the initial value problem, and thus our calculation for the amount of salt in the tank at any time t, is

$$x(t) = 14 - 9e^{-\frac{3}{70}t}.$$

As $t \to \infty$, we see that the exponential term goes to zero. This leaves us with 14 lbs of salt in the tank after a long time. This makes some sense because this would give us a concentration of $\frac{14}{70} = 0.2$ lb/gal, and that was exactly the concentration of the in-flow stream. It makes sense that after a long time of mixing and removing water from the tank, the concentration of the tank would match that of the incoming stream. ⌟

The same principle works for other types of examples, including those where the volume of the tank is not constant in time.

Example 1.10.2: A 100 liter tank contains 10 kilograms of salt dissolved in 60 liters of water. Solution of water and salt (brine) with concentration of 0.1 kilograms per liter is flowing in at the rate of 5 liters a minute. The solution in the tank is well stirred and flows out at a rate of 3 liters a minute. How much salt is in the tank when the tank is full?



Solution: We can again use the accumulation equation to write

$$\frac{dx}{dt} = (\text{flow in} \times \text{concentration in}) - (\text{flow out} \times \text{concentration out}).$$

In this example, we have

$$\begin{aligned} \text{flow in} &= 5, \\ \text{concentration in} &= 0.1, \\ \text{flow out} &= 3, \\ \text{concentration out} &= \frac{x}{\text{volume}} = \frac{x}{60+(5-3)t}. \end{aligned}$$

Our equation is, therefore,

$$\frac{dx}{dt} = (5 \times 0.1) - \left(3\frac{x}{60+2t}\right).$$

Or in the form (1.3)

$$\frac{dx}{dt} + \frac{3}{60+2t}x = 0.5.$$

Let us solve. The integrating factor is

$$r(t) = \exp\left(\int \frac{3}{60+2t}dt\right) = \exp\left(\frac{3}{2}\ln(60+2t)\right) = (60+2t)^{3/2}.$$

We multiply both sides of the equation to get

$$\begin{aligned} (60+2t)^{3/2}\frac{dx}{dt} + (60+2t)^{3/2}\frac{3}{60+2t}x &= 0.5(60+2t)^{3/2}, \\ \frac{d}{dt}\left[(60+2t)^{3/2}x\right] &= 0.5(60+2t)^{3/2}, \\ (60+2t)^{3/2}x &= \int 0.5(60+2t)^{3/2}dt + C, \\ x &= (60+2t)^{-3/2}\int \frac{(60+2t)^{3/2}}{2}dt + C(60+2t)^{-3/2}, \\ x &= (60+2t)^{-3/2}\frac{1}{10}(60+2t)^{5/2} + C(60+2t)^{-3/2}, \\ x &= \frac{60+2t}{10} + C(60+2t)^{-3/2}. \end{aligned}$$

We need to find C. We know that at $t = 0$, $x = 10$. So

$$10 = x(0) = \frac{60}{10} + C(60)^{-3/2} = 6 + C(60)^{-3/2},$$

or

$$C = 4(60^{3/2}) \approx 1859.03.$$

We are interested in x when the tank is full. The tank is full when $60 + 2t = 100$, or when $t = 20$. So

$$\begin{aligned} x(20) &= \frac{60 + 40}{10} + C(60 + 40)^{-3/2} \\ &\approx 10 + 1859.03(100)^{-3/2} \approx 11.86. \end{aligned}$$

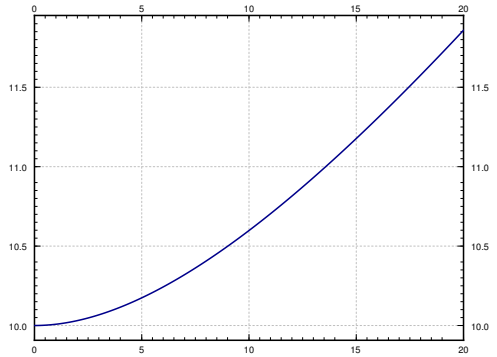


Figure 1.31: Graph of the solution x kilograms of salt in the tank at time t.

See Figure 1.31 for the graph of x over t.

The concentration when the tank is full is approximately $0.1186\,{}^{\text{kg}}/_{\text{liter}}$, and we started with $1/6$ or $0.167\,{}^{\text{kg}}/_{\text{liter}}$.

For the previous example, we obtained the solution

$$x(t) = \frac{60 + 2t}{10} + 1859.03(60 + 2t)^{-3/2},$$

which is valid and well defined for all positive values of t (it has an issue at $t = -30$, but we aren't concerned about that here). However, as a differential equation that represents a physical situation, it is not valid for all positive values of t. The issue here is that the tank is full at $t = 20$. Therefore, beyond this point, while the function still exists, it is not a valid model for this physical system. Once the tank fills, you can't keep adding and removing water at the same rates that you have been up until this point, because something is going to break with the system. The same goes for if you are removing water from the tank at a faster rate than you are adding it, because then the tank will empty at some point in time and beyond that, the model equation no longer represents the system.

The same ideas apply to problems involving interest compounded continuously. For an interest rate of r, the "rate in," or the rate at which the money in the account is increasing, is rP where P is the amount of money in the account. Taking this along with other factors that may affect the balance of the account allows us to write a differential equation, which we can solve to determine what the balance will be over time.

Example 1.10.3: A bank account with an interest rate of 6% per year, compounded continuously, starts with a balance of \$30000. The owner of the account withdraws \$50 from the account each month. Find and solve a differential equation for the account balance over time. What is the largest amount that the owner could withdraw each month without the account eventually reaching \$0?

Solution: We will use the function $P(t)$ to model the balance of the account over time, where t is in *years*. Since the owner withdraws \$50 per month, this means that they withdraw

$600 over the course of the year. This means that the differential equation we want is

$$\frac{dP}{dt} = 0.06P - 600 \qquad P(0) = 30000.$$

We can solve this equation by the integrating factor method.

$$\begin{aligned} P' - 0.06P &= -600 \\ (e^{-0.06t}P)' &= -600e^{-0.06t} \\ e^{-0.06t}P &= 10000e^{-0.06t} + C \\ P &= 10000 + Ce^{0.06t} \end{aligned}$$

For $P(0) = 30000$, we need to take $C = 20000$. Thus, the solution to the initial value problem is

$$P(t) = 10000 + 20000e^{0.06t}.$$

Since the coefficient in front of $e^{0.06t}$ is positive, this means that the account balance here will grow in time.

For the second part, we need to adjust the withdrawal amount to see how the solution changes. If we let K be the monthly withdrawal amount, then we have the differential equation

$$\frac{dP}{dt} = 0.06P - 12K \qquad P(0) = 30000.$$

The same solution method gives us

$$P(t) = \frac{12K}{0.06} + Ce^{0.06t}.$$

If $C < 0$, then the account balance will eventually go to zero. Therefore, we need $C \geq 0$, and since $P(0) = 30000$, we have that

$$30000 = \frac{12K}{0.06} + C \quad \text{or} \quad C = 30000 - \frac{12K}{0.06}.$$

For this to be equal to zero, we need

$$\frac{12K}{0.06} = 30000 \qquad K = 150.$$

Thus, the owner can withdraw $150 per month and keep the account balance positive. ⌟

To end this section, we will analyze the example that was presented at the very beginning of the book.

Example 1.10.4: An object falling through the air has its velocity affected by two factors: gravity and a drag force. The velocity downward is increased at a rate of 9.8 m/s^2 due to gravity, and it is decreased by a rate equation to 0.3 times the current velocity of the object. If the ball is initially thrown downwards at a speed of 2 m/s, what will the velocity be 10 seconds later?

Solution: As described in that first section, we know that the differential equation that we can write for this situation is

$$\frac{dv}{dt} = 9.8 - 0.3v$$

and that the initial condition for the velocity if $v(0) = 2$. Since we have gravity as a positive 9.8, this means that the downward direction is positive, so the object being thrown downward at 2 m/s means that it is positive. We then need to solve this initial value problem, which we can do using first order linear methods. The equation can be written as

$$v' + 0.3v = 9.8$$

which has integrating factor $e^{0.3t}$. After multiplying this to both sides and integrating, we get that

$$e^{0.3t}v = \frac{9.8}{0.3}e^{0.3t} + C$$

or that

$$v(t) = \frac{9.8}{0.3} + Ce^{-0.3t}.$$

Using the initial condition, we get that

$$v(0) = \frac{98}{3} + C = 2$$

so that $C = -\frac{92}{3}$ and the solution to the initial value problem is

$$v(t) = \frac{98}{3} - \frac{92}{3}e^{-0.3t}.$$

Then, to determine the velocity at $t = 10$, we can plug 10 into this formula to get that

$$v(10) = \frac{98}{3} - \frac{92}{3}e^{-3} \approx 31.14 \ m/s.$$

All of these examples are based around the same idea of the accumulation equation. We need to determine the quantity that is changing as well as all of the factors that cause it to increase and decrease. These get combined into a differential equation which we can solve in order to analyze the situation and answer whatever questions you want about that physical problem. Keeping these ideas in mind will help you approach a wide variety of problems both in this class as well as future applications in engineering classes and beyond.

1.10.3 Exercises

***Exercise* 1.10.1:** *Suppose there are two lakes located on a stream. Clean water flows into the first lake, then the water from the first lake flows into the second lake, and then water from the second lake flows further downstream. The in and out flow from each lake is 500 liters per hour. The first lake contains 100 thousand liters of water and the second lake contains 200 thousand liters of water. A truck with 500 kg of toxic substance crashes into the first lake. Assume that the water is being continually mixed perfectly by the stream.*

a) *Find the concentration of toxic substance as a function of time in both lakes.*

b) *When will the concentration in the first lake be below 0.001 kg per liter?*

c) *When will the concentration in the second lake be maximal?*

***Exercise* 1.10.2:** *Newton's law of cooling states that* $\frac{dx}{dt} = -k(x - A)$ *where* x *is the temperature,* t *is time,* A *is the ambient temperature, and* $k > 0$ *is a constant. Suppose that* $A = A_0 \cos(\omega t)$ *for some constants* A_0 *and* ω*. That is, the ambient temperature oscillates (for example night and day temperatures).*

a) *Find the general solution.*

b) *In the long term, will the initial conditions make much of a difference? Why or why not?*

***Exercise* 1.10.3:** *Initially 5 grams of salt are dissolved in 20 liters of water. Brine with concentration of salt 2 grams of salt per liter is added at a rate of 3 liters per minute. The tank is mixed well and is drained at 3 liters per minute. How long does the process have to continue until there are 20 grams of salt in the tank?*

***Exercise* 1.10.4:** *Initially a tank contains 10 liters of pure water. Brine of unknown (but constant) concentration of salt is flowing in at 1 liter per minute. The water is mixed well and drained at 1 liter per minute. In 20 minutes there are 15 grams of salt in the tank. What is the concentration of salt in the incoming brine?*

Exercise* 1.10.5: *Suppose a water tank is being pumped out at* $3\,{}^{L}\!/_{\text{min}}$*. The water tank starts at 10 L of clean water. Water with toxic substance is flowing into the tank at* $2\,{}^{L}\!/_{\text{min}}$*, with concentration* $20t\,{}^{g}\!/_{L}$ *at time* t*. When the tank is half empty, how many grams of toxic substance are in the tank (assuming perfect mixing)?*

***Exercise* 1.10.6:** *A 300 gallon well-mixed water tank initially starts with 200 gallons of water and 15 lbs of salt. One stream with salt concentration one pound per gallon flows into the tank at a rate of 3 gallons per minute and water is removed from the well-mixed tank at a rate of 2 gallons per minute.*

a) *Write and solve an initial value problem for the volume of water in the tank at any time* t*.*

b) *Set up an initial value problem for the amount of salt in the tank at any time* t*. You do not need to solve it (yet), but should make sure to state it fully.*

c) *Is the solution to this initial value problem a valid representation of the physical model for all times* $t > 0$*? If so, use the information in the equation to determine the long-time behavior of the solution. If not, explain why, determine the time when the representation breaks down, and what happens at that point in time.*

d) *Solve the initial value problem above and compare this to your answer to the previous part.*

***Exercise* 1.10.7:** *A 500 gallon well-mixed water tank initially starts with 300 gallons of water and 200 lbs of salt. One stream with salt concentration of 0.5 $^{lb}/_{gal}$ flows into the tank at a rate of 5 $^{gal}/_{min}$ and water is removed from the well-mixed tank at a rate of 7 $^{gal}/_{min}$.*

a) *Write and solve an initial value problem for the volume of water in the tank at any time t.*

b) *Set up an initial value problem for the amount of salt in the tank at any time t. You do not need to solve it (yet), but should make sure to state it fully.*

c) *Is the solution to this initial value problem a valid representation of the physical model for all times $t > 0$? If so, use the information in the equation to determine the long-time behavior of the solution. If not, explain why, determine the time when the representation breaks down, and what happens at that point in time.*

d) *Solve the initial value problem above and compare this to your answer to the previous part.*

***Exercise* 1.10.8:** *A 200 gallon well-mixed water tank initially starts with 150 gallons of water and 50 lbs of salt. One stream with salt concentration of 0.2 $^{lb}/_{gal}$ flows into the tank at a rate of 4 $^{gal}/_{min}$ and water is removed from the well-mixed tank at a rate of 4 $^{gal}/_{min}$.*

a) *Write and solve an initial value problem for the volume of water in the tank at any time t.*

b) *Set up an initial value problem for the amount of salt in the tank at any time t. You do not need to solve it (yet), but should make sure to state it fully.*

c) *Is the solution to this initial value problem a valid representation of the physical model for all times $t > 0$? If so, use the information in the equation to determine the long-time behavior of the solution. If not, explain why, determine the time when the representation breaks down, and what happens at that point in time.*

d) *Solve the initial value problem above and compare this to your answer to the previous part.*

Exercise* 1.10.9: *Suppose we have bacteria on a plate and suppose that we are slowly adding a toxic substance such that the rate of growth is slowing down. That is, suppose that $\frac{dP}{dt} = (2 - 0.1\,t)P$. If $P(0) = 1000$, find the population at $t = 5$.*

Exercise* 1.10.10: *A cylindrical water tank has water flowing in at I cubic meters per second. Let A be the area of the cross section of the tank in meters. Suppose water is flowing from the bottom of the tank at a rate proportional to the height of the water level. Set up the differential equation for h, the height of the water, introducing and naming constants that you need. You should also give the units for your constants.*

***Exercise* 1.10.11:** *An object in free fall has a velocity that increases at a rate of 32 ft/s^2. Due to drag, the velocity decreases at a rate of 0.1 times the velocity of the object squared, when written in feet per second.*

a) *Write a differential equation to model the velocity of this object over time.*

b) *This equation is autonomous, so draw a phase diagram for this equation and classify all critical points.*

c) *What will happen to the velocity if the object is dropped at $t = 0$? What about if the object is thrown downwards at a rate of $10ft/s$?*

***Exercise* 1.10.12:** *The number of people in a town that support a given measure decays at a constant rate of 10 people per day. However, the support for the measure can be increased by individuals discussing the issue. This results in an increase of the support at a rate of $ay(1000 - y)$ people per day, where y is the number of people who support the measure, and a is a constant depending on the way in which the issue is being discussed. Write a differential equation to model this situation, and determine the amount of people who will support the measure long-term if a is set to 2.*

***Exercise* 1.10.13:** *Newton's Law of Procrastination states that the rate at which one accomplishes a chore is proportional to the amount of the chore not yet done. Unbeknownst to Newton, this applies to robots too. A Roomba is attempting to vacuum a house measuring 1000 square feet. When none of the house is clean, the roomba can clean 200 square feet per hour. What makes this problem fun is that there is also a dog. It's whatever kind of dog you like, take your pick. The dog dirties the house at a constant rate of 50 square feet per hour.*

a) *Assume that none of the house is clean at $t = 0$. Write a DE for the number of square feet that are clean as a function of time, and solve for that quantity.*

b) *How long will it take before the house is half clean? Will it ever be entirely clean? (Explain briefly.)*

***Exercise* 1.10.14:** *A student has a loan for \$50000 with 5% interest. The student makes \$300 payments on the loan each month.*

a) *With this setup, how long does it take the student to pay off the loan? How much money does the student pay over this period of time?*

b) *What is the minimal amount the student should pay each month if they want to pay off the loan within 5 years? How much does the student pay over this period?*

***Exercise* 1.10.15:** *A factory pumps 6 tons of sludge per day into a nearby pond. The pond initially contains 100,000 gallons of water, and no sludge. Each day, 3,000 gallons of rain water falls into the pond, and 1,000 gallons per day leave the pond via a river. Assume, for no good reason, that the water leaving the pond has the same concentration of sludge as the pond as a whole. How much sludge will there be in the pond after 150 days?*

***Exercise* 1.10.16:** *In this exercise, we compare two different young people and their investment strategies. Both of these people are investing in an account with 7.5% annual rate of return. Person 1 invests \$50 a month starting at age 20, and Person 2 invests \$100 per month starting at age 30. Write differential equations to model each of these account balances over time, and compute the amount of money in each account at age 50. Who has more money in the account? Who has invested more money? What would person 2 have to invest each month in order for the two balances to be equal at age 50?*

***Exercise* 1.10.17:** *Radioactive decay follows similar rules to interest, where a certain portion of the material decays over time, resulting in an equation of the form*

$$\frac{dy}{dt} = -ky$$

for some constant k. The half-life *of a material is the amount of time that it takes for half of the material to have decayed away. Assume that the* half-life *of a given substance is T minutes. Find a formula for k, the coefficient in the decay equation, in terms of T.*

1.11 Modeling and Parameter Estimation

Learning Objectives

After this section, you will be able to:

- Use parameter estimation to approximate physical parameters from data.

One of the most common ways that the mathematical modeling structure can be used to analyze physical problems is the idea of parameter estimation. The situation is that we have physical principles that give rise to a differential equation that defines how a physical system should behave, but there are one or more constants in the problem that we do not know. Two simpler examples of this are Newton's Law of Cooling

$$\frac{dT}{dt} = -k(T - T_s)$$

which models the temperature of an object in an environment of temperature T_s over time, and velocity affected by drag

$$\frac{dv}{dt} = 9.8 - \alpha v^2$$

modeling the velocity of a falling object where the drag force is proportional to the square of the velocity. In both of these cases, the models are well established, but for a given object, we likely do not know the k or α values in the problem. These are these *parameters* of the problem, and would be determined by the shape and structure of the objects, the material that it is made of, and many other factors, so it could be hard to figure out what they are in advance. How can we find these values? We can use data from the actual physical problem to try to estimate these parameters.

The easier version of this is to use a single value at a later time to calculate the constant.

Example 1.11.1: An object that obeys Newton's Law of Cooling is placed in an environment at a constant temperature of 20^o C. The object starts at 50^o C, and after 10 minutes, it has reached a temperature of 40^o C. Find a function for the temperature as a function of time.

Solution: Based on Newton's Law of Cooling, we know that the temperature satisfies the differential equation

$$\frac{dT}{dt} = -k(T - T_s) = -k(T - 20)$$

with initial condition $T(0) = 50$, but we do not know the value of k. In order to work this out, we should solve the differential equation with unknown constant k, then figure out which value of k gives us the appropriate temperature after 10 minutes. This is a first order linear equation, which can be rewritten as

$$T' + kT = 20k.$$

The integrating factor we need is e^{kt}, which turns the equation into

$$(e^{kt}T)' = 20ke^{kt}.$$

Integrating both sides and solving for T gives

$$T(t) = 20 + Ce^{-kt}.$$

To satisfy the initial condition, we need that $T(0) = 50$, or $C = 30$. Thus, our solution, still with an unknown constant k, is

$$T(t) = 20 + 30e^{-kt}.$$

To determine the value of k, we need to utilize the other given piece of information: that $T(10) = 40$. Plugging this in gives that

$$40 = 20 + 30e^{-10k}$$

which we can solve for k using logarithms. This will give that

$$\frac{2}{3} = e^{-10k} \qquad \Rightarrow \qquad k = -\frac{1}{10}\ln\frac{2}{3}.$$

Finally, we can plug that constant into our equation to get the solution for the temperature at any time value,

$$T(t) = 20 + 30e^{-\frac{t}{10}\ln\frac{2}{3}}.$$

This method works great if we have the exact measurement from the object at one point in time. However, if the measurements at multiple points in time are known, and if the data is not likely to be exact, then a different method is more applicable. The idea is that we want to minimize the "error" between our predicted result and the physical data that we gather. The method used to minimize the error is the "Least Squared Error" method.

Assume that we want to do this for the drag coefficient problem,

$$\frac{dv}{dt} = 9.8 - \alpha v^2$$

where we do not know, and want to estimate, the value of α. For this method, the data that we gather is a set of velocity values $v_1, v_2, ..., v_n$ that are obtained at times $t_1, t_2, ..., t_n$. For any given value of α, we can solve, either numerically or analytically, the solution v_α to the given differential equation with that value of α. From this solution, we can compute $v_\alpha(t_1), v_\alpha(t_2), ..., v_\alpha(t_n)$, the value of this solution at each of the times that we gathered data originally. Now, we want to compute the error that we made in choosing this parameter α. This is computed by

$$E(\alpha) = (v_1 - v_\alpha(t_1))^2 + (v_2 - v_\alpha(t_2))^2 + \cdots + (v_n - v_\alpha(t_n))^2$$

which is the sum of the squares of the differences between the gathered data and the predicted solution. In order to find the best possible value of α, we want to minimze this error by choosing different values of α

$$E_{min} = \min_\alpha E(\alpha) = \min_\alpha \sum_{i=1}^{n} (v_i - v_\alpha(t_i))^2$$

and whatever value of α gives us this minimum is the optimal choice for that parameter.

The function that we want to minimize here is usually a very complicated function, and we may not even be able to solve the differential equation analytically for any α. Thus, computers are used most often here to solve these types of problems.

Example 1.11.2: An object is falling under the force of gravity, and has a drag component that is proportional to the square of the velocity. Data is gathered on the falling object, and the velocity at a variety of times are given in Table 1.3.

t (s)	**v** (m/s)
0	0
0.1	0.9797
0.3	2.8625
0.5	4.4750
0.8	6.3828
0.9	6.8360
1.0	7.0334
1.5	8.1612

Table 1.3: Data for estimating drag coefficient using least squared errors.

Use this data to estimate the coefficient of proportionality on the drag term in the equation

$$\frac{dv}{dt} = 9.8 - \alpha v^2.$$

Solution: To solve this problem, we will use the least squared error method implemented in MATLAB. The code we need for this is the following, which makes use of the Optimization Toolbox.

```
global tVals
global vVals

tVals = [0, 0.1, 0.3, 0.5, 0.8, 0.9, 1.0, 1.5];
vVals = [0,0.9797,2.8625,4.4750,6.3828,6.8360,7.0334,8.1612];

[aVal, errVal] = fminbnd(@(a) EstSqError(a), 0, 4)
```

This bit of code inputs the necessary values and uses the `fminbnd` function to find the minimum of the error function on a defined interval. These problems need to be done on a bounded interval, but in most physical situations there is some reasonable window for where the parameter could be. The rest of the code is the definition of the `EstSqError` function.

```
function err = EstSqError(al)

global tVals
global vVals

fun = @(t,v) 9.8 - al.*v.^2;
sol = ode45(fun, [0,3], 0);
vTest = deval(sol, tVals);

err = sum((vVals - vTest).^2)
end
```

The main point of this code is that it takes in a value of α, over which we are trying to minimize, it numerically solves the differential equation with that value of α over a desired range of values, and then compares the inputted `vVals` with the generated `vTest` array, computing the sum of squared errors, and returning the error value.

Running this code results in an α value of 0.1256, with an error of 0.0345. That means that, based on this data, the best approximation to α is 0.1256. ⌟

Note that in the above example, the total error was not zero, and doesn't actually match the coefficient used to generate the data, which was 0.123. This is because noise was added to the data before trying to compute the drag coefficient. In a real world problem, noise would not be added, but a similar effect would arise from slightly inaccurate measurements or round-off errors in the data. While we may not have found the constant exactly, we got really close to it, and could use this as a starting point for further experiments and data validation.

1.11.1 Exercises

***Exercise* 1.11.1:** *Bob is getting coffee from a restaurant and knows that the temperature of the coffee will follow Newton's Law of Cooling, which says that*

$$\frac{dT}{dt} = k(T_0 - T)$$

where T_0 is the ambient temperature and k is a constant depending on the object and geometry. His car is held at a constant 20° C, and when he receives the coffee, he measures the temperature to be 90° C. Two minutes later, the temperature is 81°C.

a) Use these two points of data to determine the value of k for this coffee.

b) Bob only wants to drink his coffee once it reaches 65° C. How long does he have to wait for this to happen?

c) If the coffee is too cold for Bob's taste once it reaches 35° C, how long is the acceptable window for Bob to drink his coffee?

***Exercise* 1.11.2:** *Assume that a falling object has a velocity (in meters per second) that obeys the differential equation*

$$\frac{dv}{dt} = 9.8 - \alpha v$$

where α represents the drag coefficient of the object.

a) *Solve this differential equation with initial condition $v(0) = 0$ to get a solution that depends on α.*

b) *Assume that you drop an object from a height of 10 meters and it hits the ground after 3 seconds. What is the value of α here? (Note: You solved for $v(t)$ in the previous part, and now you need to get to position. What does that require?)*

c) *Assume that a second object hits the ground in 6 seconds. How does this change the value of α?*

***Exercise* 1.11.3:** *A restaurant is trying to analyze the to-go coffee cups that it uses in order to best serve their customers. They assume that the coffee follows Newton's Law of Cooling and place it in a room with ambient temperature 22° C. They record the following data for the temperature of the coffee as a function of time.*

t (min)	**T** (° C)
0	80
0.5	77.1624
1.1	73.8082
1.7	70.6800
2.3	67.7996

a) *Use code to determine the best-fit value of k for this data.*

b) *The restaurant determines that to avoid any potential legal issues, the coffee can be no warmer than 60 °C when it is served. If the coffee comes out of the machine at 90 °C, how long do they have to wait before they can serve the coffee?*

Exercise* 1.11.4: *Assume that an object falling has a velocity that follows the differential equation*

$$\frac{dv}{dt} = 9.8 - \alpha v^2$$

where the velocity is in $^m/_s$ and α represents the drag coefficient. During a physics experiment, a student measures data for the velocity of a falling object over time given in the table below.

Use this data (and code) to estimate the value of α for this object.

t (s)	**v** (m/s)
0	0
0.1	0.9762
0.2	1.9341
0.4	3.6597
0.6	5.1613
0.9	6.7847
1.1	7.4103
1.3	7.9471
1.5	8.2975
1.8	8.5739
2.1	8.7769

Table 1.4: Data for Exercise 1.11.4.

t (d)	**P** (thousands)
0	50
7	58.6556
14	68.4521
28	91.4883
37	108.5750
50	135.7148
78	197.3520
100	239.9479

Table 1.5: Data for Exercise 1.11.5.

***Exercise* 1.11.5:** *Assume that a species of fish in a lake has a population that is modeled by the differential equation*

$$\frac{dP}{dt} = \frac{1}{100} rP(K - P) - \alpha$$

where r, K, and α are parameters, r representing the growth rate, K the carrying capacity, and α the harvesting rate, and the population P is in thousands., with t given in years. From previous studies, you know that the best value of r is 3.12. After studying the population over a period of time, you get the data given above.

a) *Your friend tells you that in a previous study, he found that the value of K for this particular lake is 255.2. Use code to determine the best value of α for this situation. Note that the equation expects t in years, but the data is given in days.*

b) *That answer doesn't look great. Plot the solution with these parameters along with the data and compare them.*

c) *The fit does not look great, so maybe your friend's value was not quite right. Run code to find best values for K and α simultaneously.*

1.12 Substitution

Attribution: [JL], §1.5.

Learning Objectives

After this section, you will be able to:

- Use substitution to solve more complicated first order equations,
- Use a Bernoulli substitution to solve appropriate first order equations, and
- Use a homogeneity transformation to solve appropriate first order equations.

The equation

$$y' = (x - y + 1)^2$$

is neither separable nor linear. What can we do? One technique that worked for helping us in evaluating integrals was substitution, or change of variables. For example, in order to evaluate the integral

$$\int 2x(x^2 + 4)^5 \, dx$$

we can do so by defining $u = x^2 + 4$ so that $du = 2x \, dx$, and then evaluate the integral as

$$\int u^5 \, du = \frac{u^6}{6} + C = \frac{(x^2 + 4)^6}{6} + C$$

after we have plugged our original function back in.

We can try to do the same thing here, and be careful with how we set things up. Our general strategy will be to pick a new dependent variable, find a differential equation that this new variable solves (which will come from the old equation), solve that equation, then convert back to the original variable. We will take v as our new dependent variable here, which is as function $v(x)$. Let us try

$$v = x - y + 1,$$

which is the "inside" function (it's inside the square) of our example. In order to get to a differential equation that v satisfies, we need to figure out y' in terms of v', v and x. We differentiate (in x) to obtain $v' = 1 - y'$. So $y' = 1 - v'$. We plug this into the equation to get

$$1 - v' = v^2.$$

In other words, $v' = 1 - v^2$. Such an equation we know how to solve by separating variables:

$$\frac{1}{1 - v^2} \, dv = dx.$$

So

$$\frac{1}{2} \ln \left| \frac{v+1}{v-1} \right| = x + C, \qquad \text{or} \qquad \left| \frac{v+1}{v-1} \right| = e^{2x+2C}, \qquad \text{or} \qquad \frac{v+1}{v-1} = De^{2x},$$

for some constant D. Note that $v = 1$ and $v = -1$ are also solutions; they are the *singular solutions* in this problem. (This solution method requires partial fractions; for a review of that topic, see § B.3.)

Now we need to "unsubstitute" to obtain

$$\frac{x - y + 2}{x - y} = De^{2x},$$

and also the two solutions $x - y + 1 = 1$ or $y = x$, and $x - y + 1 = -1$ or $y = x + 2$. We solve the first equation for y.

$$\begin{aligned} x - y + 2 &= (x - y)De^{2x}, \\ x - y + 2 &= Dxe^{2x} - yDe^{2x}, \\ -y + yDe^{2x} &= Dxe^{2x} - x - 2, \\ y\left(-1 + De^{2x}\right) &= Dxe^{2x} - x - 2, \\ y &= \frac{Dxe^{2x} - x - 2}{De^{2x} - 1}. \end{aligned}$$

Note that $D = 0$ gives $y = x + 2$, but no value of D gives the solution $y = x$.

Substitution in differential equations is applied in much the same way that it is applied in calculus. You guess. Several different substitutions might work. There are some general patterns to look for. We summarize a few of these in a table.

When you see	Try substituting
yy'	$v = y^2$
y^2y'	$v = y^3$
$(\cos y)y'$	$v = \sin y$
$(\sin y)y'$	$v = \cos y$
$y'e^y$	$v = e^y$

Usually you try to substitute in the "most complicated" part of the equation with the hopes of simplifying it. The table above is just a rule of thumb. You might have to modify your guesses. If a substitution does not work (it does not make the equation any simpler), try a different one.

1.12.1 Bernoulli equations

There are some forms of equations where there is a general rule for substitution that always works. One such example is the so-called *Bernoulli equation*[*]:

$$y' + p(x)y = q(x)y^n.$$

[*]There are several things called Bernoulli equations, this is just one of them. The Bernoullis were a prominent Swiss family of mathematicians. These particular equations are named for Jacob Bernoulli (1654–1705).

This equation looks a lot like a linear equation except for the y^n. If $n = 0$ or $n = 1$, then the equation is linear and we can solve it. Otherwise, the substitution $v = y^{1-n}$ transforms the Bernoulli equation into a linear equation. Note that n need not be an integer.

Example 1.12.1: Find the general solution of

$$y' - \frac{4}{3x}y = -\frac{2}{3}y^4$$

Solution: This equation fits the Bernoulli equation structure with $p(x) = -\frac{4}{3x}$ and $q(x) = -\frac{2}{3}$. Since there is a y^4 on the right-hand side, we take $n = 4$ and make the substitution $v = y^{1-4} = y^{-3}$. With this, we see that

$$v' = -3y^{-4}y'$$

or $y' = -1/3y^4v'$. Plugging this into the equation gives

$$\begin{aligned} -\frac{1}{3}y^4v' - \frac{4}{3x}y &= -\frac{2}{3}y^4 \\ -\frac{1}{3}v' - \frac{4}{3x}y^{-3} &= -\frac{2}{3} \\ v' + \frac{4}{x}v &= 2 \end{aligned}$$

This last equation is now a first order linear equation, so we can solve it. The integrating factor we are looking for is

$$\mu(x) = e^{\int p(x)\ dx} = e^{\int \frac{4}{x}\ dx} = e^{4\ln x} = x^4,$$

which results in the euation

$$x^4v' + 4x^3v = 2x^4.$$

The left-hand side is $(x^4v)'$, so we can integrate both sides to get

$$x^4v = \frac{2}{5}x^5 + C,$$

or, solving for v,

$$v(x) = \frac{2}{5}x + \frac{C}{x^4}.$$

However, our original equation was for y, not v. Using the fact that $v = y^{-3}$, we can solve for y as $y = v^{-1/3}$, giving

$$y(x) = \left(\frac{2}{5}x + \frac{C}{x^4}\right)^{-1/3} = \frac{1}{\sqrt[3]{\frac{2}{5}x + \frac{C}{x^4}}}$$

as the general solution to this equation. ⌟

Even if we need to use integrals to write out the solution to these Bernoulli equations, we can still use the substitution method and solve back out for the desired solution at the end.

Example 1.12.2: Solve

$$xy' + y(x+1) + xy^5 = 0, \qquad y(1) = 1.$$

Solution: First, the equation is Bernoulli ($p(x) = (x+1)/x$ and $q(x) = -1$). We substitute

$$v = y^{1-5} = y^{-4}, \qquad v' = -4y^{-5}y'.$$

In other words, $(-1/4)\, y^5 v' = y'$. So

$$\begin{aligned} xy' + y(x+1) + xy^5 &= 0, \\ \frac{-xy^5}{4}v' + y(x+1) + xy^5 &= 0, \\ \frac{-x}{4}v' + y^{-4}(x+1) + x &= 0, \\ \frac{-x}{4}v' + v(x+1) + x &= 0, \end{aligned}$$

and finally

$$v' - \frac{4(x+1)}{x}v = 4.$$

The equation is now linear. We can use the integrating factor method. In particular, we use formula (1.4). Let us assume that $x > 0$ so $|x| = x$. This assumption is OK, as our initial condition is $x = 1$. Let us compute the integrating factor. Here $p(s)$ from formula (1.4) is $\frac{-4(s+1)}{s}$.

$$\begin{aligned} e^{\int_1^x p(s)\,ds} &= \exp\left(\int_1^x \frac{-4(s+1)}{s}\,ds\right) = e^{-4x-4\ln(x)+4} = e^{-4x+4}x^{-4} = \frac{e^{-4x+4}}{x^4}, \\ e^{-\int_1^x p(s)\,ds} &= e^{4x+4\ln(x)-4} = e^{4x-4}x^4. \end{aligned}$$

We now plug in to (1.4)

$$\begin{aligned} v(x) &= e^{-\int_1^x p(s)\,ds}\left(\int_1^x e^{\int_1^t p(s)\,ds} 4\,dt + 1\right) \\ &= e^{4x-4}x^4\left(\int_1^x 4\frac{e^{-4t+4}}{t^4}\,dt + 1\right). \end{aligned}$$

The integral in this expression is not possible to find in closed form. As we said before, it is perfectly fine to have a definite integral in our solution. Now "unsubstitute"

$$\begin{aligned} y^{-4} &= e^{4x-4}x^4\left(4\int_1^x \frac{e^{-4t+4}}{t^4}\,dt + 1\right), \\ y &= \frac{e^{-x+1}}{x\left(4\int_1^x \frac{e^{-4t+4}}{t^4}\,dt + 1\right)^{1/4}}. \end{aligned}$$

1.12.2 Homogeneous equations

Another type of equations we can solve by substitution are the so-called *homogeneous equations*. Note that this is *not* the same as a homogeneous linear equation. These equations do not have to be linear, and are solved in a very different way. Suppose that we can write the differential equation as

$$y' = F\left(\frac{y}{x}\right).$$

Here we try the substitutions

$$v = \frac{y}{x} \qquad \text{and therefore} \qquad y' = v + xv'.$$

We note that the equation is transformed into

$$v + xv' = F(v) \qquad \text{or} \qquad xv' = F(v) - v \qquad \text{or} \qquad \frac{v'}{F(v) - v} = \frac{1}{x}.$$

Hence an implicit solution is

$$\int \frac{1}{F(v) - v}\, dv = \ln|x| + C.$$

Example 1.12.3: Solve

$$x^2 y' = y^2 + xy, \qquad y(1) = 1.$$

Solution: We put the equation into the form $y' = (y/x)^2 + y/x$. We substitute $v = y/x$ to get the separable equation

$$xv' = v^2 + v - v = v^2,$$

which has a solution

$$\begin{aligned} \int \frac{1}{v^2}\, dv &= \ln|x| + C, \\ \frac{-1}{v} &= \ln|x| + C, \\ v &= \frac{-1}{\ln|x| + C}. \end{aligned}$$

We unsubstitute

$$\begin{aligned} \frac{y}{x} &= \frac{-1}{\ln|x| + C}, \\ y &= \frac{-x}{\ln|x| + C}. \end{aligned}$$

We want $y(1) = 1$, so

$$1 = y(1) = \frac{-1}{\ln|1| + C} = \frac{-1}{C}.$$

Thus $C = -1$ and the solution we are looking for is

$$y = \frac{-x}{\ln|x| - 1}.$$

1.12.3 Exercises

Hint: Answers need not always be in closed form.

***Exercise* 1.12.1:** *Solve* $y' + y(x^2 - 1) + xy^6 = 0$, *with* $y(1) = 1$.

Exercise* 1.12.2: *Solve* $xy' + y + y^2 = 0$, $y(1) = 2$.

***Exercise* 1.12.3:** *Solve* $2yy' + 1 = y^2 + x$, *with* $y(0) = 1$.

Exercise* 1.12.4: *Solve* $xy' + y + x = 0$, $y(1) = 1$.

***Exercise* 1.12.5:** *Solve* $y' + xy = y^4$, *with* $y(0) = 1$.

***Exercise* 1.12.6:** *Solve* $y' + 3y = 2xy^4$.

***Exercise* 1.12.7:** *Solve* $xy' - 2y = (3x^2 - x^{-3})y^5$ *with* $y(1) = 2$.

***Exercise* 1.12.8:** *Solve* $y' + 5y = \frac{e^{2x}}{y^2}$.

Exercise* 1.12.9: *Solve* $y^2y' = y^3 - 3x$, $y(0) = 2$.

***Exercise* 1.12.10:** *Solve* $yy' + x = \sqrt{x^2 + y^2}$.

***Exercise* 1.12.11:** *Solve* $y' = (x + y - 1)^2$.

***Exercise* 1.12.12:** *Solve* $y' = \frac{x^2 - y^2}{xy}$, *with* $y(1) = 2$.

Exercise* 1.12.13: *Solve* $2yy' = e^{y^2 - x^2} + 2x$.

***Exercise* 1.12.14:** *Consider the DE*

$$\frac{dy}{dt} = \left(y - \frac{1}{t}\right)^2 - \frac{1}{t^2}. \tag{1.10}$$

a) *Explain why* (1.10) *is not a linear equation.*

b) *Use a Bernoulli substitution to solve* (1.10).

Chapter 2

Higher order linear ODEs

As addressed in Chapter 1, we have a lot of different techniques for solving first order equations. However, not all differential equations are first order. A lot of physical systems in the world operate using higher order equations, particularly second order. Consider the system of a mass hanging from a spring. Newton's second law tells us that the net force on the object equals the mass of the object times its acceleration. However, Hooke's law for springs says that the force the spring exerts on the object is proportional to the distance this object is from the equilibrium position. Therefore, we get a relation between the acceleration of the object and the position. Since the acceleration is the second derivative (in time) of the position of the object, this naturally gives rise to a second order equation.

This means that we want to see what we can do with higher order equations as well. If we can manage to find solutions to these equations as well, then we can address more types of physical problems as well. However, increasing the order of the equation makes it significantly more difficult to find solutions. Even for linear equations, where in first order, we had an explicit method and formula for solutions, we need to put many more restrictions on higher order linear equations in order to have a direct method to generate solutions.

2.1 Second order linear ODEs

Attribution: [JL], §2.1.

Learning Objectives

After this section, you will be able to:

- Identify the general second order linear differential equation,
- Determine the characteristic equation for constant coefficient equations,
- Find the general solution for constant coefficient equations in the real and distinct roots case, and
- Determine if two functions are linearly independent.

The general second order ordinary differential equation is of the form

$$y'' = F(x, y, y')$$

for F an arbitrary function of three variables. As with first order equations, if the function F is not in a nice or simple form, there really isn't a hope to find a solution for this. For second order equations, we need to be even more specific about the structure of these equations in order to find solutions than we did for first order.

Definition 2.1.1

The general *second order linear differential equation* is of the form

$$A(x)y'' + B(x)y' + C(x)y = F(x).$$

This equation can be written in *standard form* by dividing throuhg by $A(x)$ to get

$$y'' + p(x)y' + q(x)y = f(x), \tag{2.1}$$

where $p(x) = {}^{B(x)}/_{A(x)}$, $q(x) = {}^{C(x)}/_{A(x)}$, and $f(x) = {}^{F(x)}/_{A(x)}$.

The word *linear* means that the equation contains no powers nor functions of y, y', and y''. In the special case when $f(x) = 0$, we have a so-called *homogeneous* equation

$$y'' + p(x)y' + q(x)y = 0. \tag{2.2}$$

We have already seen some second order linear homogeneous equations.

$$\begin{aligned} &y'' + k^2 y = 0 && \text{Two solutions are:} \quad y_1 = \cos(kx), \quad y_2 = \sin(kx). \\ &y'' - k^2 y = 0 && \text{Two solutions are:} \quad y_1 = e^{kx}, \quad y_2 = e^{-kx}. \end{aligned}$$

With the examples above, we were able to find solutions. However, notice that these equations don't have functions of x as coefficients of the y term. This means they are constant coefficient equations. It turns out that one of the few ways we can have a guaranteed method for finding solutions to these equation is if they have constant coefficients. For first order, we had a method for every linear equation, but for second order, we only have a formulaic method for constant coefficient and homogeneous linear equations.

If we know two solutions of a linear homogeneous equation, we know many more of them.

Theorem 2.1.1 (Superposition)

Suppose y_1 and y_2 are two solutions of the homogeneous equation (2.2). Then

$$y(x) = C_1 y_1(x) + C_2 y_2(x),$$

also solves (2.2) for arbitrary constants C_1 and C_2.

That is, we can add solutions together and multiply them by constants to obtain new and different solutions. We call the expression $C_1 y_1 + C_2 y_2$ a *linear combination* of y_1 and y_2.

Let us prove this theorem; the proof is very enlightening and illustrates how linear equations work.

Proof: Let $y = C_1 y_1 + C_2 y_2$. Then

$$\begin{aligned} y'' + py' + qy &= (C_1 y_1 + C_2 y_2)'' + p(C_1 y_1 + C_2 y_2)' + q(C_1 y_1 + C_2 y_2) \\ &= C_1 y_1'' + C_2 y_2'' + C_1 p y_1' + C_2 p y_2' + C_1 q y_1 + C_2 q y_2 \\ &= C_1(y_1'' + p y_1' + q y_1) + C_2(y_2'' + p y_2' + q y_2) \\ &= C_1 \cdot 0 + C_2 \cdot 0 = 0. \quad \square \end{aligned}$$

The proof becomes even simpler to state if we use the operator notation. An *operator* is an object that eats functions and spits out functions (kind of like what a function is, but a function eats numbers and spits out numbers). Define the operator L by

$$L[y] = y'' + py' + qy.$$

The differential equation now becomes $L[y] = 0$. The operator (and the equation) L being *linear* means that $L[C_1 y_1 + C_2 y_2] = C_1 L[y_1] + C_2 L[y_2]$. The proof above becomes

$$L[y] = L[C_1 y_1 + C_2 y_2] = C_1 L[y_1] + C_2 L[y_2] = C_1 \cdot 0 + C_2 \cdot 0 = 0.$$

***Exercise* 2.1.1:** *This fact does not hold if the equation is non-linear. Show that* $y_1(t) = e^t$ *and* $y_2(t) = 1$ *solve*

$$y'' = \sqrt{y \cdot y'}$$

but $y(t) = e^t + 1$ *does not.*

Two different solutions to the second equation $y'' - k^2 y = 0$ are $y_1 = \cosh(kx)$ and $y_2 = \sinh(kx)$. Let us remind ourselves of the definition, $\cosh x = \frac{e^x + e^{-x}}{2}$ and $\sinh x = \frac{e^x - e^{-x}}{2}$. Therefore, these are solutions by superposition as they are linear combinations of the two exponential solutions.

The functions sinh and cosh are sometimes more convenient to use than the exponential. Let us review some of their properties:

$$\begin{aligned} &\cosh 0 = 1, && \sinh 0 = 0, \\ &\frac{d}{dx}\Big[\cosh x\Big] = \sinh x, && \frac{d}{dx}\Big[\sinh x\Big] = \cosh x, \\ &\cosh^2 x - \sinh^2 x = 1. \end{aligned}$$

***Exercise* 2.1.2:** *Derive these properties using the definitions of* sinh *and* cosh *in terms of exponentials.*

2.1.1 Intial Value Problems

For first order equations, a lot of problems were stated as Initial Value Problems, containing both a differential equation and an initial condition of the value of y at some point x_0. What do these initial condition(s) look like for second order equations?

Example 2.1.1: Solve the second-order differential equation

$$y'' = x.$$

Solution: We can attempt to find a solution to this problem by integrating both sides twice. A first integration gives

$$y' = \frac{x^2}{2} + C$$

and a second integration leads to

$$y = \frac{x^3}{6} + Cx + D$$

for any two constants C and D. We can check that differentiating this y function twice gives us back the function x that we wanted. ⌟

In the previous example, we ended up with two unknown constants in our answer, whereas for first order equations, we only had one. In order to specify these two constants, we will need to give two additional facts about this function. This could be the value of the function at two points, but more traditionally, it is given as the value of the function y and its first derivative y' at a value x_0. Fairly often, this value x_0 is 0, but it could be any other number.

Example 2.1.2: Solve the initial value problem

$$y'' = x, \qquad y(1) = 2,\ y'(1) = 3$$

Solution: We previously found our solution with unknown constants as

$$y = \frac{x^3}{6} + Cx + D$$

and also found that

$$y' = \frac{x^2}{2} + C.$$

To find the values of C and D, we need to plug in the two initial conditions into their corresponding functions. The initial value of the derivative gives that

$$3 = y'(1) = \frac{1^2}{2} + C = C + \frac{1}{2}$$

so that we have $C = \frac{5}{2}$. We can then use the initial value of y, along with this C value, to conclude that

$$2 = y(1) = \frac{1^3}{6} + \frac{5}{2}(1) + D = \frac{1}{6} + \frac{5}{2} + D = \frac{16}{6} + D.$$

Solving this out gives that $D = -\frac{4}{6} = -\frac{2}{3}$. Putting these constants in gives that the solution to the initial value problem is

$$y = \frac{x^3}{6} + \frac{5}{2}x - \frac{2}{3}.$$

⌟

For first-order equations, we have theorems that told us that solutions existed and were unique, at least on small intervals. Linear first-order equations in particular had a very nice existence and uniqueness theorem (Theorem 1.5.1), guaranteeing existence on a full interval wherever the coefficient functions are continuous. Linear second-order equations have an existence and uniqueness theorem that gives the same type of result when the initial condition is stated properly.

Theorem 2.1.2 (Existence and uniqueness)

Suppose p, q, f are continuous functions on some interval I, a is a number in I, and a, b_0, b_1 are constants. The equation

$$y'' + p(x)y' + q(x)y = f(x),$$

has exactly one solution $y(x)$ defined on the same interval I satisfying the initial conditions

$$y(a) = b_0, \qquad y'(a) = b_1.$$

For example, the equation $y'' + k^2 y = 0$ with $y(0) = b_0$ and $y'(0) = b_1$ has the solution

$$y(x) = b_0 \cos(kx) + \frac{b_1}{k} \sin(kx).$$

The equation $y'' - k^2 y = 0$ with $y(0) = b_0$ and $y'(0) = b_1$ has the solution

$$y(x) = b_0 \cosh(kx) + \frac{b_1}{k} \sinh(kx).$$

Using cosh and sinh in this solution allows us to solve for the initial conditions in a cleaner way than if we have used the exponentials.

As it did for first order equations, this theorem tells us what the proper form is for initial value problems for second order equations. The take-away here is that in order to fully specify a solution to an initial value problem, a second order equation requires two initial conditions. They are usually given in the form $y(a)$ and $y'(a)$, but could be given as $y(a_1)$ and $y(a_2)$ in other applications. In any case, two pieces of information are needed to determine a problem of second order, where we only needed one for first order.

2.1.2 Constant Coefficient Equations - Real and Distinct Roots

Now we want to try to solve some of these equations. As discussed earlier in this section, there is no explicit solution method possible for second order equations. However, if we restrict to a very simple case (which is also one that shows up frequently in physical systems) we can start to develop a method for solving these equations. The type of equation we restrict to is linear and constant coefficient equations. *Constant coefficients* means that the functions in front of y'', y', and y are constants, they do not depend on x. The most general second order, linear, constant coefficient equation is

$$ay'' + by' + cy = g(x)$$

for real constants a, b, c and an arbitrary function $g(x)$. We will study the solution of nonhomogeneous equations (with $g(x) \neq 0$) in § 2.5. We will first focus on finding general solutions to homogeneous equations, which are of the form

$$ay'' + by' + cy = 0.$$

Consider the problem

$$y'' - 6y' + 8y = 0.$$

This is a second order linear homogeneous equation with constant coefficients, so it fits the type of equation where we want to hunt for solutions. To guess a solution, think of a function that stays essentially the same when we differentiate it, so that we can take the function and its derivatives, add some multiples of these together, and end up with zero. Yes, we are talking about the exponential.

Let us try* a solution of the form $y = e^{rx}$. Then $y' = re^{rx}$ and $y'' = r^2 e^{rx}$. Plug in to get

$$\begin{aligned} y'' - 6y' + 8y &= 0, \\ \underbrace{r^2 e^{rx}}_{y''} - 6\underbrace{re^{rx}}_{y'} + 8\underbrace{e^{rx}}_{y} &= 0, \\ r^2 - 6r + 8 &= 0 \qquad \text{(divide through by } e^{rx}\text{)}, \\ (r-2)(r-4) &= 0. \end{aligned}$$

Hence, if $r = 2$ or $r = 4$, then e^{rx} is a solution. So let $y_1 = e^{2x}$ and $y_2 = e^{4x}$.

***Exercise* 2.1.3:** *Check that y_1 and y_2 are solutions.*

So we have found two solutions to this differential equation! That's great, but there may be a few concerning ideas at this point:

(1) Did we just get lucky with this particular equation?

(2) How do we know that there aren't other solutions that aren't of the form e^{rx}? We made that assumption, so we could have missed something.

The second point comes back to the existence and uniqueness theorem. This differential equation satisfies the conditions of the existence and uniqueness theorem. That means that as long as we find *a* solution that can meet any initial condition, then we know that the solution we have found is the *only* solution. We have not yet verified the part about meeting initial conditions yet (that's coming later), but once we do, we'll know that making this assumption is completely fine, because it got us to a solution that works, and the uniqueness theorem tells us that this is the only solution.

For the first point, let's try to generalize the calculation we did above into a method that will work for more equations. Suppose that we have an equation

$$ay'' + by' + cy = 0, \tag{2.3}$$

*Making an educated guess with some parameters to solve for is such a central technique in differential equations, that people sometimes use a fancy name for such a guess: *ansatz*, German for "initial placement of a tool at a work piece." Yes, the Germans have a word for that.

where a, b, c are constants. We can take our same assumption that the solution is of the form $y = e^{rx}$ to obtain

$$ar^2e^{rx} + bre^{rx} + ce^{rx} = 0.$$

Divide by e^{rx} to obtain the so-called *characteristic equation* of the ODE:

$$ar^2 + br + c = 0.$$

Solve for the r by using the quadratic formula.

$$r_1, r_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

There are three cases that can arise based on this equation.

(1) If $b^2 - 4ac > 0$, then we have r_1 and r_2 as two real roots to the equation. This is the same as the example above, and we get e^{r_1x} and e^{r_2x} as two solutions. This is the larger class of problems to which this exact process applies.

(2) If $b^2 - 4ac < 0$, then r_1 and r_2 are complex numbers. We can still use e^{r_1x} and e^{r_2x} as solutions, but this runs into some issues, which will be addressed in Section 2.2.

(3) If $b^2 - 4ac = 0$, then we only get one root, since $r_1 = r_2$. We do get that e^{r_1x} as a solution, but that's all we get. This is another issue, which is addressed in Section 2.3.

So, as long as we have $b^2 - 4ac > 0$, this method will work to give us two solutions to this differential equation.

Example 2.1.3: Find two values of r so that e^{rx} is a solution to

$$y'' + 3y' - 10y = 0$$

Our first step is to find the characteristic equation by plugging e^{rx} into the equation. This gives that

$$r^2 + 3r - 10 = 0$$

This polynomial factors as $(r - 2)(r + 5)$, so we know that values of $r = 2$ and $r = -5$ will work. This means (check this!) that e^{2x} and e^{-5x} solve this differential equation.

2.1.3 Linear Independence

Since e^{2x} and e^{-5x} solve the linear differential equation in the previous example, we know that superposition applies, so that $C_1e^{2x} + C_2e^{-5x}$ solves the differential equation for any C_1 and C_2. The last thing to check is that we can pick C_1 and C_2 in order to meet any initial condition that we want. If this is possible, then we know that our method using the characteristic equation to find e^{2x} and e^{-5x} as solutions was enough to always solve this problem. The end of this argument is done using the existence and uniqueness theorem as described previously.

Let's work this out. Assume that we are given b_0 and b_1 and want to solve the initial value problem

$$y'' + 3y' - 10y = 0 \qquad y(0) = b_0,\ y'(0) = b_1.$$

We want to do this by picking C_1 and C_2 in the expression $y = C_1e^{2x} + C_2e^{-5x}$ Since

$$y' = 2C_1e^{2x} - 5C_2e^{-5x}$$

we can plug zero into this equation and the equation for y to get that we would need to have

$$\begin{aligned} b_0 = y(0) &= C_1 + C_2 \\ b_1 = y'(0) &= 2C_1 - 5C_2 \end{aligned}.$$

We can solve this system of equations by elimination. Multiplying the first equation by 5 adding them together gives

$$5b_0 + b_1 = 7C_1$$

so that

$$C_1 = \frac{5b_0 + b_1}{7}.$$

We can then compute the value of C_2 as

$$C_2 = b_0 - C_1 = b_0 - \frac{5b_0 + b_1}{7} = \frac{2b_0 - b_1}{7}.$$

Therefore, we can appropriate values of C_1 and C_2 that will meet the initial conditions for arbitrary values b_0 and b_1. This is great! This means that our method of finding solutions was sufficient for this problem.

Let's look at this situation in more generality. Assume that we have two solutions y_1 and y_2 that solve a second order linear, homogeneous differential equation, and we want to know if $C_1y_1 + C_2y_2$ can meet any initial condition for this problem. We have two unknowns and two equations ($y(x_0)$ and $y'(x_0)$ for some value x_0), so it should work out.

We can carry out the same steps as above. If we have initial conditions $y(x_0) = b_0$ and $y'(x_0) = b_1$, we want to satisfy

$$\begin{aligned} b_0 = y(x_0) &= C_1y_1(x_0) + C_2y_2(x_0) \\ b_1 = y'(x_0) &= C_1y_1'(x_0) + C_2y_2'(x_0) \end{aligned}'$$

which we get by taking the derivative of $y(x) = C_1y_1(x) + C_2y_2(x)$ and plugging in x_0. We will again use elimination to solve this. We can multiply the first equation by $y_1'(x_0)$, multiply the second by $y_1(x_0)$, and subtract them. This will cancel out the C_1 term, leaving us with

$$b_0y_1'(x_0) - b_1y_1(x_0) = C_2(y_1'(x_0)y_2(x_0) - y_1(x_0)y_2'(x_0)).$$

We want to solve for C_2 here, and once we do that, solving for C_1 happens by plugging back into one of the original equations. Most of the time, this will be completely fine, but there's one issue left. We can't divide by zero. So to be able to solve these equations for C_1 and C_2, we need to know that

$$y_1'(x_0)y_2(x_0) - y_1(x_0)y_2'(x_0) \neq 0. \tag{2.4}$$

The left side of this equation is often called the *Wronskian* of the functions y_1 and y_2 at the point x_0. In general, the Wronskian is the function $y_1'(x)y_2(x) - y_2'(x)y_1(x)$ for two solutions to a second order differential equation. This relation (the Wronskian being non-zero) tells us that the two solutions y_1 and y_2 are different enough to allow us to meet every initial condition for the differential equation. This condition is so important to the study of second order linear equations that we give it a name. We say that two solutions y_1 and y_2 are *linearly independent at* x_0 if (2.4) holds, that is, if the Wronskian of the solutions is non-zero at that point. For two solutions of a differential equation (which is more specific than just having two random functions), two solutions being linearly independent is equivalent to 2.4 holding for any* value x_0 where they are defined. Our work and calculations above leads to the following theorem:

Theorem 2.1.3

Let p, q be continuous functions. Let y_1 and y_2 be two linearly independent solutions to the homogeneous equation (2.2). Then every other solution is of the form

$$y = C_1y_1 + C_2y_2$$

for some constants C_1 and C_2. That is, $y = C_1y_1 + C_2y_2$ is the general solution.

Note that this theorem works for all linear homogeneous equations, not just constant coefficients ones. However, the methods that we have described here (and will in future sections) for *finding* these solutions will generally only work for constant coefficient equations.

This idea of linear independence can also be expressed in a different way: two solutions y_1 and y_2 are linearly independent if only way to make the expression

$$c_1y_1 + c_2y_2 = 0$$

is by setting both $c_1 = 0$ and $c_2 = 0$. This comes from the idea of linear independence from linear algebra (see Chapter 3) and uniqueness of solutions to differential equations. If there are such constants, we can also rearrange the equation to give

$$y_1 = -\frac{c_2}{c_1}y_2$$

which says that y_1 is a constant multiple of y_2, which holds for all values of x. Thus, if we have y_1 and y_2, and there is no constant A so that $y_1 = Ay_2$, then these functions are linearly independent.

Example 2.1.4: Find the general solution of the differential equation $y'' + y = 0$.

Solution: One of the four fundamental equations in § 0.1.4 showed that the two functions $y_1 = \sin x$ and $y_2 = \cos x$ are solutions to the equation $y'' + y = 0$. It is not hard to see that sine and cosine are not constant multiples of each other. If $\sin x = A\cos x$ for some **constant** A, we let $x = 0$ and this would imply $A = 0$. But then $\sin x = 0$ for all x, which

*Abel's Theorem, another theoretical result, says that the Wronskian $y_1'y_2 - y_1y_2'$ is either always zero or never zero. That means that any one value can be checked to determine if two solutions are linearly independent. Picking 0 is usually a convenient choice.

is preposterous. So y_1 and y_2 are linearly independent. We could also have checked this by taking derivatives and plugging in zero. Since

$$y_1(0) = 0 \quad y_1'(0) = 1 \quad y_2(0) = 1 \quad y_2'(0) = 0$$

we have that

$$y_1'(0)y_2(0) - y_1(0)y_2'(0) = (1)(1) - (0)(0) = 1 \neq 0$$

so these solutions are linearly independent. Hence,

$$y = C_1 \cos x + C_2 \sin x$$

is the general solution to $y'' + y = 0$. □

For two functions, checking linear independence is rather simple. Let us see another example using non-constant coefficient equations. Consider $y'' - 2x^{-2}y = 0$. Then $y_1 = x^2$ and $y_2 = 1/x$ are solutions. To see that they are linearly indepedent, suppose one is a multple of the other: $y_1 = Ay_2$, we just have to find out that A cannot be a constant. In this case we have $A = y_1/y_2 = x^3$, this most decidedly not a constant. So $y = C_1x^2 + C_2 1/x$ is the general solution.

Now, back to our discussion of constant coefficient equations. If $b^2 - 4ac > 0$, then we have two distinct real roots r_1 and r_2, giving rise to solutions of the form $y_1(x) = e^{r_1x}$ and $y_2(x) = e^{r_2x}$. Using condition 2.4 with $x_0 = 0$, we compute

$$y_1'(0)y_2(0) - y_1(0)y_2'(0) = (r_1)(1) - (1)(r_2) = r_1 - r_2.$$

Since $r_1 \neq r_2$, this expression is not zero, so the two solutions are linearly independent. Therefore, in this case, we know that the general solution will be

$$y = C_1e^{r_1x} + C_2e^{r_2x}.$$

Using the other formulation of linear independence of two functions, we would need to show that there is no constant A so that

$$e^{r_1x} = Ae^{r_2x}.$$

Since this can be rewritten as $A = e^{(r_2-r_1)x}$ and we know that $r_1 \neq r_2$, this is not a constant, so we again know that these functions are linearly independent and give rise to a general solution.

Example 2.1.5: Solve the initial value problem

$$y'' + 2y' - 3y = 0 \qquad y(0) = 2,\ y'(0) = 1.$$

Solution: To start, we find the characteristic equation of this differential equation and look for the roots. The characteristic equation here is

$$r^2 + 2r - 3 = 0$$

and this factors as $(r+3)(r-1)=0$. Thus, the two roots are $r=1$ and $r=-3$, so that the general solution (and we know it is the general solution because these are different exponents and so the solutions are linearly independent) is

$$y(x) = C_1 e^x + C_2 e^{-3x}.$$

In order to find the values of C_1 and C_2, we need to use the initial conditions. Plugging zero into $y(x)$ gives

$$y(0) = 2 = C_1 + C_2$$

and since the derivative $y'(x) = C_1 e^x - 3C_2 e^{-3x}$, the second condition gives that

$$y'(0) = 1 = C_1 - 3C_2.$$

Subtracting the second equation from the first gives that

$$1 = 4C_2$$

so that $C_2 = 1/4$ and $C_1 = 7/4$. Thus, the solution to the initial value problem is

$$y(x) = \frac{7}{4}e^x + \frac{1}{4}e^{-3x}.$$

In this second example, we solve a problem in the same way, but the roots of the characteristic equation do not work out as nicely. Even with that, the structure and process for the problem is identical to the previous example.

Example 2.1.6: Solve the initial value problem

$$y'' - 2y' - y = 0 \qquad y(0) = 2,\ y'(0) = 3.$$

Solution: We start by looking for the characteristic equation of this differential equation and finding its roots. The characteristic equation is

$$r^2 - 2r - 1 = 0$$

which has roots

$$r = \frac{2 \pm \sqrt{(-2)^2 - 4(1)(-1)}}{2} = \frac{2 \pm \sqrt{8}}{2} = 1 \pm \sqrt{2}.$$

There are two real and distinct roots, so we know that the two solutions $y_1(x) = e^{(1+\sqrt{2})x}$ and $y_2(x) = e^{(1-\sqrt{2})x}$ are linearly independent, so we have that the general solution to this problem is

$$y(x) = C_1 e^{(1+\sqrt{2})x} + C_2 e^{(1-\sqrt{2})x}.$$

Next, we need to find the constants C_1 and C_2 to meet the initial conditions. We can see that, by computing the first derivative,

$$\begin{aligned} y(x) &= C_1 e^{(1+\sqrt{2})x} + C_2 e^{(1-\sqrt{2})x}, \\ y'(x) &= (1+\sqrt{2})C_1 e^{(1+\sqrt{2})x} + (1-\sqrt{2})C_2 e^{(1-\sqrt{2})x}, \end{aligned}$$

and plugging in $x = 0$ gives that we want C_1 and C_2 to solve

$$\begin{aligned} 2 &= C_1 + C_2, \\ 3 &= (1+\sqrt{2})C_1 + (1-\sqrt{2})C_2. \end{aligned}$$

We can solve this by any method. One trick at the start is to subtract equation 1 from equation 2, giving that

$$\begin{aligned} 2 &= C_1 + C_2, \\ 1 &= \sqrt{2}C_1 - \sqrt{2}C_2, \end{aligned}$$

which can be rewritten as

$$\begin{aligned} 2 &= C_1 + C_2, \\ \frac{1}{\sqrt{2}} &= C_1 - C_2. \end{aligned}$$

Adding these equations together and dividing by 2 gives that

$$2C_1 = 2 + \frac{1}{\sqrt{2}}$$

so that $C_1 = 1 + \frac{1}{2\sqrt{2}}$, and since $C_1 + C_2 = 2$, we have that $C_2 = 1 - \frac{1}{2\sqrt{2}}$. Therefore, the solution to the desired initial value problem is

$$y(x) = \left(1 + \frac{1}{2\sqrt{2}}\right) e^{(1+\sqrt{2})x} + \left(1 - \frac{1}{2\sqrt{2}}\right) e^{(1-\sqrt{2})x}.$$

2.1.4 Exercises

Exercise 2.1.4: *Show that $y = e^x$ and $y = e^{2x}$ are linearly independent.*

Exercise 2.1.5:* *Are $\sin(x)$ and e^x linearly independent? Justify.*

Exercise 2.1.6:* *Are e^x and e^{x+2} linearly independent? Justify.*

Exercise 2.1.7:* *Guess a solution to $y'' + y' + y = 5$.*

Exercise 2.1.8: *Take $y'' + 5y = 10x + 5$. Find (guess!) a solution.*

Exercise 2.1.9: *Verify that $y_1(t) = e^t \cos(2t)$ and $y_2(t) = e^t \sin(2t)$ both solve $y'' - 2y' + 5y = 0$. Are these two solutions linearly independent? What does that mean about the general solution to $y'' - 2y' + 5y = 0$?*

Exercise 2.1.10: *Prove the superposition principle for nonhomogeneous equations. Suppose that y_1 is a solution to $Ly_1 = f(x)$ and y_2 is a solution to $Ly_2 = g(x)$ (same linear operator L). Show that $y = y_1 + y_2$ solves $Ly = f(x) + g(x)$.*

Exercise 2.1.11: *Determine the maximal interval of existence of the solution to the differential equation*

$$(t-5)y'' + \frac{1}{t+1}y' + e^t y = \frac{\cos(t)}{t^2+1}$$

with initial condition $y(3) = 8$. What about if the initial condition is $y(-3) = 4$?

Exercise 2.1.12: *For the equation $x^2y'' - xy' = 0$, find two solutions, show that they are linearly independent and find the general solution. Hint: Try $y = x^r$.*

Exercise 2.1.13:* *Find the general solution to $xy'' + y' = 0$. Hint: It is a first order ODE in y'.*

Exercise 2.1.14: *Find the general solution of $2y'' + 2y' - 4y = 0$.*

Exercise 2.1.15: *Solve $y'' + 9y' = 0$ with $y(0) = 1$, $y'(0) = 1$.*

Exercise 2.1.16: *Find the general solution of $y'' + 9y' - 10y = 0$.*

Exercise 2.1.17: *Find the general solution to $y'' - 3y' - 4y = 0$.*

Exercise 2.1.18: *Find the general solution to $y'' + 6y' + 8y = 0$.*

Exercise 2.1.19: *Find the solution to $y'' - 3y' + 2y = 0$ with $y(0) = 3$ and $y'(0) = -1$.*

Exercise 2.1.20: *Find the solution to $y'' + y' - 12y = 0$ with $y(0) = 1$ and $y'(0) = -2$.*

Exercise 2.1.21:* *Find the general solution to $y'' + 4y' + 2y = 0$.*

Exercise 2.1.22:* *Find the solution to $2y'' + y' - 3y = 0$, $y(0) = a$, $y'(0) = b$.*

Exercise 2.1.23:* *Find the solution to $y'' - (\alpha+\beta)y' + \alpha\beta y = 0$, $y(0) = a$, $y'(0) = b$, where α, β, a, and b are real numbers, and $\alpha \neq \beta$.*

Exercise 2.1.24:* *Write down an equation (guess) for which we have the solutions e^x and e^{2x}. Hint: Try an equation of the form $y'' + Ay' + By = 0$ for constants A and B, plug in both e^x and e^{2x} and solve for A and B.*

Exercise 2.1.25:* *Construct an equation such that $y = C_1e^{3x} + C_2e^{-2x}$ is the general solution.*

Exercise 2.1.26: *Give an example of a 2nd-order DE whose general solution is $y = c_1e^{-2t} + c_2e^{-4t}$.*

Equations of the form $ax^2y'' + bxy' + cy = 0$ are called *Euler's equations* or *Cauchy–Euler equations*. They are solved by trying $y = x^r$ and solving for r (assume that $x \geq 0$ for simplicity).

Exercise 2.1.27: *Suppose that $(b-a)^2 - 4ac > 0$.*

a) *Find a formula for the general solution of $ax^2y'' + bxy' + cy = 0$. Hint: Try $y = x^r$ and find a formula for r.*

b) *What happens when $(b-a)^2 - 4ac = 0$ or $(b-a)^2 - 4ac < 0$?*

We will revisit the case when $(b-a)^2 - 4ac < 0$ later.

Exercise 2.1.28: *Same equation as in Exercise 2.1.27. Suppose $(b-a)^2 - 4ac = 0$. Find a formula for the general solution of $ax^2y'' + bxy' + cy = 0$. Hint: Try $y = x^r \ln x$ for the second solution.*

2.2 Complex Roots and Euler's Formula

Attribution: [JL], §2.2.

Learning Objectives

After this section, you will be able to:

- Understand the basics of complex numbers,
- Use complex numbers to find complex solutions to second order constant coefficient equations, and
- Use Euler's formula to find real-valued general solutions to these second order equations.

The next case to consider for constant coefficient second order equations is the one where $b^2 - 4ac < 0$. This results in two roots r_1 and r_2, but they are complex roots. In order to solve differential equations with $b^2 - 4ac < 0$, we need to be able to manipulate and use some properties of complex numbers. Complex numbers may seem a strange concept, especially because of the terminology. There is nothing imaginary or really complicated about complex numbers. For more bachground information on complex numbers, see Appendix B.2.

To start with, we define $i = \sqrt{-1}$. Since this is the square root of a negative number, this i is not a real number. A complex number is written in the form $z = x + iy$ where x and y are real numbers. For a complex number $x + iy$ we call x the *real part* and y the *imaginary part* of the number. Often the following notation is used,

$$\mathrm{Re}(x + iy) = x \qquad \text{and} \qquad \mathrm{Im}(x + iy) = y.$$

The real numbers are contained in the complex numbers as those complex numbers with the imaginary part being zero.

When trying to do arithmetic with complex numbers, we treat i as though it is a variable, and do computations just as we would with polynomials. The important fact that we will use to simplify is the fact that since $i = \sqrt{-1}$, we have that $i^2 = -1$. So whenever we see i^2, we replace it by -1. For example,

$$(2 + 3i)(4i) - 5i = (2 \times 4)i + (3 \times 4)i^2 - 5i = 8i + 12(-1) - 5i = -12 + 3i.$$

The numbers i and $-i$ are the two roots of $r^2 + 1 = 0$. Engineers often use the letter j instead of i for the square root of -1. We use the mathematicians' convention and use i.

***Exercise* 2.2.1:** *Make sure you understand (that you can justify) the following identities:*

a) $i^2 = -1$, $i^3 = -i$, $i^4 = 1$,

b) $\dfrac{1}{i} = -i$,

c) $(3 - 7i)(-2 - 9i) = \cdots = -69 - 13i$,

d) $(3-2i)(3+2i) = 3^2 - (2i)^2 = 3^2 + 2^2 = 13$,

e) $\frac{1}{3-2i} = \frac{1}{3-2i}\frac{3+2i}{3+2i} = \frac{3+2i}{13} = \frac{3}{13} + \frac{2}{13}i$.

In order to solve differential equations where the characteristic equation has complex roots, we need to deal with the exponential e^{a+bi} of complex numbers. We do this by writing down the Taylor series and plugging in the complex number. Because most properties of the exponential can be proved by looking at the Taylor series, these properties still hold for the complex exponential. For example the very important property: $e^{x+y} = e^x e^y$. This means that $e^{a+ib} = e^a e^{ib}$. Hence if we can compute e^{ib}, we can compute e^{a+ib}. For e^{ib}, we use the so-called *Euler's formula*.

Theorem 2.2.1 (Euler's formula)

$$e^{i\theta} = \cos\theta + i\sin\theta \qquad \text{and} \qquad e^{-i\theta} = \cos\theta - i\sin\theta.$$

In other words, $e^{a+ib} = e^a\big(\cos(b) + i\sin(b)\big) = e^a\cos(b) + ie^a\sin(b)$.

***Exercise* 2.2.2:** *Using Euler's formula, check the identities:*

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} \qquad \textit{and} \qquad \sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}.$$

***Exercise* 2.2.3:** *Double angle identities: Start with $e^{i(2\theta)} = \big(e^{i\theta}\big)^2$. Use Euler on each side and deduce:*

$$\cos(2\theta) = \cos^2\theta - \sin^2\theta \qquad \textit{and} \qquad \sin(2\theta) = 2\sin\theta\cos\theta.$$

2.2.1 Complex roots

Suppose the equation $ay'' + by' + cy = 0$ has the characteristic equation $ar^2 + br + c = 0$ that has complex roots. By the quadratic formula, the roots are $\frac{-b\pm\sqrt{b^2-4ac}}{2a}$. These roots are complex if $b^2 - 4ac < 0$. In this case the roots are

$$r_1, r_2 = \frac{-b}{2a} \pm i\frac{\sqrt{4ac - b^2}}{2a}.$$

As you can see, we always get a pair of roots of the form $\alpha \pm i\beta$. In this case we can still write the solution as

$$y = C_1 e^{(\alpha+i\beta)x} + C_2 e^{(\alpha-i\beta)x}.$$

However, the exponential is now complex-valued, and so (real) linear combinations of these solutions will be complex valued. If we are using these equations to model physical problems, the answer should be real-valued, as the position of a mass-on-a-spring can not be a complex number. To do this, we need to determine two real-valued, linearly independent solutions to this differential equation.

To do this, we use the following result.

Theorem 2.2.2

Consider the differential equation

$$y'' + p(x)y' + q(x)y = 0$$

where $p(t)$ and $q(t)$ are *real-valued* continuous functions on some interval I. If y is a complex-valued solution to this differential equation and we can split $y(x) = u(x) + iv(x)$ into its real and imaginary parts u and v, then u and v are both solutions to $y'' + p(x)y' + q(x)y = 0$.

Proof. This is based on the fact that the differential equation is linear. We can compute derivatives of y

$$\begin{aligned} y(x) &= u(x) + iv(x) \\ y'(x) &= u'(x) + iv'(x) \\ y''(x) &= u''(x) + iv''(x) \end{aligned} .$$

Then, we can plug this into the differential equation

$$\begin{aligned} 0 &= y'' + p(x)y' + q(x)y \\ &= u''(x) + iv''(x) + p(x)(u'(x) + iv'(x)) + q(x)(u(x) + iv(x)) \\ 0 &= u''(x) + p(x)u'(x) + q(x)u(x) + i(v''(x) + p(x)v'(x) + q(x)v(x)) \end{aligned} .$$

Since the equation at the end of this chain is equal to zero, it must be zero as a complex number, which means that both the real and imaginary parts must be zero. This means that

$$\begin{aligned} u''(x) + p(x)u'(x) + q(x)u(x) &= 0 \\ v''(x) + p(x)v'(x) + q(x)v(x) &= 0 \end{aligned}'$$

so that both u and v solve the original differential equation. □

To use this to solve the problem at hand, we have our solution

$$y_1(x) = e^{\alpha + i\beta x}$$

and we need to split this into its real and imaginary parts. Since

$$y_1 = e^{\alpha x}\cos(\beta x) + ie^{\alpha x}\sin(\beta x),$$

the real and imaginary parts of this function are

$$\begin{aligned} u(x) &= e^{\alpha x}\cos(\beta x) \\ v(x) &= e^{\alpha x}\sin(\beta x) \end{aligned}$$

which, by the previous theorem, we know are also solutions. These are two solutions to our original differential equation that are also real-valued!

On the other hand, assume that we take the other complex solution, which will be

$$y_2(x) = e^{\alpha - i\beta x}.$$

If we split this into real and imaginary parts, we will get

$$y_2 = e^{\alpha x}\cos(\beta x) - ie^{\alpha x}\sin(\beta x),$$

so that the real and imaginary parts of this solution are

$$\begin{aligned} u_2(x) &= e^{\alpha x}\cos(\beta x) \\ v_2(x) &= -e^{\alpha x}\sin(\beta x). \end{aligned}$$

These are exactly the same as the previous real and imaginary parts, up to the minus sign on v_2. Since we are going to incorporate these with constants C_1 and C_2 eventually, they will give rise to the same general solution. So, we only need one of these two complex solutions to generate our two linearly independent real-valued solutions, and either of the two complex solutions give the same pair of real-valued solutions.

***Exercise* 2.2.4:** *For $\beta \neq 0$, check that $e^{\alpha x}\cos(\beta x)$ and $e^{\alpha x}\sin(\beta x)$ are linearly independent.*

With that fact, we have the following theorem.

Theorem 2.2.3

Take the equation

$$ay'' + by' + cy = 0.$$

If the characteristic equation has the roots $\alpha \pm i\beta$ (when $b^2 - 4ac < 0$), then the general solution is

$$y = C_1 e^{\alpha x}\cos(\beta x) + C_2 e^{\alpha x}\sin(\beta x).$$

Example 2.2.1: Find the general solution of $y'' + k^2 y = 0$, for a constant $k > 0$.

Solution: The characteristic equation is $r^2 + k^2 = 0$. Therefore, the roots are $r = \pm ik$, and by the theorem, we have the general solution

$$y = C_1\cos(kx) + C_2\sin(kx).$$

Example 2.2.2: Find the solution of $y'' - 6y' + 13y = 0$, $y(0) = 0$, $y'(0) = 10$.

Solution: The characteristic equation is $r^2 - 6r + 13 = 0$. By completing the square we get $(r-3)^2 + 2^2 = 0$ and hence the roots are $r = 3 \pm 2i$. By the theorem we have the general solution

$$y = C_1 e^{3x}\cos(2x) + C_2 e^{3x}\sin(2x).$$

To find the solution satisfying the initial conditions, we first plug in zero to get

$$0 = y(0) = C_1 e^0\cos 0 + C_2 e^0 \sin 0 = C_1.$$

Hence, $C_1 = 0$ and $y = C_2 e^{3x} \sin(2x)$. We differentiate,

$$y' = 3C_2 e^{3x} \sin(2x) + 2C_2 e^{3x} \cos(2x).$$

We again plug in the initial condition and obtain $10 = y'(0) = 2C_2$, or $C_2 = 5$. The solution we are seeking is

$$y = 5e^{3x} \sin(2x).$$

In this previous example, we can get a fairly good idea of how to sketch out the graph of this function. Since $\sin(2x)$ oscillates between -1 and 1, the graph of $y = 5e^{3x} \sin(2x)$ will oscillate between the graphs of $5e^{3x}$ and $-5e^{3x}$. These curves that surround the graph of the solution are called *envelope curves* for the solution. In Figure 2.1, this phenomenon is illustrated for the function $y = 2e^x \sin(5x)$.

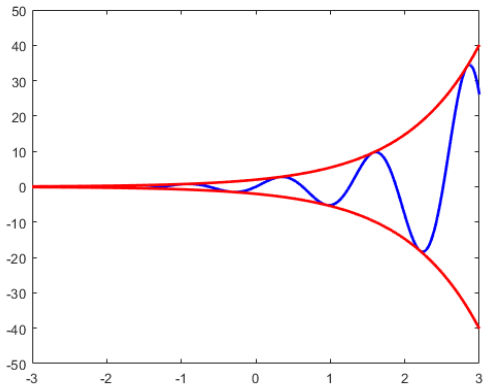


Figure 2.1: Plot of the function $y = 2e^x \sin(5x)$ with envelope curves.

This is simple when there is only one term in the function we want to draw. When both sine and cosine terms appear, this can get more tricky, but we can still work it out. In the more general case, the solution will look something like

$$y = Ae^{\alpha x} \cos(\beta x) + Be^{\alpha x} \sin(\beta x).$$

We can first factor out an $e^{\alpha x}$, and then we want to write $A\cos(\beta x) + B\sin(\beta x)$ as a single trigonometric function. The identity we want to use here is the trigonometric identity

$$\cos(\beta x - \delta) = \cos(\delta)\cos(\beta x) + \sin(\delta)\sin(\beta x).$$

If there is an angle δ so that $A = \cos(\delta)$ and $B = \sin(\delta)$, then we could write

$$A\cos(\beta x) + B\sin(\beta x) = \cos(\beta x - \delta)$$

and we would be done. However, this does not always happen; the main issue being that $\cos^2(\delta) + \sin^2(\delta) = 1$ for all δ, but it is not necessarily the case that $A^2 + B^2 = 1$. But we can

force this last condition. If we define $R = \sqrt{A^2 + B^2}$, then we can rewrite this expression as

$$\begin{aligned} A\cos(\beta x) + B\sin(\beta x) &= R\left(\frac{A}{\sqrt{A^2+B^2}}\cos(\beta x) + \frac{B}{\sqrt{A^2+B^2}}\sin(\beta x)\right) \\ &= R\left(\cos(\delta)\cos(\beta x) + \sin(\delta)\sin(\beta x)\right) \\ &= R\cos(\beta x - \delta) \end{aligned}$$

where δ is the angle so that

$$\cos(\delta) = \frac{A}{R} \qquad \sin(\delta) = \frac{B}{R}$$

and such an angle will always exist. Therefore, we can represent the original solution

$$y = Ae^{\alpha x}\cos(\beta x) + Be^{\alpha x}\sin(\beta x)$$

as

$$y = Re^{\alpha x}\cos(\beta x - \delta)$$

where

$$R = \sqrt{A^2 + B^2} \qquad \cos(\delta) = \frac{A}{R} \qquad \sin(\delta) = \frac{B}{R}.$$

Therefore, the envelope curves for this solution will be

$$y = \pm Re^{\alpha x}.$$

Note that in order to determine these envelope curves, you do not need to determine the δ value in the representation of the solution. All you need is the value of R, which can be computed as $\sqrt{A^2 + B^2}$ where A and B are the coefficients of the sine and cosine terms in the solution.

Example 2.2.3: Find the solution to the initial value problem

$$y'' + 2y' + 5y = 0 \qquad y(0) = 1,\ y'(0) = 5.$$

Determine a value T where the solution $y(x)$ satisfies $|y(x)| < 0.1$ for all $x > T$.

Solution: We solve the initial value problem by normal techniques from this section. The characteristic equation is $r^2 + 2r + 5 = 0$, which has roots $r = -1 \pm 2i$. Therefore, the general solution of the differential equation is

$$y = C_1 e^{-x}\cos(2x) + C_2 e^{-x}\sin(2x).$$

Plugging in 0 gives that $y(0) = 1 = C_1$, and the derivative of this general solution is

$$y' = -C_1 e^{-x}\cos(2x) - 2C_1 e^{-x}\sin(2x) - C_2 e^{-x}\sin(2x) + 2C_2 e^{-x}\cos(2x).$$

Plugging in 0 here gives

$$y'(0) = -C_1 + 2C_2.$$

Since $C_1 = 1$, this gives that $C_2 = 3$. So, our solution is

$$y(x) = e^{-x}\cos(2x) + 3e^{-x}\sin(2x).$$

Through the work above, we can find $R = \sqrt{1+9} = \sqrt{10}$. Therefore, the envelope curves for the solution are

$$\pm\sqrt{10}e^{-x}.$$

In order to find this threshold T where the solution will stay within 0.1 of zero, we need to figure out when this envelope curves get to the 0.1 threshold. Once the envelope curves get to that level, we know that the full solution must be trapped there as well. We can solve

$$0.1 = \sqrt{10}e^{-T} \qquad T = -\ln\left(\frac{0.1}{\sqrt{10}}\right) \approx 3.454.$$

So, for all values of x larger than 3.454, the solution will be within 0.1 of zero. This is illustrated in Figure 2.2. Note that we did not find the *best* value T here, as it probably could be made smaller using the actual solution. The issue here is that because the solution is oscillating, it may end up staying inside the 0.1 cutoff before that value of time, but this is the lowest value of T that we can prove and validate using envelope curves. ⌟

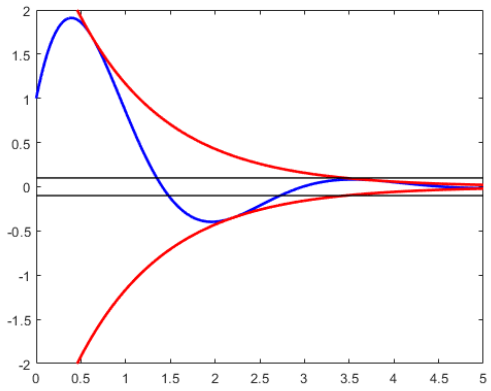


Figure 2.2: Plot of the function $e^{-x}\cos(2x) + 3e^{-x}\sin(2x)$ with envelope curves illustrating the bounds on the function for large values of x.

2.2.2 Exercises

Exercise 2.2.5:* *Write $3\cos(2x) + 3\sin(2x)$ in the form $R\cos(\beta x - \delta)$.*

Exercise 2.2.6: *Write $2\cos(3x) + \sin(3x)$ in the form $R\cos(\beta x - \delta)$.*

Exercise 2.2.7: *Write $3\cos(x) - 4\sin(x)$ in the form $R\cos(\beta x - \delta)$.*

Exercise 2.2.8: *Show that $e^{2x}\cos(x)$ and $e^{2x}\sin(x)$ are linearly independent.*

Exercise 2.2.9: *Find the general solution of* $2y'' + 50y = 0$.

Exercise 2.2.10: *Find the general solution of* $y'' - 6y' + 13y = 0$.

Exercise 2.2.11: *Find the solution to* $y'' - 2y' + 5y = 0$ *with* $y(0) = 3$ *and* $y'(0) = 2$.

Exercise 2.2.12: *Find the general solution of* $y'' + 2y' - 3y = 0$.

Exercise 2.2.13:* *Find the solution to* $2y'' + y' + y = 0$, $y(0) = 1$, $y'(0) = -2$.

Exercise 2.2.14:* *Find the solution to* $z''(t) = -2z'(t) - 2z(t)$, $z(0) = 2$, $z'(0) = -2$.

Exercise 2.2.15: *Let us revisit the Cauchy–Euler equations of Exercise 2.1.27 on page 123. Suppose now that* $(b-a)^2 - 4ac < 0$. *Find a formula for the general solution of* $ax^2y'' + bxy' + cy = 0$. *Hint: Note that* $x^r = e^{r \ln x}$.

Exercise 2.2.16: *Construct an equation such that* $y = C_1 e^{-2x}\cos(3x) + C_2 e^{-2x}\sin(3x)$ *is the general solution.*

Exercise 2.2.17:* *Find a second order, constant coefficient differential equation with general solution given by* $y(t) = C_1 e^{x}\cos(2x) + C_2 e^{2x}\sin(x)$ *or explain why there is no such thing.*

Exercise 2.2.18: *Find a second order, constant coefficient differential equation with general solution given by* $y(t) = C_1 e^{x}\cos(2x) + C_2 e^{x}\sin(2x)$ *or explain why there is no such thing.*

Exercise 2.2.19: *Find the solution to the initial value problem*

$$y'' + 4y' + 5y = 0 \qquad y(0) = 3,\ y'(0) = -1.$$

Determine a value T *so that* $|y(x)| < 0.02$ *for all* $x > T$.

Exercise 2.2.20: *Find the solution to the initial value problem*

$$y'' + 6y' + 13y = 0 \qquad y(0) = 4,\ y'(0) = 7.$$

Determine a value T *so that* $|y(x)| < 0.01$ *for all* $x > T$.

2.3 Repeated Roots and Reduction of Order

Attribution: [JL], §2.1, 2.2.

Learning Objectives

After this section, you will be able to:

- Find the general solution to a second order constant coefficient equation with repeated roots,
- Apply the method of reduction of order to generate a second solution to an equation given one solution, and
- Solve Euler equations using the method of reduction of order.

The last case we have to handle for solving all second order linear constant coefficient equations is the case where $b^2 - 4ac = 0$ in the equation

$$ay'' + by' + cy = 0.$$

When we try to find the characteristic equation and find solutions to this equation, we get a double root at r_1, so that the characteristic polynomial is $(r - r_1)^2$. For this, we get that $e^{r_1 x}$ is a solution. However, that's the only solution we get. We need to have two linearly independent solutions in order to get the general solution to the differential equation, so we need to find some method to get another solution. The standard method, and the one we apply here is *reduction of order*. Let's see how this works through an example.

Example 2.3.1: Find two linearly independent solutions to the differential equation

$$y'' + 2y' + y = 0.$$

Solution: To start, we find the first solution using our original method. The characteristic equation here is $r^2 + 2r + 1 = 0$, which is $(r + 1)^2$. Therefore, we have a double root at $r = -1$, so that $y_1(x) = e^{-x}$ is a solution.

To find a second solution, the reduction of order method suggests that we try to plug in $y = v(x)e^{-x}$ for an unknown function $v(x)$. The goal is to figure out an equation that v must satisfy to see if this leads us to a second solution to the original equation. We can compute the first two derivatives of $y = v(x)e^{-x}$

$$\begin{aligned} y(x) &= v(x)e^{-x} \\ y'(x) &= v'(x)e^{-x} - v(x)e^{-x} \\ y''(x) &= v''(x)e^{-x} - 2v'(x)e^{-x} + v(x)e^{-x} \end{aligned}$$

and then plug them into the original differential equation

$$\begin{aligned} 0 &= y'' + 2y' + y \\ &= (v''(x)e^{-x} - 2v'(x)e^{-x} + v(x)e^{-x}) + 2(v'(x)e^{-x} - v(x)e^{-x}) + v(x)e^{-x} \\ &= v''(x)e^{-x} + v'(x)(-2e^{-x} + 2e^{-x}) + v(x)(e^{-x} - 2e^{-x} + e^{-x}) \\ &= v''(x)e^{-x} \end{aligned}.$$

Since e^{-x} is never zero, this means we must have $v''(x) = 0$. This is still a second order equation, but we know how to solve it. We can integrate both sides twice to get that $v(x) = Ax + B$ for any constants A and B.

Our goal with all of this was to find a solution y of the form $v(x)e^{-x}$. The set up here means that $y = (Ax + B)e^{-x}$ will solve the differential equation. Since we already knew that Be^{-x} was a solution, the new information we gained here was that Axe^{-x}, or in particular, xe^{-x} is a solution to the differential equation. Thus, our two solutions are $y_1(x) = e^{-x}$ and $y_2(x) = xe^{-x}$. ⌟

***Exercise* 2.3.1:** *Check that e^{-x} and xe^{-x} both solve $y'' + 2y' + y = 0$, and that these solutions are linearly independent.*

The *reduction of order method* applies more generally to any second order linear homogenous equation and the goal is the same: use one solution of the differential equation to generate another one. The idea is that if we somehow found y_1 as a solution of $y'' + p(x)y' + q(x)y = 0$ we try a second solution of the form $y_2(x) = y_1(x)v(x)$. We just need to find v. We plug y_2 into the equation:

$$\begin{aligned} 0 = y_2'' + p(x)y_2' + q(x)y_2 &= y_1''v + 2y_1'v' + y_1v'' + p(x)(y_1'v + y_1v') + q(z)y_1v \\ &= y_1v'' + (2y_1' + p(x)y_1)v' + \cancelto{0}{\big(y_1'' + p(x)y_1' + q(x)y_1\big)}v. \end{aligned}$$

In other words, $y_1v'' + (2y_1' + p(x)y_1)v' = 0$. Using $w = v'$ we have the first order linear equation $y_1w' + (2y_1' + p(x)y_1)w = 0$. After solving this equation for w (integrating factor), we find v by antidifferentiating w. We then form y_2 by computing y_1v. For example, suppose we somehow know $y_1 = x$ is a solution to $y'' + x^{-1}y' - x^{-2}y = 0$. The equation for w is then $xw' + 3w = 0$. We find a solution, $w = Cx^{-3}$, and we find an antiderivative $v = \frac{-C}{2x^2}$. Hence $y_2 = y_1v = \frac{-C}{2x}$. Any C works and so $C = -2$ makes $y_2 = 1/x$. Thus, the general solution is $y = C_1x + C_2 1/x$.

The easiest way to work out these problems is to remember that we need to try $y_2(x) = y_1(x)v(x)$ and find $v(x)$ as we did above. Also, the technique works for higher order equations too: you get to reduce the order for each solution you find.

In summary, for constant coefficient equations with a repeated root, the reduction of order method will always give the equation $v'' = 0$, and so the solution is $v(x) = Ax + B$. Multiplying by the y_1 solution e^{rx} gives that xe^{rx} is the other solution. Therefore, the general solution for repeated root equations is always of the form

$$y = C_1e^{r_1x} + C_2xr^{r_1x}.$$

Example 2.3.2: Find the general solution of

$$y'' - 8y' + 16y = 0.$$

Solution: The characteristic equation is $r^2 - 8r + 16 = (r - 4)^2 = 0$. The equation has a double root $r_1 = r_2 = 4$. The general solution is, therefore,

$$y = (C_1 + C_2x)\, e^{4x} = C_1e^{4x} + C_2xe^{4x}.$$

***Exercise* 2.3.2:** *Check that e^{4x} and xe^{4x} are linearly independent.*

That e^{4x} solves the equation is clear. If xe^{4x} solves the equation, then we know we are done. Let us compute $y' = e^{4x} + 4xe^{4x}$ and $y'' = 8e^{4x} + 16xe^{4x}$. Plug in

$$y'' - 8y' + 16y = 8e^{4x} + 16xe^{4x} - 8(e^{4x} + 4xe^{4x}) + 16xe^{4x} = 0.$$

In some sense, a doubled root rarely happens. If coefficients are picked randomly, a doubled root is unlikely. There are, however, some natural phenomena where a doubled root does happen, so we cannot just dismiss this case. In addition, there are specific physical applications that involve the double root problem, which we will discuss in Section 2.4. Finally, the solution with a doubled root can be thought of as an approximation of the solution with two roots that are very close together, and the behavior of this solution will approximate "nearby" solutions as well.

Example 2.3.3: Find the solution $y(t)$ to the initial value problem

$$y'' + 6y' + 9y = 0 \qquad y(0) = 2,\ y'(0) = -3.$$

Solution: The characteristic polynomials for this differential equation is

$$r^2 + 6r + 9$$

which factors as $(r+3)^2$, so that we have a double root at -3. With the work done previously, we know that the general solution is

$$y(t) = (C_1 + C_2 t)e^{-3t} = C_1 e^{-3t} + C_2 t e^{-3t}.$$

If we use the initial conditions, we can set $t = 0$ to get that

$$2 = y(0) = C_1 e^0$$

so that $C_1 = 2$. Differentiating the general solution gives that

$$y'(t) = -3C_1 e^{-3t} + C_2 e^{-3t} - 3C_2 t e^{-3t}$$

and plugging in zero here gives

$$-3 = y'(0) = -3C_1 + C_2.$$

Since $C_1 = 2$, this implies that $C_2 = 3$. Therefore, the solution to this initial value problem is

$$y(t) = 2e^{-3t} + 3te^{-3t}.$$

2.3.1 Exercises

***Exercise* 2.3.3:** *Find the general solution to* $y'' + 4y' + 4y = 0$.

Exercise* 2.3.4: *Find the general solution to* $y'' - 6y' + 9y = 0$.

***Exercise* 2.3.5:** *Find the solution to* $y'' + 6y' + 9y = 0$ *with* $y(0) = 3$ *and* $y'(0) = -1$.

***Exercise* 2.3.6:** *Solve* $y'' - 8y' + 16y = 0$ *for* $y(0) = 2$, $y'(0) = 0$.

***Exercise* 2.3.7:** *Find the general solution of* $y'' = 0$ *using the methods of this section.*

***Exercise* 2.3.8:** *The method of this section applies to equations of other orders than two. We will see higher orders later. Try to solve the first order equation* $2y' + 3y = 0$ *using the methods of this section.*

***Exercise* 2.3.9:** *Consider the second-order DE*

$$ty'' + (4t + 2)y' + (4t + 4)y = 0. \tag{2.5}$$

a) *Does the superposition principle apply to this DE? Give a one- or two-sentence explanation wither way.*

b) *Find a value of* r *so that* $y = e^{rt}$ *is a solution to* (2.5)

c) *Using your result from the previous page, apply* **reduction of order** *to find the general solution to* (2.5).

Exercise* 2.3.10** (Euler Equations):*** *Consider the differential equation* $x^2y'' + 3xy' - 3y = 0$.

a) *Verify that* $y_1(x) = x$ *is a solution.*

b) *Use reduction of order to find a second linearly independent solution.*

c) *Write out the general solution.*

Exercise* 2.3.11** (Euler Equations):*** *Consider the differential equation* $x^2y'' + 4xy' - 2y = 0$.

a) *Verify that* $y_1(x) = \frac{1}{x}$ *is a solution.*

b) *Use reduction of order to find a second linearly independent solution.*

c) *Write out the general solution.*

Exercise* 2.3.12** (Euler Equations):*** *Consider the differential equation* $x^2y'' - 6xy' - 10y = 0$.

a) *Verify that* $y_1(x) = x^2$ *is a solution.*

b) *Use reduction of order to find a second linearly independent solution.*

c) *Write out the general solution.*

Exercise 2.3.13: *Write down a differential equation with general solution $y = at^2 + bt^{-3}$, or explain why there is no such thing.*

Exercise 2.3.14: *Find the solution to $y'' - (2\alpha)y' + \alpha^2 y = 0$, $y(0) = a$, $y'(0) = b$, where α, a, and b are real numbers.*

Exercise 2.3.15 (reduction of order)***:*** *Suppose y_1 is a solution to $y'' + p(x)y' + q(x)y = 0$. By directly plugging into the equation, show that*

$$y_2(x) = y_1(x) \int \frac{e^{-\int p(x)\,dx}}{\big(y_1(x)\big)^2}\,dx$$

is also a solution.

Exercise 2.3.16 (Chebyshev's equation of order 1)***:*** *Take $(1 - x^2)y'' - xy' + y = 0$.*

a) Show that $y = x$ is a solution.

b) Use reduction of order to find a second linearly independent solution.

c) Write down the general solution.

Exercise 2.3.17 (Hermite's equation of order 2)***:*** *Take $y'' - 2xy' + 4y = 0$.*

a) Show that $y = 1 - 2x^2$ is a solution.

b) Use reduction of order to find a second linearly independent solution. (It's OK to leave a definite integral in the formula.)

c) Write down the general solution.

The rest of these exercises can be solved using any of the methods discussed in the last three sections. Pick the appropriate method in order to solve the problem.

Exercise 2.3.18: *Find the general solution of $y'' + 5y' - 6y = 0$.*

Exercise 2.3.19: *Find the general solution of $y'' - 2y' + 2y = 0$.*

Exercise 2.3.20: *Find the general solution of $y'' + 4y' + 4y = 0$.*

Exercise 2.3.21: *Find the general solution of $y'' + 4y' + 5y = 0$.*

Exercise 2.3.22: *Find the solution to $y'' - 6y' + 13y = 0$ with $y(0) = 2$ and $y'(0) = 1$.*

Exercise 2.3.23: *Find the solution to $y'' + 4y' - 12y = 0$ with $y(0) = -1$ and $y'(0) = 3$.*

Exercise 2.3.24: *Find the solution to $y'' - 6y' + 9y = 0$ with $y(0) = -4$ and $y'(0) = -1$.*

2.4 Mechanical vibrations

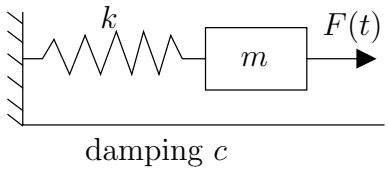
Attribution: [JL], §2.4.

Learning Objectives

After this section, you will be able to:

- Write second-order differential equations to model physical situations,
- Classify a mechanical oscillation as undamped, underdamped, critically damped, or overdamped, and
- Use the solution to a differential equation to describe the resulting physical motion.

In the last few sections, we have discussed all of the different possible solutions to constant coefficient second order differential equations, whether the roots of the characteristic polynomial real and distinct, complex, or repeated. Now, we want to look at applications of these equations, now that we know how to solve them. Since Newton's Second Law $F = ma$ involves the second derivative of position (acceleration), it is reasonable that a lot of physical systems will be defined by second order differential equations.



Our first example is a mass on a spring. Suppose we have a mass $m > 0$ (in kilograms) connected by a spring with spring constant $k > 0$ (in newtons per meter) to a fixed wall. There may be some external force $F(t)$ (in newtons) acting on the mass. Finally, there is some friction measured by $c \geq 0$ (in newton-seconds per meter) as the mass slides along the floor (or perhaps a damper is connected).

Let x be the displacement of the mass ($x = 0$ is the rest position), with x growing to the right (away from the wall). The force exerted by the spring is proportional to the compression of the spring by Hooke's law. Therefore, it is kx in the negative direction. Similarly the amount of force exerted by friction is proportional to the velocity of the mass. By Newton's second law we know that force equals mass times acceleration and hence $mx'' = F(t) - cx' - kx$ or

$$mx'' + cx' + kx = F(t).$$

This is a linear second order constant coefficient ODE. We say the motion is

(i) *forced*, if $F \not\equiv 0$ (if F is not identically zero),

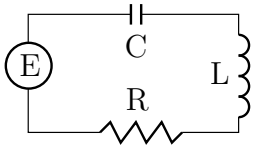
(ii) *unforced* or *free*, if $F \equiv 0$ (if F is identically zero),

(iii) *damped*, if $c > 0$, and

(iv) *undamped*, if $c = 0$.

This system appears in lots of applications even if it does not at first seem like it. Many real-world scenarios can be simplified to a mass on a spring. For example, a bungee jump

setup is essentially a mass and spring system (you are the mass). It would be good if someone did the math before you jump off the bridge, right? Let us give two other examples.



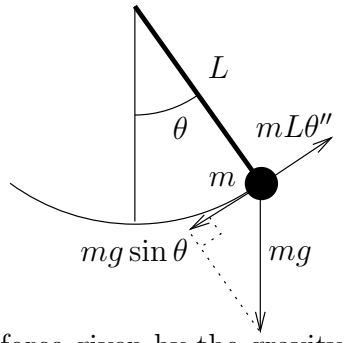
Here is an example for electrical engineers. Consider the pictured RLC circuit. There is a resistor with a resistance of R ohms, an inductor with an inductance of L henries, and a capacitor with a capacitance of C farads. There is also an electric source (such as a battery) giving a voltage of $E(t)$ volts at time t (measured in seconds). Let $Q(t)$ be the charge in coulombs on the capacitor and $I(t)$ be the current in the circuit. The relation between the two is $Q' = I$. By elementary principles we find $LI' + RI + {}^{Q}/_{C} = E$. Since $Q' = I$, this means that $I' = Q''$, and we can write this equation as

$$LQ''(t) + RQ'(t) + \frac{1}{C}Q(t) = E(t).$$

We can also write this a different way by differentiating the entire equation in t to get a second order equation for $I(t)$:

$$LI''(t) + RI'(t) + \frac{1}{C}I(t) = E'(t).$$

This is a nonhomogeneous second order constant coefficient linear equation. As L, R, and C are all positive, this system behaves just like the mass and spring system. Position of the mass is replaced by current. Mass is replaced by inductance, damping is replaced by resistance, and the spring constant is replaced by one over the capacitance. The change in voltage becomes the forcing function—for constant voltage this is an unforced motion.



Our next example behaves like a mass and spring system only approximately. Suppose a mass m hangs on a pendulum of length L. We seek an equation for the angle $\theta(t)$ (in radians). Let g be the force of gravity. Elementary physics mandates that the equation is

$$\theta'' + \frac{g}{L}\sin\theta = 0.$$

Let us derive this equation using Newton's second law: force equals mass times acceleration. The acceleration is $L\theta''$ and mass is m. So $mL\theta''$ has to be equal to the tangential component of the force given by the gravity, which is $mg\sin\theta$ in the opposite direction. So $mL\theta'' = -mg\sin\theta$. The m curiously cancels from the equation.

Now we make our approximation. For small θ we have that approximately $\sin\theta \approx \theta$. This can be seen by looking at the graph. In Figure 2.3 on the facing page we can see that for approximately $-0.5 < \theta < 0.5$ (in radians) the graphs of $\sin\theta$ and θ are almost the same.

Therefore, when the swings are small, θ is small and we can model the behavior by the simpler linear equation

$$\theta'' + \frac{g}{L}\theta = 0.$$

The errors from this approximation build up. So after a long time, the state of the real-world system might be substantially different from our solution. Also we will see that in a mass-spring system, the amplitude is independent of the period. This is not true for a pendulum.

Nevertheless, for reasonably short periods of time and small swings (that is, only small angles θ), the approximation is reasonably good.

In real-world problems it is often necessary to make these types of simplifications. We must understand both the mathematics and the physics of the situation to see if the simplification is valid in the context of the questions we are trying to answer.

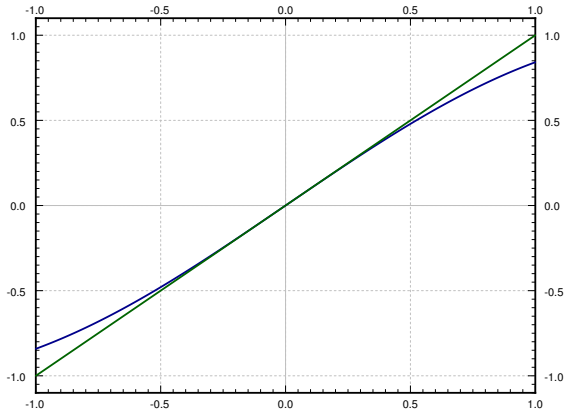


Figure 2.3: The graphs of $\sin\theta$ *and* θ *(in radians).*

2.4.1 Free undamped motion

In this section we only consider free or unforced motion, as we do not know yet how to solve nonhomogeneous equations. Let us start with undamped motion where $c = 0$. The equation is

$$mx'' + kx = 0.$$

We divide by m and let $\omega_0 = \sqrt{k/m}$ to rewrite the equation as

$$x'' + \omega_0^2 x = 0.$$

The general solution to this equation is

$$x(t) = A\cos(\omega_0 t) + B\sin(\omega_0 t).$$

By a trigonometric identity that we discussed previously in § 2.2,

$$A\cos(\omega_0 t) + B\sin(\omega_0 t) = C\cos(\omega_0 t - \delta),$$

for two constants C and γ. Earlier, we found that we can compute these constants as $C = \sqrt{A^2 + B^2}$ and $\tan\delta = B/A$. Therefore, we let C and δ be our arbitrary constants and write $x(t) = C\cos(\omega_0 t - \delta)$.

***Exercise* 2.4.1:** *Justify the identity* $A\cos(\omega_0 t) + B\sin(\omega_0 t) = C\cos(\omega_0 t - \delta)$ *and verify the equations for* C *and* δ*. Hint: Start with* $\cos(\alpha - \beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta)$ *and multiply by* C*. Then what should* α *and* β *be?*

While it is generally easier to use the first form with A and B to solve for the initial conditions, the second form is much more natural to use for interpretation of physical systems, since the constants C and δ have nice physical interpretation. Write the solution as

$$x(t) = C\cos(\omega_0 t - \delta).$$

This is a pure-frequency oscillation (a sine wave). The *amplitude* is C, ω_0 is the (angular) *frequency*, and δ is the so-called *phase shift*. The phase shift just shifts the graph left or right. We call ω_0 the *natural (angular) frequency*. This entire setup is called *simple harmonic motion*.

Let us pause to explain the word *angular* before the word *frequency*. The units of ω_0 are radians per unit time, not cycles per unit time as is the usual measure of frequency. Because one cycle is 2π radians, the usual frequency is given by $\frac{\omega_0}{2\pi}$. It is simply a matter of where we put the constant 2π, and that is a matter of taste.

The *period* of the motion is one over the frequency (in cycles per unit time) and hence $\frac{2\pi}{\omega_0}$. That is the amount of time it takes to complete one full cycle.

Example 2.4.1: Suppose that $m = 2\,\text{kg}$ and $k = 8\,\text{N/m}$. The whole mass and spring setup is sitting on a truck that was traveling at $1\,\text{m/s}$. The truck crashes and hence stops. The mass was held in place 0.5 meters forward from the rest position. During the crash the mass gets loose. That is, the mass is now moving forward at $1\,\text{m/s}$, while the other end of the spring is held in place. The mass therefore starts oscillating. What is the frequency of the resulting oscillation? What is the amplitude? The units are the mks units (meters-kilograms-seconds).

Solution: The setup means that the mass was at half a meter in the positive direction during the crash and relative to the wall the spring is mounted to, the mass was moving forward (in the positive direction) at $1\,\text{m/s}$. This gives us the initial conditions.

So the equation with initial conditions is

$$2x'' + 8x = 0, \qquad x(0) = 0.5, \qquad x'(0) = 1.$$

We directly compute $\omega_0 = \sqrt{k/m} = \sqrt{4} = 2$. Hence the angular frequency is 2. The usual frequency in Hertz (cycles per second) is $2/2\pi = 1/\pi \approx 0.318$.

The general solution is

$$x(t) = A\cos(2t) + B\sin(2t).$$

Letting $x(0) = 0.5$ means $A = 0.5$. Then $x'(t) = -2(0.5)\sin(2t) + 2B\cos(2t)$. Letting $x'(0) = 1$ we get $B = 0.5$. Therefore, the amplitude is $C = \sqrt{A^2 + B^2} = \sqrt{0.25 + 0.25} = \sqrt{0.5} \approx 0.707$. The solution is

$$x(t) = 0.5\cos(2t) + 0.5\sin(2t).$$

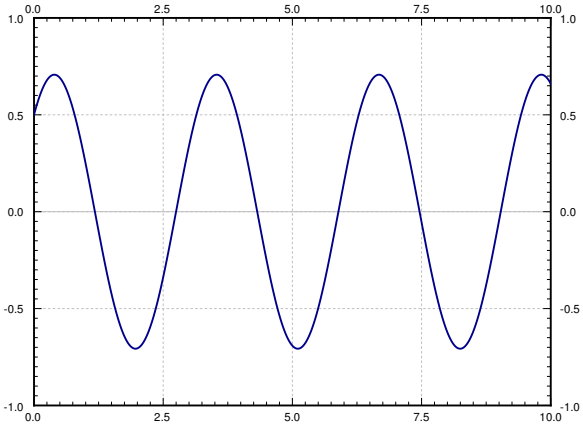


Figure 2.4: Simple undamped oscillation.

A plot of $x(t)$ is shown in Figure 2.4. □

In general, for free undamped motion, a solution of the form

$$x(t) = A\cos(\omega_0 t) + B\sin(\omega_0 t),$$

corresponds to the initial conditions $x(0) = A$ and $x'(0) = \omega_0 B$. Therefore, it is easy to figure out A and B from the initial conditions. The amplitude and the phase shift can then be computed from A and B. In the example, we have already found the amplitude C. Let us compute the phase shift. We know that $\tan\delta = B/A = 1$. We take the arctangent of 1 and get $\pi/4$ or approximately 0.785. We still need to check if this δ is in the correct quadrant

(and add π to δ if it is not). Since both A and B are positive, then δ should be in the first quadrant, $\pi/4$ radians is in the first quadrant, so $\delta = \pi/4$.

Note: Many calculators and computer software have not only the `atan` function for arctangent, but also what is sometimes called `atan2`. This function takes two arguments, B and A, and returns a δ in the correct quadrant for you.

2.4.2 Free damped motion

Let us now focus on damped motion. Let us rewrite the equation

$$mx'' + \gamma x' + kx = 0,$$

as

$$x'' + 2px' + \omega_0^2 x = 0,$$

where

$$\omega_0 = \sqrt{\frac{k}{m}}, \qquad p = \frac{\gamma}{2m}.$$

The characteristic equation is

$$r^2 + 2pr + \omega_0^2 = 0.$$

Using the quadratic formula we get that the roots are

$$r = -p \pm \sqrt{p^2 - \omega_0^2}.$$

The form of the solution depends on whether we get complex or real roots. We get real roots if and only if the following number is nonnegative:

$$p^2 - \omega_0^2 = \left(\frac{\gamma}{2m}\right)^2 - \frac{k}{m} = \frac{\gamma^2 - 4km}{4m^2}.$$

The sign of $p^2 - \omega_0^2$ is the same as the sign of $\gamma^2 - 4km$. Thus we get real roots if and only if $\gamma^2 - 4km$ is nonnegative, or in other words if $\gamma^2 \geq 4km$. If these look familiar, that is not surprising, as they are the same as the conditions we had for the different types of roots in second order constant coefficient equations.

Overdamping

When $\gamma^2 - 4km > 0$, the system is *overdamped*. In this case, there are two distinct real roots r_1 and r_2. Both roots are negative: As $\sqrt{p^2 - \omega_0^2}$ is always less than p, then $-p \pm \sqrt{p^2 - \omega_0^2}$ is negative in either case.

The solution is

$$x(t) = C_1 e^{r_1 t} + C_2 e^{r_2 t}.$$

Since r_1, r_2 are negative, $x(t) \to 0$ as $t \to \infty$. Thus the mass will tend towards the rest position as time goes to infinity. For a few sample plots for different initial conditions, see Figure 2.5 on the following page.

No oscillation happens. In fact, the graph crosses the x-axis at most once. To see why, we try to solve $0 = C_1e^{r_1t}+C_2e^{r_2t}$. Therefore, $C_1e^{r_1t} = -C_2e^{r_2t}$ and using laws of exponents we obtain

$$\frac{-C_1}{C_2} = e^{(r_2-r_1)t}.$$

This equation has at most one solution $t \geq 0$. For some initial conditions the graph never crosses the x-axis, as is evident from the sample graphs.

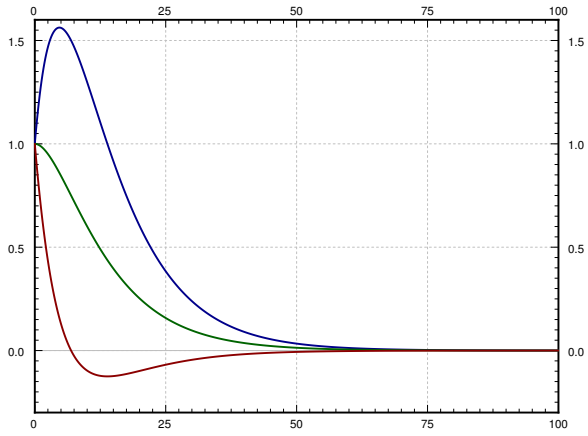


Figure 2.5: Overdamped motion for several different initial conditions.

Example 2.4.2: Suppose the mass is released from rest. That is $x(0) = x_0$ and $x'(0) = 0$. Then

$$x(t) = \frac{x_0}{r_1 - r_2}\left(r_1e^{r_2t} - r_2e^{r_1t}\right).$$

It is not hard to see that this satisfies the initial conditions.

Critical damping

When $\gamma^2 - 4km = 0$, the system is *critically damped*. In this case, there is one root of multiplicity 2 and this root is $-p$. Our solution is

$$x(t) = C_1e^{-pt} + C_2te^{-pt}.$$

The behavior of a critically damped system is very similar to an overdamped system. After all a critically damped system is in some sense a limit of overdamped systems. Even though our models are only approximations of the real world problem, the idea of critical damping can be helpful in optimizing systems. Figure 2.6 shows how the solution to

$$x'' + \gamma x' + x = 0$$

for different values of γ and initial conditions $x(0) = 4$ and $x'(0) = 0$. This solution is critically damped if $\gamma = 2$, as that will give us a repeated root in the characteristic equation. Comparing these solutions, we see that the critically damped solution gets back to equilibrium faster than any of the more overdamped solution. When trying to design a

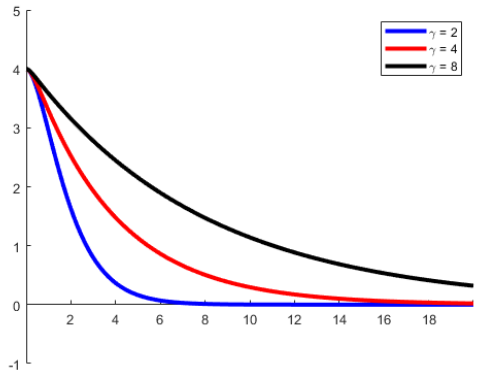


Figure 2.6: Overdamped and critically damped motion for $x'' + \gamma x' + x = 0$ for $\gamma = 2, 4, 8$.

system, if we want it to settle back to the zero point as quickly as possible, then we should try to get as closed to critically damped as possible. Even though we are always a little bit underdamped or a little bit overdamped, getting as close as possible will give the best possible result for returning to equilibrium.

Underdamping

When $\gamma^2 - 4km < 0$, the system is *underdamped*. In this case, the roots are complex.

$$\begin{aligned} r &= -p \pm \sqrt{p^2 - \omega_0^2} \\ &= -p \pm \sqrt{-1}\sqrt{\omega_0^2 - p^2} \\ &= -p \pm i\omega_1, \end{aligned}$$

where $\omega_1 = \sqrt{\omega_0^2 - p^2}$. Our solution is

$$x(t) = e^{-pt}\big(A\cos(\omega_1 t) + B\sin(\omega_1 t)\big),$$

or

$$x(t) = Ce^{-pt}\cos(\omega_1 t - \delta).$$

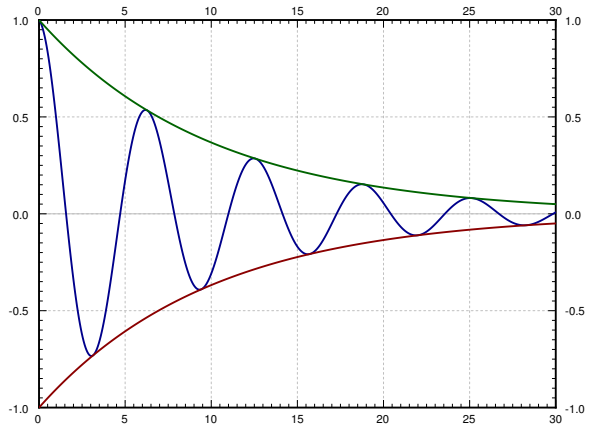


Figure 2.7: Underdamped motion with the envelope curves shown.

An example plot is given in Figure 2.7. Note that we still have that $x(t) \to 0$ as $t \to \infty$.

The figure also shows the *envelope curves* Ce^{-pt} and $-Ce^{-pt}$. The solution is the oscillating line between the two envelope curves. The envelope curves give the maximum amplitude of the oscillation at any given point in time. For example, if you are bungee jumping, you are really interested in computing the envelope curve as not to hit the concrete with your head.

The phase shift δ shifts the oscillation left or right, but within the envelope curves (the envelope curves do not change if δ changes).

Notice that the angular *pseudo-frequency*[*] or *quasi-frequency* becomes smaller when the damping γ (and hence p) becomes larger. This makes sense. When we change the damping just a little bit, we do not expect the behavior of the solution to change dramatically. If we keep making γ larger, then at some point the solution should start looking like the solution for critical damping or overdamping, where no oscillation happens. So if γ^2 approaches $4km$, we want ω_1 to approach 0. Since $\omega_1 = \sqrt{\omega_0^2 - p^2}$ with $p = \frac{\gamma}{2m}$ and $\omega_0 = \sqrt{km}$, we have that

$$\omega_1 = \sqrt{\frac{k}{m} - \frac{\gamma^2}{4m^2}} = \sqrt{\frac{4mk - \gamma^2}{4m^2}},$$

which does go to zero as γ^2 gets closer to $4mk$.

On the other hand, when γ gets smaller, ω_1 approaches ω_0 (ω_1 is always smaller than ω_0), and the solution looks more and more like the steady periodic motion of the undamped case. The envelope curves become flatter and flatter as γ (and hence p) goes to 0.

[*]We do not call ω_1 a frequency since the solution $x(t)$ is not really a periodic function.

2.4.3 Exercises

***Exercise* 2.4.2:** *Consider a mass and spring system with a mass $m = 2$, spring constant $k = 3$, and damping constant $\gamma = 1$.*

a) *Set up and find the general solution of the system.*

b) *Is the system underdamped, overdamped or critically damped?*

c) *If the system is not critically damped, find a γ that makes the system critically damped.*

***Exercise* 2.4.3:** *Do Exercise 2.4.2 for $m = 3$, $k = 12$, and $\gamma = 12$.*

***Exercise* 2.4.4:** *Using the mks units (meters-kilograms-seconds), suppose you have a spring with spring constant $4\,\mathrm{N/m}$. You want to use it to weigh items. Assume no friction. You place the mass on the spring and put it in motion.*

a) *You count and find that the frequency is 0.8 Hz (cycles per second). What is the mass?*

b) *Find a formula for the mass m given the frequency ω in Hz.*

Exercise* 2.4.5: *A mass of 2 kilograms is on a spring with spring constant k newtons per meter with no damping. Suppose the system is at rest and at time $t = 0$ the mass is kicked and starts traveling at 2 meters per second. How large does k have to be to so that the mass does not go further than 3 meters from the rest position?*

***Exercise* 2.4.6:** *Suppose we add possible friction to Exercise 2.4.4. Further, suppose you do not know the spring constant, but you have two reference weights 1 kg and 2 kg to calibrate your setup. You put each in motion on your spring and measure the quasi-frequency. For the 1 kg weight you measured 1.1 Hz, for the 2 kg weight you measured 0.8 Hz.*

a) *Find k (spring constant) and γ (damping constant).*

b) *Find a formula for the mass in terms of the frequency in Hz.* Note that there may be more than one possible mass for a given frequency.

c) *For an unknown object you measured 0.2 Hz, what is the mass of the object? Suppose that you know that the mass of the unknown object is more than a kilogram.*

***Exercise* 2.4.7:** *Suppose you wish to measure the friction a mass of 0.1 kg experiences as it slides along a floor (you wish to find γ). You have a spring with spring constant $k = 5\,\mathrm{N/m}$. You take the spring, you attach it to the mass and fix it to a wall. Then you pull on the spring and let the mass go. You find that the mass oscillates with quasi-frequency 1 Hz. What is the friction?*

Exercise* 2.4.8: *A 5000 kg railcar hits a bumper (a spring) at $1\,\mathrm{m/s}$, and the spring*

compresses by 0.1 m. Assume no damping.

a) *Find k.*

b) *How far does the spring compress when a 10000 kg railcar hits the spring at the same speed?*

c) *If the spring would break if it compresses further than 0.3 m, what is the maximum mass of a railcar that can hit it at 1 $^{m}/_{s}$?*

d) *What is the maximum mass of a railcar that can hit the spring without breaking at 2 $^{m}/_{s}$?*

***Exercise* 2.4.9:** *When attached to a spring, a 2 kg mass stretches the spring by 0.49 m.*

a) *What is the spring constant of this spring? Use 9.8 $^{m}/_{s^2}$ as the gravity constant.*

b) *This mass is allowed to come to rest, lifted up by 0.4 m and then released. If there is no damping, set up and solve an initial value problem for the position of the mass as a function of time.*

c) *For a next experiment, you attach a dampener of coefficient 16 $^{Ns}/_{m}$ to the system, and give the same initial condition. Set up and solve an initial value problem for the position of the mass. What type of "dampening" would be used to characterize this situation?*

Exercise* 2.4.10: *A mass of m kg is on a spring with $k = 3\,^{N}/_{m}$ and $c = 2\,^{Ns}/_{m}$. Find the mass m_0 for which there is critical damping. If $m < m_0$, does the system oscillate or not, that is, is it underdamped or overdamped?*

Exercise* 2.4.11: *Suppose we have an RLC circuit with a resistor of 100 milliohms (0.1 ohms), inductor of inductance of 50 millihenries (0.05 henries), and a capacitor of 5 farads, with constant voltage.*

a) *Set up the ODE equation for the current I.*

b) *Find the general solution.*

c) *Solve for $I(0) = 10$ and $I'(0) = 0$.*

***Exercise* 2.4.12:** *For RLC circuits, we can use either charge or current to set up the equation. Let's see how the two of those compare.*

a) *Assume that we have an RLC circuit with a 30 millihenry inductor, a 120 milliohm resistor, and a capacitor with capacitance $^{20}/_{3}$ F. Set up a differential equation for the charge on the capacitor as a function of time.*

b) *Use the same circuit to set up a differential equation for the current through the circuit as a function of time. How do these equations relate?*

c) *Find the general solution to each of these equations.*

d) *Solve the initial value problem for the charge with* $Q(0) = {}^1\!/_2 C$ *and* $Q'(0) = 0$.

e) *Using the fact that* $I = Q'$, *determine the appropriate initial conditions needed for* I *in order to solve for the current in this same setup (with those initial values for charge).*

f) *Now, we'll do the same in the other direction. Solve the initial value problem for current with* $I(0) = 2A$ *and* $I'(0) = 1{}^{\mathrm{A}}\!/_{\mathrm{s}}$, *and see what the initial conditions would be for* $Q(t)$ *for this setup.*

***Exercise* 2.4.13:** *Assume that the system* $my'' + \gamma y' + ky = 0$ *is either critically or overdamped. Prove that the solution can pass through zero* at most once, *regardless of initial conditions.* Hint: *Try to find all values of* t *for which* $y(t) = 0$, *given the form of the solution.*

2.5 Nonhomogeneous equations

Attribution: [JL], §2.5.

Learning Objectives

After this section, you will be able to:

- Find the corresponding homogeneous equation for a non-homogeneous equation,
- Use the method of undetermined coefficients to solve non-homogeneous equations,
- Use variation of parameters to solve non-homogeneous equations, and
- Solve for the necessary coefficients to solve initial value problems for non-homogeneous equations.

2.5.1 Solving nonhomogeneous equations

We have solved linear constant coefficient homogeneous equations. What about nonhomogeneous linear ODEs? For example, the equations for forced mechanical vibrations, where we add a "forcing" term, which is a function on the right-hand side of the equation. That is, suppose we have an equation such as

$$y'' + 5y' + 6y = 2x + 1. \tag{2.6}$$

We will write $L[y] = 2x+1$, where $L[y]$ represents the entire left-hand side of $y''+5y'+6y$, when the exact form of the operator is not important. We solve (2.6) in the following manner. First, we find the general solution y_c to the *associated homogeneous equation*

$$y'' + 5y' + 6y = 0. \tag{2.7}$$

We call y_c the *complementary solution*. Next, we find a single *particular solution* y_p to (2.6) in some way (that is the point of this section). Then

$$y = y_c + y_p$$

is the general solution to (2.6). We have $L[y_c] = 0$ and $L[y_p] = 2x + 1$. As L is a *linear operator* we verify that y is a solution, $L[y] = L[y_c + y_p] = L[y_c] + L[y_p] = 0 + (2x + 1)$. Let us see why we obtain the *general* solution.

Let y_p and $\tilde{y}_p$ be two different particular solutions to (2.6). Write the difference as $w = y_p - \tilde{y}_p$. Then plug w into the left-hand side of the equation to get

$$w'' + 5w' + 6w = (y_p'' + 5y_p' + 6y_p) - (\tilde{y}_p'' + 5\tilde{y}_p' + 6\tilde{y}_p) = (2x + 1) - (2x + 1) = 0.$$

Using the operator notation the calculation becomes simpler. As L is a linear operator we write

$$L[w] = L[y_p - \tilde{y}_p] = L[y_p] - L[\tilde{y}_p] = (2x + 1) - (2x + 1) = 0.$$

So $w = y_p - \tilde{y}_p$ is a solution to (2.7), that is $Lw = 0$. However, we know what all solutions to $Lw = 0$ look like, as this is a homogeneous equation that we have solved previously. Therefore, any two solutions of (2.6) differ by a solution to the homogeneous equation (2.7). The solution $y = y_c + y_p$ includes *all* solutions to (2.6), since y_c is the general solution to the associated homogeneous equation.

Theorem 2.5.1

Let $L[y] = f(x)$ be a linear ODE (not necessarily constant coefficient). Let y_c be the complementary solution (the general solution to the associated homogeneous equation $L[y] = 0$) and let y_p be any particular solution to $L[y] = f(x)$. Then the general solution to $L[y] = f(x)$ is

$$y = y_c + y_p.$$

The moral of the story is that we can find the particular solution in any old way. If we find a different particular solution (by a different method, or simply by guessing), then we still get the same general solution. The formula may look different, and the constants we have to choose to satisfy the initial conditions may be different, but it is the same solution.

2.5.2 Undetermined coefficients

The trick is to somehow, in a smart way, guess one particular solution to (2.6). Note that $2x + 1$ is a polynomial, and the left-hand side of the equation (with all of the derivatives) will still be a polynomial if we let y be a polynomial of the same degree. Let us try

$$y_p = Ax + B.$$

We plug y_p into the left hand side to obtain

$$\begin{aligned} y_p'' + 5y_p' + 6y_p &= (Ax + B)'' + 5(Ax + B)' + 6(Ax + B) \\ &= 0 + 5A + 6Ax + 6B = 6Ax + (5A + 6B). \end{aligned}$$

So $6Ax + (5A + 6B) = 2x + 1$. If we match up the coefficients of x in this equation, we get that $6A = 2$ or $A = 1/3$. In order for the constant terms to match, we need that $5A + 6B = 1$. Since we know the value of A, this tells us that $B = -1/9$. That means $y_p = \frac{1}{3}x - \frac{1}{9} = \frac{3x-1}{9}$. Solving the complementary problem (exercise!) we get

$$y_c = C_1 e^{-2x} + C_2 e^{-3x}.$$

Hence the general solution to (2.6) is

$$y = C_1 e^{-2x} + C_2 e^{-3x} + \frac{3x - 1}{9}.$$

Now suppose we are further given some initial conditions. For example, $y(0) = 0$ and $y'(0) = 1/3$. First find $y' = -2C_1 e^{-2x} - 3C_2 e^{-3x} + 1/3$. Then

$$0 = y(0) = C_1 + C_2 - \frac{1}{9}, \qquad \frac{1}{3} = y'(0) = -2C_1 - 3C_2 + \frac{1}{3}.$$

We solve to get $C_1 = 1/3$ and $C_2 = -2/9$. The particular solution we want is

$$y(x) = \frac{1}{3}e^{-2x} - \frac{2}{9}e^{-3x} + \frac{3x-1}{9} = \frac{3e^{-2x} - 2e^{-3x} + 3x - 1}{9}.$$

***Exercise* 2.5.1:** *Check that y really solves the equation (2.6) and the given initial conditions.*

Note: A common mistake is to solve for constants using the initial conditions with y_c and only add the particular solution y_p after that. That will *not* work. You need to first compute $y = y_c + y_p$ and *only then* solve for the constants using the initial conditions.

A right-hand side consisting of exponentials, sines, and cosines can be handled similarly.

Example 2.5.1: One example of this is

$$y'' + 2y' + 2y = \cos(2x).$$

Solution: Let us find some y_p. We start by guessing that the solution includes some multiple of $\cos(2x)$. We try

$$y_p = A\cos(2x).$$

Plugging this into the differential equation gives

$$\underbrace{-4A\cos(2x)}_{y_p''} + 2\underbrace{\big(-2A\sin(2x)\big)}_{y_p'} + 2\underbrace{\big(A\cos(2x)\big)}_{y_p} = \cos(2x).$$

Simplifying this expression gives

$$-2A\cos(2x) - 4A\sin(2x) = \cos(2x)$$

and we have a problem. Since there is no sine term on the right-hand side, we are forced to pick $A = 0$, which means our non-homogeneous solution is zero, and that's not good. What happened here? In the previous example, when we differentiated a polynomial (as part of the y_p guess) the function stayed a polynomial, and so we did not add any new types of terms. In this case, however, when we differentiate the cosine term in our guess, it becomes a sine, which we did *not* have in our initial guess.

Thus, we will also want to add a multiple of $\sin(2x)$ to our guess since derivatives of cosine are sines. We try

$$y_p = A\cos(2x) + B\sin(2x).$$

We plug y_p into the equation and we get

$$\underbrace{-4A\cos(2x) - 4B\sin(2x)}_{y_p''} + 2\underbrace{\big(-2A\sin(2x) + 2B\cos(2x)\big)}_{y_p'} + 2\underbrace{\big(A\cos(2x) + 2B\sin(2x)\big)}_{y_p} = \cos(2x),$$

or

$$(-4A + 4B + 2A)\cos(2x) + (-4B - 4A + 2B)\sin(2x) = \cos(2x).$$

The left-hand side must equal to right-hand side. Namely, $-4A + 4B + 2A = 1$ and $-4B - 4A + 2B = 0$. So $-2A + 4B = 1$ and $2A + B = 0$. We can solve this system of equations to get that $A = -1/10$ and $B = 1/5$. So

$$y_p = A\cos(2x) + B\sin(2x) = \frac{-\cos(2x) + 2\sin(2x)}{10}.$$

Similarly, if the right-hand side contains exponentials we try exponentials. If

$$L[y] = e^{3x},$$

we try $y = Ae^{3x}$ as our guess and try to solve for A.

When the right-hand side is a multiple of sines, cosines, exponentials, and polynomials, we can use the product rule for differentiation to come up with a guess. We need to guess a form for y_p such that $L[y_p]$ is of the same form, and has all the terms needed to for the right-hand side. For example,

$$L[y] = (1 + 3x^2)\, e^{-x} \cos(\pi x).$$

For this equation, we guess

$$y_p = (A + Bx + Cx^2)\, e^{-x} \cos(\pi x) + (D + Ex + Fx^2)\, e^{-x} \sin(\pi x).$$

We plug in and then hopefully get equations that we can solve for A, B, C, D, E, and F. As you can see this can make for a very long and tedious calculation very quickly. C'est la vie!

There is one hiccup in all this. It could be that our guess actually solves the associated homogeneous equation. That is, suppose we have

$$y'' - 9y = e^{3x}.$$

We would love to guess $y = Ae^{3x}$, but if we plug this into the left-hand side of the equation we get

$$y'' - 9y = 9Ae^{3x} - 9Ae^{3x} = 0 \neq e^{3x}.$$

There is no way we can choose A to make the left-hand side be e^{3x}. The trick in this case is to multiply our guess by x to get rid of duplication with the complementary solution. That is first we compute y_c (solution to $L[y] = 0$)

$$y_c = C_1 e^{-3x} + C_2 e^{3x},$$

and we note that the e^{3x} term is a duplicate with our desired guess. We modify our guess to $y = Axe^{3x}$ so that there is no duplication anymore. Let us try: $y' = Ae^{3x} + 3Axe^{3x}$ and $y'' = 6Ae^{3x} + 9Axe^{3x}$, so

$$y'' - 9y = 6Ae^{3x} + 9Axe^{3x} - 9Axe^{3x} = 6Ae^{3x}.$$

Thus $6Ae^{3x}$ is supposed to equal e^{3x}. Hence, $6A = 1$ and so $A = 1/6$. We can now write the general solution as

$$y = y_c + y_p = C_1 e^{-3x} + C_2 e^{3x} + \frac{1}{6}\, xe^{3x}.$$

Notice that the term of the form xe^{3x} does not show up on the left-hand side after differentiating the equation, and the only term that survives is the e^{3x} term that showed up from the derivatives. This works out because e^{3x} solves the homogeneous problem. With that though, make sure to remember to include the xe^{3x} when you write out the general solution at the end of the problem, because it does appear there.

It is possible that multiplying by x does not get rid of all duplication. For example,

$$y'' - 6y' + 9y = e^{3x}.$$

The complementary solution is $y_c = C_1 e^{3x} + C_2 xe^{3x}$. Guessing $y = Axe^{3x}$ would not get us anywhere. In this case we want to guess $y_p = Ax^2 e^{3x}$. Basically, we want to multiply our guess by x until all duplication is gone. *But no more!* Multiplying too many times will not work (in that case, the derivatives won't actually get down to the plain e^{3x} term that you need in order to solve the problem).

Finally, what if the right-hand side has several terms, such as

$$L[y] = e^{2x} + \cos x.$$

In this case we find u that solves $L[u] = e^{2x}$ and v that solves $L[v] = \cos x$ (that is, do each term separately). Then note that if $y = u + v$, then $L[y] = e^{2x} + \cos x$. This is because L is linear; we have $L[y] = L[u + v] = L[u] + L[v] = e^{2x} + \cos x$.

To summarize all of this, we can make a table of the different guesses we should make given the form of the right hand side.

Right hand side	**Guess**
$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$	$Ax^n + Bx^{n-1} + \cdots + Nx + P$
e^{ax}	Ae^{ax}
$\cos ax$	$A\cos ax + B\sin ax$
$\sin ax$	$A\cos ax + B\sin ax$

- If there is a product of above terms, guess the product of the guesses. So, for a right hand side of xe^{ax}, the guess should be $(Ax+B)e^{ax}$, and for a right hand side of $x\cos ax$, the guess should be $(Ax + B)\cos ax + (Cx + D)\sin ax$.

- If any part solves the homogeneous problem, multiply that entire component by x until nothing does.

Example 2.5.2: Find the solution to the initial value problem

$$y'' - 3y' - 4y = 2e^{-x} + 4\sin(x) \qquad y(0) = -2,\ y'(0) = 1$$

Solution: To start this problem, we look for the solution to the homogeneous problem. The characteristic equation for the left hand side is $r^2 - 3r - 4$, which factors as $(r-4)(r+1)$. Therefore the general solution to the homogeneous problem (or the complementary solution) is

$$y_c(x) = C_1 e^{4x} + C_2 e^{-x}.$$

Next, we want to use undetermined coefficients to solve the non-homogeneous problem. Note that we have to wait until after this part to meet the initial conditions. Since our right-hand side is $2e^{-x} + 4\sin(x)$, we need to guess two components for the two different terms in this function. For the first term, we would want to guess Ae^{-x}, but this function solves the homogeneous problem. Therefore, we need to multiply by x to use Axe^{-x} as our guess. For the sine term, we need to guess both sine and cosine, so we add $B\sin(x) + C\cos(x)$ to our guess. Therefore, our total guess for the non-homogeneous solution is

$$y_p(x) = Axe^{-x} + B\sin(x) + C\cos(x).$$

We take two derivatives of this function and then plug it into the differential equation

$$\begin{aligned} y_p(x) &= Axe^{-x} + B\sin(x) + C\cos(x) \\ y_p'(x) &= Ae^{-x} - Axe^{-x} + B\cos(x) - C\sin(x) \\ y_p''(x) &= Axe^{-x} - 2Ae^{-x} - B\sin(x) - C\cos(x) \end{aligned}$$

so that

$$\begin{aligned} y_p'' - 3y_p' - 4y_p = &(Axe^{-x} - 2Ae^{-x} - B\sin(x) - C\cos(x)) \\ &- 3(Ae^{-x} - Axe^{-x} + B\cos(x) - C\sin(x)) \\ &- 4(Axe^{-x} + B\sin(x) + C\cos(x)) \end{aligned}$$

which can be simplified to

$$y_p'' - 3y_p' - 4y_p = -5Ae^{-x} + (3B + 3C)\sin(x) + (3C - 3B)\cos(x).$$

Since we want this to equal $2e^{-x} + 4\sin(x)$, this means that we need $-5A = 2$, so $A = -2/5$, as well as $3B + 3C = 4$ and $3C - 3B = 0$. The second of these implies that $B = C$, and the first equation then gives that $6B = 4$ or $B = C = 2/3$. Therefore, the general solution to this non-homogeneous problem is

$$y(x) = C_1 e^{4x} + C_2 e^{-x} - \frac{2}{5}xe^{-x} + \frac{2}{3}\sin(x) + \frac{2}{3}\cos(x).$$

Now we can look to meet the initial conditions. We want to differentiate this expression to get

$$y'(x) = 4C_1 e^{4x} - C_2 e^{-x} - \frac{2}{5}e^{-x} + \frac{2}{5}xe^{-x} + \frac{2}{3}\cos(x) - \frac{2}{3}\sin(x)$$

and then plug zero into both y and y' to get that

$$\begin{aligned} y(0) &= C_1 + C_2 + \frac{2}{3} = -2 \\ y'(0) &= 4C_1 - C_2 - \frac{2}{5} + \frac{2}{3} = 1 \end{aligned}$$

which gives rise to the system

$$C_1 + C_2 = -\frac{8}{3} \qquad 4C_1 - C_2 = \frac{11}{15}.$$

Adding the equations together gives $5C_1 = -\frac{29}{15}$ so that $C_1 = -\frac{29}{75}$ and then $C_2 = -\frac{171}{75}$. Therefore the solution to the initial value problem is

$$y(x) = -\frac{29}{75}e^{4x} - \frac{171}{75}e^{-x} - \frac{2}{5}xe^{-x} + \frac{2}{3}\sin(x) + \frac{2}{3}\cos(x).$$

***Exercise* 2.5.2:** *Verify that this $y(x)$ solves the initial value problem!*

2.5.3 Variation of parameters

The method of undetermined coefficients works for many basic problems that crop up. But it does not work all the time. It only works when the right-hand side of the equation $L[y] = f(x)$ has finitely many linearly independent derivatives, so that we can write a guess that consists of them all. Some equations are a bit tougher. Consider

$$y'' + y = \tan x.$$

Each new derivative of $\tan x$ looks completely different and cannot be written as a linear combination of the previous derivatives. If we start differentiating $\tan x$, we get:

$$\sec^2 x, \quad 2\sec^2 x \tan x, \quad 4\sec^2 x \tan^2 x + 2\sec^4 x,$$
$$8\sec^2 x \tan^3 x + 16\sec^4 x \tan x, \quad 16\sec^2 x \tan^4 x + 88\sec^4 x \tan^2 x + 16\sec^6 x, \quad \ldots$$

This equation calls for a different method. We present the method of *variation of parameters*, which handles any equation of the form $L[y] = f(x)$, provided we can solve certain integrals. For simplicity, we restrict ourselves to second order constant coefficient equations, but the method works for higher order equations just as well (the computations become more tedious). The method also works for equations with nonconstant coefficients, provided we can solve the associated homogeneous equation.

Perhaps it is best to explain this method by example. Let us try to solve the equation

$$L[y] = y'' + y = \tan x.$$

First we find the complementary solution (solution to $L[y_c] = 0$). We get $y_c = C_1y_1 + C_2y_2$, where $y_1 = \cos x$ and $y_2 = \sin x$. To find a particular solution to the nonhomogeneous equation we try

$$y_p = y = u_1y_1 + u_2y_2,$$

where u_1 and u_2 are *functions* and not constants. We are trying to satisfy $L[y] = \tan x$. That gives us one condition on the functions u_1 and u_2. Compute (note the product rule!)

$$y' = (u_1'y_1 + u_2'y_2) + (u_1y_1' + u_2y_2').$$

We can still impose one more condition at our discretion to simplify computations (we have two unknown functions, so we should be allowed two conditions). We require that $(u_1'y_1 + u_2'y_2) = 0$. This makes computing the second derivative easier.

$$\begin{aligned} y' &= u_1y_1' + u_2y_2', \\ y'' &= (u_1'y_1' + u_2'y_2') + (u_1y_1'' + u_2y_2''). \end{aligned}$$

Since y_1 and y_2 are solutions to $y'' + y = 0$, we find $y_1'' = -y_1$ and $y_2'' = -y_2$. (If the equation was a more general $y'' + p(x)y' + q(x)y = 0$, we would have $y_i'' = -p(x)y_i' - q(x)y_i$.) So

$$y'' = (u_1'y_1' + u_2'y_2') - (u_1y_1 + u_2y_2).$$

We have $(u_1y_1 + u_2y_2) = y$ and so

$$y'' = (u_1'y_1' + u_2'y_2') - y,$$

and hence

$$y'' + y = L[y] = u_1'y_1' + u_2'y_2'.$$

For y to satisfy $L[y] = f(x)$ we must have $f(x) = u_1'y_1' + u_2'y_2'$.

What we need to solve are the two equations (conditions) we imposed on u_1 and u_2:

$$\boxed{\begin{aligned} u_1'y_1 + u_2'y_2 &= 0, \\ u_1'y_1' + u_2'y_2' &= f(x). \end{aligned}}$$

We solve for u_1' and u_2' in terms of $f(x)$, y_1 and y_2. We always get these formulas for any $L[y] = f(x)$, where $L[y] = y'' + p(x)y' + q(x)y$. There is a general formula for the solution we could just plug into, but instead of memorizing that, it is better, and easier, to just repeat what we do below. In our case the two equations are

$$\begin{aligned} u_1'\cos(x) + u_2'\sin(x) &= 0, \\ -u_1'\sin(x) + u_2'\cos(x) &= \tan(x). \end{aligned}$$

Hence

$$\begin{aligned} u_1'\cos(x)\sin(x) + u_2'\sin^2(x) &= 0, \\ -u_1'\sin(x)\cos(x) + u_2'\cos^2(x) &= \tan(x)\cos(x) = \sin(x). \end{aligned}$$

And thus

$$\begin{aligned} u_2'\big(\sin^2(x) + \cos^2(x)\big) &= \sin(x), \\ u_2' &= \sin(x), \\ u_1' &= \frac{-\sin^2(x)}{\cos(x)} = -\tan(x)\sin(x). \end{aligned}$$

We integrate u_1' and u_2' to get u_1 and u_2.

$$u_1 = \int u_1' \, dx = \int -\tan(x)\sin(x) \, dx = \frac{1}{2}\ln\left|\frac{\sin(x)-1}{\sin(x)+1}\right| + \sin(x),$$
$$u_2 = \int u_2' \, dx = \int \sin(x) \, dx = -\cos(x).$$

So our particular solution is

$$y_p = u_1y_1 + u_2y_2 = \frac{1}{2}\cos(x)\ln\left|\frac{\sin(x)-1}{\sin(x)+1}\right| + \cos(x)\sin(x) - \cos(x)\sin(x) =$$
$$= \frac{1}{2}\cos(x)\ln\left|\frac{\sin(x)-1}{\sin(x)+1}\right|.$$

The general solution to $y'' + y = \tan x$ is, therefore,

$$y = C_1\cos(x) + C_2\sin(x) + \frac{1}{2}\cos(x)\ln\left|\frac{\sin(x)-1}{\sin(x)+1}\right|.$$

In more generality, we can take the system of equations

$$\boxed{\begin{aligned} u_1'y_1 + u_2'y_2 &= 0, \\ u_1'y_1' + u_2'y_2' &= f(x). \end{aligned}}$$

and solve out for u_1' and u_2' using elimination. If we do that, we get that

$$u_1' = -\frac{y_2(x)f(x)}{y_1(x)y_2'(x) - y_1'(x)y_2(x)} \qquad u_2' = \frac{y_1(x)f(x)}{y_1(x)y_2'(x) - y_1'(x)y_2(x)}.$$

We know that solving the equations this way will work out because we start with the assumption that y_1 and y_2 are linearly independent solutions, and the denominator of both of these fractions is exactly what we know is not zero from this assumption. Therefore, both of these functions can be written this way, we can integrate both of them, and set up our particular solution of the form $y_p(x) = u_1y_1 + u_2y_2$ to get

$$y_p(x) = -y_1(x)\int_{x_0}^{x} \frac{y_2(r)f(r)}{y_1(r)y_2'(r) - y_1'(r)y_2(r)} \, dr + y_2(x)\int_{x_0}^{x} \frac{y_1(r)f(r)}{y_1(r)y_2'(r) - y_1'(r)y_2(r)} \, dr \quad (2.8)$$

where x_0 is any conveniently chosen value (usually zero). Notice the use of r as a dummy variable here to separate the functions being integrated from the actual variable that shows up in the solution. This formula will always work for finding a particular solution to a non-homogeneous equation given that we know the solution to the homogeneous equation, but we may not be able to work out the integrals explicitly. This is the downside of this method, it may always work, but can be very tedious and may not result in nice, closed-form expressions like we might get from other methods.

Example 2.5.3: Find the general solution to the differential equation

$$y'' + 4y' + 3y = e^{3x} + 2$$

using both undetermined coefficients and variation of parameters.

Solution: For both methods of solving non-homogeneous equations, we need the solution to the homogeneous problem. For this equation, the characteristic polynomial is $r^2 + 4r + 3$, which factors as $(r + 1)(r + 3)$, so the general solution to the homogeneous problem is

$$y_c(x) = C_1 e^{-x} + C_2 e^{-3x}.$$

To use undetermined coefficients, we need to get the appropriate guess for the right-hand side, which in this case is $y_p(x) = Ae^{3x} + B$. Plugging this in to the differential equation gives

$$9Ae^{3x} + 4(3Ae^{3x}) + 3(Ae^{3x} + B) = e^{3x} + 2$$

which simplifies to

$$24Ae^{3x} + 3B = e^{3x} + 2$$

so that $A = 1/24$ and $B = 2/3$. Thus, the general solution to the non-homogeneous equation is

$$y(x) = C_1 e^{-x} + C_2 e^{-3x} + \frac{1}{24}e^{3x} + \frac{2}{3}.$$

In order to use variation of parameters, we let $y_1(x) = e^{-x}$ and $y_2(x) = e^{-3x}$ be the two linearly independent solutions that we found to the homogeneous problem. Our right-hand side function is $f(x) = e^{3x} + 2$ and we can compute the expression

$$y_1(x)y_2'(x) - y_1'(x)y_2(x) = e^{-x}(-3e^{-3x}) - (-e^{-x})e^{-3x} = -2e^{-4x}.$$

Therefore, we can use the formulas from the method of variation of parameters to compute that

$$u_1' = -\frac{y_2(x)f(x)}{y_1(x)y_2'(x) - y_1'(x)y_2(x)} = -\frac{e^{-3x}(e^{3x} + 2)}{-2e^{-4x}} = \frac{1}{2}e^{4x} + e^x$$
$$u_2' = \frac{y_1(x)f(x)}{y_1(x)y_2'(x) - y_1'(x)y_2(x)} = \frac{e^{-x}(e^{3x} + 2)}{-2e^{-4x}} = -\frac{1}{2}e^{6x} - e^{3x}.$$

Then we can compute

$$u_1 = \frac{1}{8}e^{4x} + e^x + C_1 \qquad u_2 = -\frac{1}{12}e^{6x} - \frac{1}{3}e^{3x} + C_2.$$

Then, we can write out the full general solution as $y(x) = u_1(x)y_1(x) + u_2(x)y_2(x)$ or

$$\begin{aligned} y(x) &= e^{-x}\left(\frac{1}{8}e^{4x} + e^x + C_1\right) + e^{-3x}\left(-\frac{1}{12}e^{6x} - \frac{1}{3}e^{3x} + C_2\right) \\ &= \frac{1}{8}e^{3x} + 1 + C_1e^{-x} - \frac{1}{12}e^{3x} - \frac{1}{3} + C_2e^{-3x} \end{aligned}$$

which, after combining the terms, is the same as the solution that we obtained via undetermined coefficients. ⌟

2.5.4 Exercises

Exercise 2.5.3: *Find a particular solution of* $y'' - y' - 6y = e^{2x}$.

Exercise 2.5.4: *Find a particular solution of* $y'' - 4y' + 4y = e^{2x}$.

Exercise 2.5.5:* Find a particular solution to* $y'' - y' + y = 2\sin(3x)$

Exercise 2.5.6: *Solve the initial value problem* $y'' + 9y = \cos(3x) + \sin(3x)$ *for* $y(0) = 2$, $y'(0) = 1$.

Exercise 2.5.7: *Set up the form of the particular solution but do not solve for the coefficients for* $y^{(4)} - 2y''' + y'' = e^x$.

Exercise 2.5.8: *Set up the form of the particular solution but do not solve for the coefficients for* $y^{(4)} - 2y''' + y'' = e^x + x + \sin x$.

Exercise 2.5.9:* Solve* $y'' + 2y' + y = x^2$, $y(0) = 1$, $y'(0) = 2$.

Exercise 2.5.10: *Use the method of undetermined coefficients to solve the DE* $y'' + 4y' = 2t + 30$.

Exercise 2.5.11:

a) Using variation of parameters find a particular solution of $y'' - 2y' + y = e^x$.

b) Find a particular solution using undetermined coefficients.

c) Are the two solutions you found the same? See also Exercise 2.5.27.

Exercise 2.5.12:*

a) Find a particular solution to $y'' + 2y = e^x + x^3$.

b) Find the general solution.

Exercise 2.5.13: *Find the general solution to* $y'' - 3y' - 4y = e^{2t} + 1$.

Exercise 2.5.14: *Find the general solution to* $y'' - 2y' - 5y = \sin(3t) + 2\cos(3t)$.

Exercise 2.5.15: *Find the general solution to* $y'' - 4y' - 21y = e^{-3t} + e^{4t}$.

Exercise 2.5.16: *Find the general solution to* $y'' - 2y' + y = e^t - t$.

Exercise 2.5.17: *Find the general solution to* $y'' + 4y = \sec(2t)$ *using variation of parameters.*

Exercise 2.5.18: *Find the solution of the initial value problem* $y'' - 2y' - 15y = e^{5t} + 3$, $y(0) = 2$, $y'(0) = -1$.

Exercise 2.5.19: *Find the solution of the initial value problem* $y'' + 4y' + 5y = \cos(3t) + t$, $y(0) = 0$, $y'(0) = 2$.

***Exercise* 2.5.20:** *The following differential equations are all related. Find the general solution to each of them and compare and contrast the different solutions and the methods used to approach them.*

a) $y'' - 2y' - 15y = e^t + 5e^{-4t}$

b) $y'' - 2y' - 15y = 2e^{2t} + 3e^{-t}$

c) $y'' - 2y' - 15y = 3\cos(2t)$

d) $y'' - 2y' - 15y = 2e^{5t} - \sin(t)$

***Exercise* 2.5.21:** *The following differential equations are all related. Find the general solution to each of them and compare and contrast the different solutions and the methods used to approach them.*

a) $y'' + 4y' + 3y = e^{2t} + 3e^{4t}$

b) $y'' - 2y' + 5y = e^{2t} + 3e^{4t}$

c) $y'' - 3y' - 10y = e^{2t} + 3e^{4t}$

d) $y'' - 8y' + 16y = e^{2t} + 3e^{4t}$

***Exercise* 2.5.22:** *Find a particular solution of* $y'' - 2y' + y = \sin(x^2)$. *It is OK to leave the answer as a definite integral.*

***Exercise* 2.5.23:** *Use variation of parameters to find a particular solution of* $y'' - y = \frac{1}{e^x + e^{-x}}$.

***Exercise* 2.5.24:** *Recall that a homogeneous* Euler equation *is one of the form* $t^2y'' + aty' + by = 0$ *and is solved by using the guess* $y(t) = t^r$ *and solving for the potential values of* r.

a) Solve $t^2y'' - 2ty' - 10y = 0$.

b) Let y_1 *and* y_2 *be a fundamental set for the above equation. Use the variation of parameters equations* $u_1 = -\displaystyle\int \frac{y_2\, g(t)}{y_1y_2' - y_2y_1'}\, dt$, $y_2 = \displaystyle\int \frac{y_1\, g(t)}{y_1y_2' - y_2y_1'}\, dt$ *to solve the non-homogeneous equation* $y'' - \dfrac{2}{t} - \dfrac{10}{t^2} = t^3$.
(Do not attempt method of undetermined coefficients instead; it won't work.)

***Exercise* 2.5.25:** *For an arbitrary constant* c *find the general solution to* $y'' - 2y = \sin(x + c)$.

***Exercise* 2.5.26:** *For an arbitrary constant* c *find a particular solution to* $y'' - y = e^{cx}$. *Hint: Make sure to handle every possible real* c.

***Exercise* 2.5.27:**

a) Using variation of parameters find a particular solution of $y'' - y = e^x$.

b) Find a particular solution using undetermined coefficients.

c) Are the two solutions you found the same? What is going on?

2.6 Forced oscillations and resonance

Attribution: [JL], §2.6.

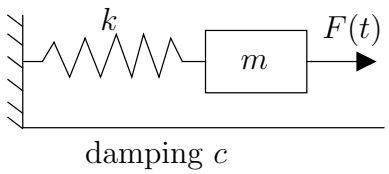
Learning Objectives

After this section, you will be able to:

- Write differential equations to model forced oscillators (like masses on springs),
- Identify when beats, pure resonance, and practical resonance can occur, and
- Use proper terminology around transient and steady periodic solutions when discussing these problems.

Let us return back to the example of a mass on a spring. We examine the case of forced oscillations, which we did not yet handle. That is, we consider the equation

$$mx'' + \gamma x' + kx = F(t),$$



for some nonzero $F(t)$. The setup is again: m is mass, γ is friction, k is the spring constant, and $F(t)$ is an external force acting on the mass.

We are interested in periodic forcing, such as noncentered rotating parts, or perhaps loud sounds, or other sources of periodic force.

2.6.1 Undamped forced motion and resonance

First let us consider undamped ($\gamma = 0$) motion. We have the equation

$$mx'' + kx = F_0 \cos(\omega t).$$

This equation has the complementary solution (solution to the associated homogeneous equation)

$$x_c = C_1 \cos(\omega_0 t) + C_2 \sin(\omega_0 t),$$

where $\omega_0 = \sqrt{k/m}$ is the *natural frequency* (angular). It is the frequency at which the system "wants to oscillate" without external interference.

Suppose that $\omega_0 \neq \omega$. We try the solution $x_p = A\cos(\omega t)$ and solve for A. We do not need a sine in our trial solution as after plugging in we only have cosines. If you include a sine, it is fine; you will find that its coefficient is zero (I could not find a second rhyme).

We solve using the method of undetermined coefficients. We find that

$$x_p = \frac{F_0}{m(\omega_0^2 - \omega^2)} \cos(\omega t).$$

We leave it as an exercise to do the algebra required.

The general solution is

$$\boxed{x = C_1 \cos(\omega_0 t) + C_2 \sin(\omega_0 t) + \frac{F_0}{m(\omega_0^2 - \omega^2)} \cos(\omega t).}$$

Written another way

$$x = C\cos(\omega_0 t - \gamma) + \frac{F_0}{m(\omega_0^2 - \omega^2)}\cos(\omega t).$$

The solution is a superposition of two cosine waves at different frequencies.

Example 2.6.1: Take

$$0.5x'' + 8x = 10\cos(\pi t), \qquad x(0) = 0, \qquad x'(0) = 0.$$

Solution: Let us compute. First we read off the parameters: $\omega = \pi$, $\omega_0 = \sqrt{8/0.5} = 4$, $F_0 = 10$, $m = 0.5$. The general solution is

$$x = C_1\cos(4t) + C_2\sin(4t) + \frac{20}{16 - \pi^2}\cos(\pi t).$$

Solve for C_1 and C_2 using the initial conditions: $C_1 = \frac{-20}{16-\pi^2}$ and $C_2 = 0$. Hence

$$x = \frac{20}{16 - \pi^2}\big(\cos(\pi t) - \cos(4t)\big).$$

Notice the "beating" behavior in Figure 2.8. First use the trigonometric identity

$$2\sin\left(\frac{A-B}{2}\right)\sin\left(\frac{A+B}{2}\right) = \cos B - \cos A$$

to get

$$x = \frac{20}{16 - \pi^2}\left(2\sin\left(\frac{4-\pi}{2}t\right)\sin\left(\frac{4+\pi}{2}t\right)\right).$$

The function x is a high frequency wave modulated by a low frequency wave.

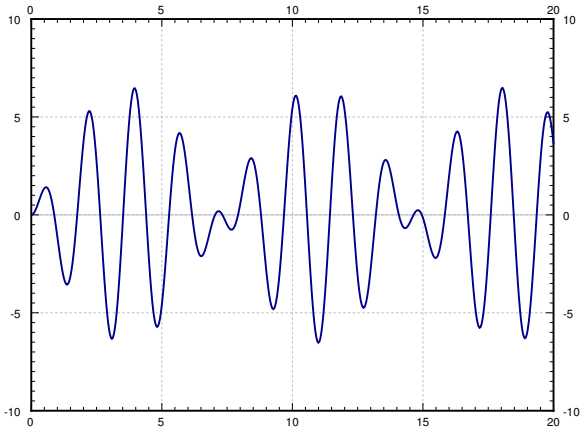


Figure 2.8: Graph of $\frac{20}{16-\pi^2}\big(\cos(\pi t) - \cos(4t)\big)$.

The beating behavior can be experienced even more readily by considering a higher frequency and viewing the resulting function as a sound wave. A sound wave of frequency 440 Hz produces and A4 sound, which is the A above middle C on a piano. This means that the function

$$x_p(t) = \sin(2\pi \cdot 440t)$$

will produce a sound wave equivalent to this A4 sound. In MATLAB, this can be done with the code

```
omega0 = 440*2*pi;
tVals = linspace(0, 5, 5*8192);

testSound = sin(omega0*tVals);
sound(testSound);
```

which will play this pitch for 5 seconds. Now, we want to see what happens if we take a mass-on-a-spring with this natural frequency and apply a forcing function with frequency close to this value. The following code assumes a forcing function of frequency 444 Hz. The multiple of ω_0 in front of the forcing function is only for scaling purposes; otherwise the resulting sound would be too quiet.

```
omega = 444*2*pi;

syms ys(t);
[V] = odeToVectorField(diff(ys, 2) + omega0^2*ys == omega0*cos(omega*t));
MS = matlabFunction(V, 'vars', {'t', 'Y'});
soln = ode45(MS, [0,10], [0,0]);

ySound = deval(soln, tVals);
ySound = ySound(1, :);
sound(ySound);
```

A graph of the solution `ySound` can be found in Figure 2.9. This exhibits the beating behavior before on a large scale. The sound played during this code also shows the beating or amplitude modulation that can happen in these sorts of solutions. In terms of tuning instruments, these beats are some of the main things musicians will listen for to know if their instrument is close to the right pitch, but just slightly off.

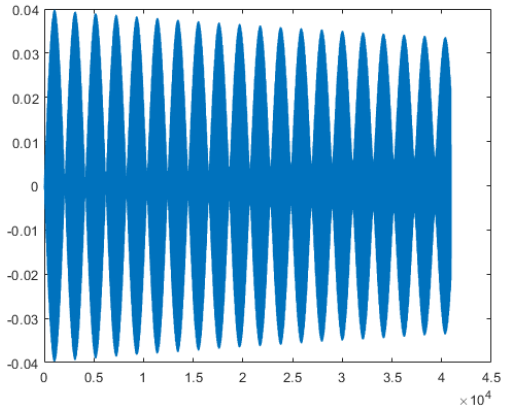


Figure 2.9: Plot of `ySound` *illustrating the beating behavior of interacting sound waves.*

Now suppose $\omega_0 = \omega$. We cannot try the solution $A\cos(\omega t)$ and then use the method of undetermined coefficients, since we notice that $\cos(\omega t)$ solves the associated homogeneous equation. Therefore, we try $x_p = At\cos(\omega t) + Bt\sin(\omega t)$. This time we need the sine term, since the second derivative of $t\cos(\omega t)$ contains sines. We write the equation

$$x'' + \omega^2 x = \frac{F_0}{m}\cos(\omega t).$$

Plugging x_p into the left-hand side we get

$$2B\omega\cos(\omega t) - 2A\omega\sin(\omega t) = \frac{F_0}{m}\cos(\omega t).$$

Hence $A = 0$ and $B = \frac{F_0}{2m\omega}$. Our particular solution is $\frac{F_0}{2m\omega}\, t\sin(\omega t)$ and our general solution is

$$x = C_1\cos(\omega t) + C_2\sin(\omega t) + \frac{F_0}{2m\omega}\, t\sin(\omega t).$$

The important term is the last one (the particular solution we found). This term grows without bound as $t \to \infty$. In fact it oscillates between $\frac{F_0 t}{2m\omega}$ and $\frac{-F_0 t}{2m\omega}$. The first two terms only oscillate between $\pm\sqrt{C_1^2 + C_2^2}$, which becomes smaller and smaller in proportion to the oscillations of the last term as t gets larger. In Figure 2.10 we see the graph with $C_1 = C_2 = 0$, $F_0 = 2$, $m = 1$, $\omega = \pi$.

By forcing the system in just the right frequency we produce very wild oscillations. This kind of behavior is called *resonance* or perhaps *pure resonance*. Sometimes resonance is desired. For example, remember when as a kid you could start swinging by just moving back and forth on the swing seat in the "correct frequency"? You were trying to achieve resonance. The force of each one of your moves was small, but after a while it produced large swings.

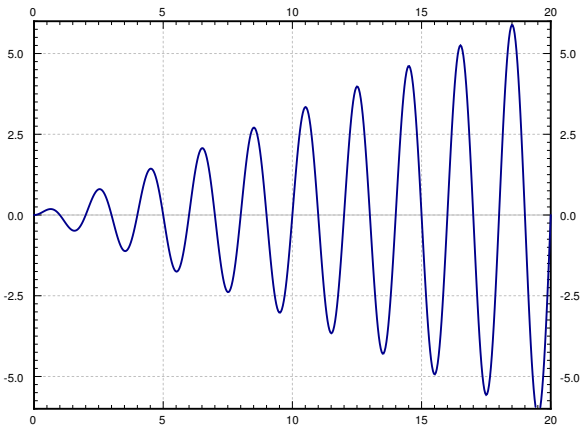


Figure 2.10: Graph of $\frac{1}{\pi}t\sin(\pi t)$.

On the other hand resonance can be destructive. In an earthquake some buildings collapse while others may be relatively undamaged. This is due to different buildings having different resonance frequencies. So figuring out the resonance frequency can be very important.

A common (but wrong) example of destructive force of resonance is the Tacoma Narrows bridge failure. It turns out there was a different phenomenon at play*.

2.6.2 Damped forced motion and practical resonance

In real life things are not as simple as they were above. There is, of course, some damping. Our equation becomes

$$mx'' + \gamma x' + kx = F_0\cos(\omega t), \tag{2.9}$$

for some $\gamma > 0$. We solved the homogeneous problem before. We let

$$p = \frac{\gamma}{2m}, \qquad \omega_0 = \sqrt{\frac{k}{m}}.$$

*K. Billah and R. Scanlan, *Resonance, Tacoma Narrows Bridge Failure, and Undergraduate Physics Textbooks*, American Journal of Physics, 59(2), 1991, 118–124, http://www.ketchum.org/billah/Billah-Scanlan.pdf

We replace equation (2.9) with

$$x'' + 2px' + \omega_0^2 x = \frac{F_0}{m}\cos(\omega t).$$

The roots of the characteristic equation of the associated homogeneous problem are $r_1, r_2 = -p \pm \sqrt{p^2 - \omega_0^2}$. The form of the general solution of the associated homogeneous equation depends on the sign of $p^2 - \omega_0^2$, or equivalently on the sign of $\gamma^2 - 4km$, as before:

$$x_c = \begin{cases} C_1 e^{r_1 t} + C_2 e^{r_2 t} & \text{if } \gamma^2 > 4km, \\ C_1 e^{-pt} + C_2 t e^{-pt} & \text{if } \gamma^2 = 4km, \\ e^{-pt}\big(C_1\cos(\omega_1 t) + C_2 \sin(\omega_1 t)\big) & \text{if } \gamma^2 < 4km, \end{cases}$$

where $\omega_1 = \sqrt{\omega_0^2 - p^2}$. In any case, we see that $x_c(t) \to 0$ as $t \to \infty$.

Let us find a particular solution. There can be no conflicts when trying to solve for the undetermined coefficients by trying $x_p = A\cos(\omega t) + B\sin(\omega t)$, because the solution to the homogeneous problem will always have exponential factors (since we have damping) and so there is no ω where this will exactly match the form of the homogeneous solution. Let us plug in and solve for A and B. We get (the tedious details are left to reader)

$$\big((\omega_0^2 - \omega^2)B - 2\omega pA\big)\sin(\omega t) + \big((\omega_0^2 - \omega^2)A + 2\omega pB\big)\cos(\omega t) = \frac{F_0}{m}\cos(\omega t).$$

We solve for A and B:

$$A = \frac{(\omega_0^2 - \omega^2)F_0}{m(2\omega p)^2 + m(\omega_0^2 - \omega^2)^2},$$
$$B = \frac{2\omega p F_0}{m(2\omega p)^2 + m(\omega_0^2 - \omega^2)^2}.$$

We also compute $C = \sqrt{A^2 + B^2}$ to be

$$C = \frac{F_0}{m\sqrt{(2\omega p)^2 + (\omega_0^2 - \omega^2)^2}}.$$

Thus our particular solution is

$$x_p = \frac{(\omega_0^2 - \omega^2)F_0}{m(2\omega p)^2 + m(\omega_0^2 - \omega^2)^2}\cos(\omega t) + \frac{2\omega p F_0}{m(2\omega p)^2 + m(\omega_0^2 - \omega^2)^2}\sin(\omega t).$$

Or in the alternative notation we have amplitude C and phase shift δ where (if $\omega \neq \omega_0$)

$$\tan\delta = \frac{B}{A} = \frac{2\omega p}{\omega_0^2 - \omega^2}.$$

Hence,

$$\boxed{x_p = \frac{F_0}{m\sqrt{(2\omega p)^2 + (\omega_0^2 - \omega^2)^2}}\cos(\omega t - \delta).}$$

If $\omega = \omega_0$, then $A = 0$, $B = C = \frac{F_0}{2m\omega p}$, and $\delta = \pi/2$.

For reasons we will explain in a moment, we call x_c the *transient solution* and denote it by x_{tr}. We call the x_p from above the *steady periodic solution* and denote it by x_{sp}. The general solution is

$$x = x_c + x_p = x_{tr} + x_{sp}.$$

The transient solution $x_c = x_{tr}$ goes to zero as $t \to \infty$, as all the terms involve an exponential with a negative exponent. So for large t, the effect of x_{tr} is negligible and we see essentially only x_{sp}. Hence the name *transient*. Notice that x_{sp} involves no arbitrary constants, and the initial conditions only affect x_{tr}. Thus, the effect of the initial conditions is negligible after some period of time. We might as well focus on the steady periodic solution and ignore the transient solution. See Figure 2.11 for a graph given several different initial conditions.

The speed at which x_{tr} goes to zero depends on p (and hence γ). The bigger p is (the bigger γ is), the "faster" x_{tr} becomes negligible. So the smaller the damping, the longer the "transient region." This is consistent with the observation that when $\gamma = 0$, the initial conditions affect the behavior for all time (i.e. an infinite "transient region").

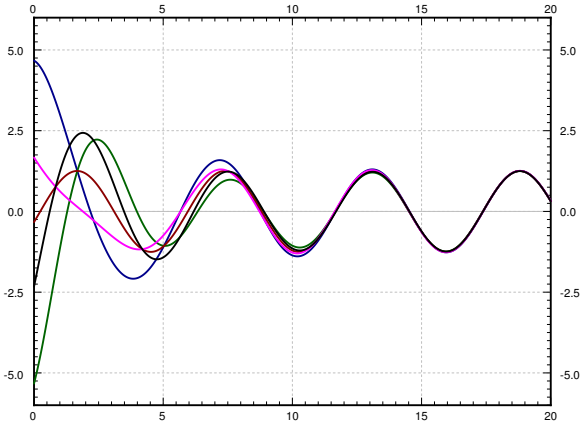


Figure 2.11: Solutions with different initial conditions for parameters $k = 1$, $m = 1$, $F_0 = 1$, $\gamma = 0.7$, and $\omega = 1.1$.

Let us describe what we mean by resonance when damping is present. Since there were no conflicts when solving with undetermined coefficient, there is no term that goes to infinity. We look instead at the maximum value of the amplitude of the steady periodic solution. Let C be the amplitude of x_{sp}. If we plot C as a function of ω (with all other parameters fixed), we can find its maximum. We call the ω that achieves this maximum the *practical resonance frequency*. We call the maximal amplitude $C(\omega)$ the *practical resonance amplitude*. Thus when damping is present we talk of *practical resonance* rather than pure resonance. A sample plot for three different values of γ is given in Figure 2.12 on the next page. As you can see the practical resonance amplitude grows as damping gets smaller, and practical resonance can disappear altogether when damping is large.

The main takeaways from Figure 2.12 on the facing page is that the amplitude can be larger than 1, which is the idea of resonance in this case. Based on Hooke's law, we know that a constant force of magnitude F_0 will stretch (or compress) a spring with constant k a length of F_0/k. If we take $F_0 = 1$ and $k = 1$, as is done in Figure 2.12 on the next page, then the resulting magnitude should be 1. However, if we don't use a constant force of magnitude F_0, but instead use an oscillatory force with frequency ω of the form $F(t) = F_0 \cos(\omega t)$, we get an amplitude of $C(\omega)$. This graph indicates how the forcing frequency changes the amplitude of the resulting oscillation. Since the amplitude "should" be 1 based on F_0/k, if

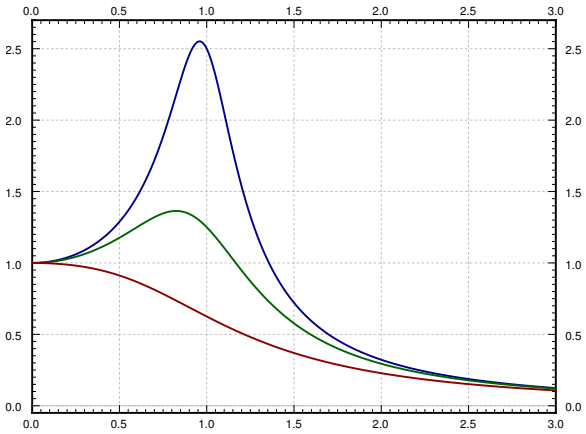


Figure 2.12: Graph of $C(\omega)$ showing practical resonance with parameters $k = 1$, $m = 1$, $F_0 = 1$. The top line is with $\gamma = 0.4$, the middle line with $\gamma = 0.8$, and the bottom line with $\gamma = 1.6$.

$C(\omega) > 1$, then the frequency chosen is causing an increase in the amplitude, which is the idea of practical resonance.

To find the maximum, or determine if there is a maximum, we need to find the derivative $C'(\omega)$. Computation shows

$$C'(\omega) = \frac{-2\omega(2p^2 + \omega^2 - \omega_0^2)F_0}{m\big((2\omega p)^2 + (\omega_0^2 - \omega^2)^2\big)^{3/2}}.$$

This is zero either when $\omega = 0$ or when $2p^2 + \omega^2 - \omega_0^2 = 0$. In other words, $C'(\omega) = 0$ when

$$\boxed{\omega = \sqrt{\omega_0^2 - 2p^2} \quad \text{or} \quad \omega = 0.}$$

If $\omega_0^2 - 2p^2$ is positive, then there is a positive value of ω, namely $\omega = \sqrt{\omega_0^2 - 2p^2}$ where the amplitude attains a maximum value. Since we know that the amplitude is $F_0/(m\omega_0^2)$ or F_0/k when $\omega = 0$, the maximum will be larger than this. As described above, this value, F_0/k is the expected amplitude, that is, the amplitude you would get with no oscillation, so that if the amplitude is larger than this for some value of ω, this means that the oscillation at frequency ω is resonating with the system to create a larger oscillation. This is the idea of *practical resonance*. It is practical because there is damping, so the situation is more physically relevant (to contrast with pure resonance), and still results in larger amplitudes of oscillation.

Our previous work indicates that a system will exhibit practical resonance for some values of ω whenever $\omega_0^2 - 2p^2$ is positive, and the frequency where the amplitude hits the maximum value is at $\sqrt{\omega_0^2 - 2p^2}$. This follows by the first derivative test for example as then $C'(\omega) > 0$ for small ω in this case. If on the other hand $\omega_0^2 - 2p^2$ is not positive, then $C(\omega)$ achieves its maximum at $\omega = 0$, and there is no practical resonance since we assume $\omega > 0$ in our system. In this case the amplitude gets larger as the forcing frequency gets smaller.

If practical resonance occurs, the peak frequency is smaller than ω_0. As the damping γ (and hence p) becomes smaller, the peak practical resonance frequency goes to ω_0. So when damping is very small, ω_0 is a good estimate of the peak practical resonance frequency. This behavior agrees with the observation that when $\gamma = 0$, then ω_0 is the resonance frequency.

Another interesting observation to make is that when $\omega \to \infty$, then $C \to 0$. This means that if the forcing frequency gets too high it does not manage to get the mass moving in the mass-spring system. This is quite reasonable intuitively. If we wiggle back and forth really fast while sitting on a swing, we will not get it moving at all, no matter how forceful. Fast vibrations just cancel each other out before the mass has any chance of responding by moving one way or the other.

The behavior is more complicated if the forcing function is not an exact cosine wave, but for example a square wave. A general periodic function will be the sum (superposition) of many cosine waves of different frequencies. The reader is encouraged to come back to this section once we have learned about the ideas of Fourier series.

2.6.3 Exercises

Exercise 2.6.1: *Write* $\cos(3x) - \cos(2x)$ *as a product of two sine functions.*

Exercise 2.6.2: *Write* $\cos(5x) - \cos(3x)$ *as a product of two sine functions.*

Exercise 2.6.3: *Write* $\cos(3x) - \cos(\pi x)$ *as a product of two sine functions.*

Exercise 2.6.4: *Derive a formula for* x_{sp} *if the equation is* $mx'' + \gamma x' + kx = F_0 \sin(\omega t)$. *Assume* $\gamma > 0$.

Exercise 2.6.5: *Derive a formula for* x_{sp} *if the equation is* $mx'' + \gamma x' + kx = F_0 \cos(\omega t) + F_1 \cos(3\omega t)$. *Assume* $\gamma > 0$.

Exercise 2.6.6:* *Derive a formula for* x_{sp} *for* $mx'' + \gamma x' + kx = F_0 \cos(\omega t) + A$, *where* A *is some constant. Assume* $\gamma > 0$.

Exercise 2.6.7: *Take* $mx'' + \gamma x' + kx = F_0 \cos(\omega t)$. *Fix* $m > 0$, $k > 0$, *and* $F_0 > 0$. *Consider the function* $C(\omega)$. *For what values of* γ *(solve in terms of* m, k, *and* F_0*) will there be no practical resonance (that is, for what values of* γ *is there no maximum of* $C(\omega)$ *for* $\omega > 0$*)?*

Exercise 2.6.8: *Take* $mx'' + \gamma x' + kx = F_0 \cos(\omega t)$. *Fix* $\gamma > 0$, $k > 0$, *and* $F_0 > 0$. *Consider the function* $C(\omega)$. *For what values of* m *(solve in terms of* γ, k, *and* F_0*) will there be no practical resonance (that is, for what values of* m *is there no maximum of* $C(\omega)$ *for* $\omega > 0$*)?*

Exercise 2.6.9:* *A mass of* $4\,\text{kg}$ *on a spring with* $k = 4\,\text{N/m}$ *and a damping constant* $c = 1\,\text{Ns/m}$. *Suppose that* $F_0 = 2\,\text{N}$. *Using forcing function* $F_0 \cos(\omega t)$, *find the* ω *that causes the maximum amount of practical resonance and find the amplitude.*

Exercise 2.6.10: *An infant is bouncing in a spring chair. The infant has a mass of* $8\,\text{kg}$, *and the chair functions as a spring with spring constant* $72\,\text{N/m}$. *The bouncing of the infant applies a force of the form* $3\cos(\omega t)$ *for some frequency* ω. *Assume that the infant starts at rest at the equilibrium position of the chair.*

a) *If there is no dampening coefficient, what frequency would the infant need to force at in order to generate pure resonance?*

b) *Assume that the chair is built with a dampener with coefficient* $5\,{}^{\text{Ns}}/_{\text{m}}$. *Set up an initial value problem for this situation if the child behaves in the same way.*

c) *Solve this initial value problem.*

d) *There are several options for chairs you can buy. There is the one with dampening coefficient* $5\,{}^{\text{Ns}}/_{\text{m}}$, *one with* $1\,{}^{\text{Ns}}/_{\text{m}}$, *and one with* $20\,{}^{\text{Ns}}/_{\text{m}}$. *Which of these would be most 'fun' for the infant? How do you know?*

***Exercise* 2.6.11:** *A water tower in an earthquake acts as a mass-spring system. Assume that the container on top is full and the water does not move around. The container then acts as the mass and the support acts as the spring, where the induced vibrations are horizontal. The container with water has a mass of* $m = 10,000\,\text{kg}$. *It takes a force of 1000 newtons to displace the container 1 meter. For simplicity assume no friction. When the earthquake hits the water tower is at rest (it is not moving). The earthquake induces an external force* $F(t) = mA\omega^2\cos(\omega t)$.

a) *What is the natural frequency of the water tower?*

b) *If* ω *is not the natural frequency, find a formula for the maximal amplitude of the resulting oscillations of the water container (the maximal deviation from the rest position). The motion will be a high frequency wave modulated by a low frequency wave, so simply find the constant in front of the sines.*

c) *Suppose* $A = 1$ *and an earthquake with frequency 0.5 cycles per second comes. What is the amplitude of the oscillations? Suppose that if the water tower moves more than 1.5 meter from the rest position, the tower collapses. Will the tower collapse?*

Exercise* 2.6.12: *Suppose there is no damping in a mass and spring system with* $m = 5$, $k = 20$, *and* $F_0 = 5$. *Suppose* ω *is chosen to be precisely the resonance frequency.*

a) *Find* ω.

b) *Find the amplitude of the oscillations at time* $t = 100$, *given the system is at rest at* $t = 0$.

***Exercise* 2.6.13:** *Assume that a 2 kg mass is attached to a spring that is acted on by a forcing function* $F(t) = 5\cos(2t)$. *Assume that there is no dampening on the spring.*

a) *What should the spring constant* k *be in order for this system to exhibit pure resonance?*

b) *If we wanted the system to exhibit practical resonance instead, what do or can we change about it to get this?*

c) *Assume that we set* k *to be the value determined in (a), and that the rest of the problem is situated so that the system exhibits practical resonance. What would we expect to see for the amplitude of the solution? This should be a generic comment, not a specific value.*

***Exercise* 2.6.14:** *Assume that we have a mass-on-a-spring system defined by the equation*

$$3y'' + 2y' + 18y = 4\cos(5t).$$

a) *Identify the mass, dampening coefficient, and spring constant for the system.*

b) *Use the entire equation to find the natural frequency, forcing frequency, and quasi-frequency of this oscillation.*

c) *Two of these frequencies will show up in the general solution to this problem. Which are they, and in which part (transient, steady-periodic) do they appear?*

d) *Find the general solution of this problem.*

***Exercise* 2.6.15:** *A circuit is built with an L Henry inductor, and R Ohm resistor, and a C Farad capacitor. All of the units are correct, but you do not know any of their values. To study this circuit, you apply an external voltage source of $F(t) = 4\cos\left(\frac{1}{2}t\right)$, and the circuit starts with no initial charge or current.*

a) *Write an initial value problem to model this situation.*

b) *Your friend (who knows more about this circuit than you do) takes a reading from this circuit after it is running and says "The amplitude of the charge oscillation is greater than 100 coulombs, which means this circuit is exhibiting practical resonance." There are* ***three*** *facts that you can learn about this circuit from the statement here that will tell you about the values of L, R, and C.*

   (i) *This statement seems to imply that the expected amplitude of the oscillation is 100 coulombs. What does this mean about the value of C?*

   (ii) *Your friend says that this circuit is in practical resonance. What does this tell you about the value of R in this case?*

   (iii) *Finally, being in practical resonance says something about how the forcing frequency compares to the natural frequency of this system. What is that, and how does it relate to the value of L?*

c) *What is the frequency of the steady-periodic oscillation that your friend mentioned above?*

2.7 Higher order linear ODEs

Attribution: [JL], §2.3.

Learning Objectives

After this section, you will be able to:

- Find the general solution to a linear, constant coefficient, homogeneous differential equation of higher order and
- Solve non-homogeneous higher order equations using the method of undetermined coefficients.

In this section, we will briefly study higher order equations. Equations appearing in applications tend to be second order. Higher order equations do appear from time to time, but generally the world around us is "second order."

The basic results about linear ODEs of higher order are essentially the same as for second order equations, with 2 replaced by n. The important concept of linear independence is somewhat more complicated when more than two functions are involved. For higher order constant coefficient ODEs, the methods developed are also somewhat harder to apply, but we will not dwell on these complications. It is also possible to use the methods for systems of linear equations from chapter 4 to solve higher order constant coefficient equations.

Let us start with a general homogeneous linear equation

$$y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = 0. \tag{2.10}$$

Theorem 2.7.1 (Superposition)

Suppose y_1, y_2, …, y_n are solutions of the homogeneous equation (2.10). Then

$$y(x) = C_1y_1(x) + C_2y_2(x) + \cdots + C_ny_n(x)$$

also solves (2.10) for arbitrary constants $C_1, C_2, \ldots, C_n$.

In other words, a *linear combination* of solutions to (2.10) is also a solution to (2.10). We also have the existence and uniqueness theorem for nonhomogeneous linear equations.

Theorem 2.7.2 (Existence and uniqueness)

Suppose p_0 through p_{n-1}, and f are continuous functions on some interval I, a is a number in I, and $b_0, b_1, \ldots, b_{n-1}$ are constants. The equation

$$y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = f(x)$$

has exactly one solution $y(x)$ defined on the same interval I satisfying the initial conditions

$$y(a) = b_0, \quad y'(a) = b_1, \quad \ldots, \quad y^{(n-1)}(a) = b_{n-1}.$$

2.7.1 Linear independence

When we had two functions y_1 and y_2 we said they were linearly independent if one was not the multiple of the other. Same idea holds for n functions. In this case, it is easier to state as follows. The functions y_1, y_2, ..., y_n are *linearly independent* if the equation

$$c_1y_1 + c_2y_2 + \cdots + c_ny_n = 0$$

has only the trivial solution $c_1 = c_2 = \cdots = c_n = 0$, where the equation must hold for all x. If we can solve equation with some constants where for example $c_1 \neq 0$, then we can solve for y_1 as a linear combination of the others. If the functions are not linearly independent, they are *linearly dependent*.

Example 2.7.1: Show that e^x, e^{2x}, e^{3x} are linearly independent.

Solution: Let us give several ways to show this fact. Many textbooks (including [EP] and [F]) introduce Wronskians for higher order equations, but it is harder to analyze them without tools from linear algebra (see Chapter 3). Once there are more than two functions involved, there is not a nice, simple formula for the Wronskian (like $y_1'y_2 - y_2'y_1$ for two functions) and linear algebra is required to analyze what is happening here. Instead, we will take a slightly different and more improvized approach to see why these functions are linearly independent.

Let us write down

$$c_1e^x + c_2e^{2x} + c_3e^{3x} = 0.$$

We use rules of exponentials and write $z = e^x$. Hence $z^2 = e^{2x}$ and $z^3 = e^{3x}$. Then we have

$$c_1z + c_2z^2 + c_3z^3 = 0.$$

The left-hand side is a third degree polynomial in z. It is either identically zero, or it has at most 3 zeros. Therefore, it is identically zero, $c_1 = c_2 = c_3 = 0$, and the functions are linearly independent.

Let us try another way. As before we write

$$c_1e^x + c_2e^{2x} + c_3e^{3x} = 0.$$

This equation has to hold for all x. We divide through by e^{3x} to get

$$c_1e^{-2x} + c_2e^{-x} + c_3 = 0.$$

As the equation is true for all x, let $x \to \infty$. After taking the limit we see that $c_3 = 0$. Hence our equation becomes

$$c_1e^x + c_2e^{2x} = 0.$$

Rinse, repeat!

How about yet another way. We again write

$$c_1e^x + c_2e^{2x} + c_3e^{3x} = 0.$$

We can evaluate the equation and its derivatives at different values of x to obtain equations for c_1, c_2, and c_3. Let us first divide by e^x for simplicity.

$$c_1 + c_2e^x + c_3e^{2x} = 0.$$

We set $x = 0$ to get the equation $c_1 + c_2 + c_3 = 0$. Now differentiate both sides

$$c_2 e^x + 2c_3 e^{2x} = 0.$$

We set $x = 0$ to get $c_2 + 2c_3 = 0$. We divide by e^x again and differentiate to get $2c_3 e^x = 0$. It is clear that c_3 is zero. Then c_2 must be zero as $c_2 = -2c_3$, and c_1 must be zero because $c_1 + c_2 + c_3 = 0$.

There is no one best way to do it. All of these methods are perfectly valid. The important thing is to understand why the functions are linearly independent. □

Exercise* 2.7.1** (not necessary on first reading):*** *Here is the linear algebra method for after reading through that chapter. Let $y_1 = e^x$, $y_2 = e^{2x}$ and $y_3 = e^{3x}$. Verify that*

$$\begin{bmatrix} y_1(0) \\ y_1'(0) \\ y_1''(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \begin{bmatrix} y_2(0) \\ y_2'(0) \\ y_2''(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} \quad \begin{bmatrix} y_3(0) \\ y_3'(0) \\ y_3''(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 9 \end{bmatrix}$$

and use that to determine that these functions are linearly independent by showing that

$$\det \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 4 & 9 \end{bmatrix} = 2 \neq 0$$

so that this matrix is invertible.

Example 2.7.2: On the other hand, the functions e^x, e^{-x}, and $\cosh x$ are linearly dependent. Simply apply definition of the hyperbolic cosine:

$$\cosh x = \frac{e^x + e^{-x}}{2} \qquad \text{or} \qquad 2\cosh x - e^x - e^{-x} = 0.$$

This second form here is a linear combination (coefficients 2, -1, and -1) of the three functions that adds to zero.

2.7.2 Constant coefficient higher order ODEs

When we have a higher order constant coefficient homogeneous linear equation, the song and dance is exactly the same as it was for second order. We just need to find more solutions. If the equation is n^{th} order, we need to find n linearly independent solutions. It is best seen by example.

Example 2.7.3: Find the general solution to

$$y''' - 3y'' - y' + 3y = 0. \tag{2.11}$$

Solution: Try: $y = e^{rx}$. We plug in and get

$$\underbrace{r^3 e^{rx}}_{y'''} - 3\underbrace{r^2 e^{rx}}_{y''} - \underbrace{r e^{rx}}_{y'} + 3\underbrace{e^{rx}}_{y} = 0.$$

We divide through by e^{rx}. Then

$$r^3 - 3r^2 - r + 3 = 0.$$

The trick now is to find the roots. There is a formula for the roots of degree 3 and 4 polynomials but it is very complicated. There is no formula for higher degree polynomials. That does not mean that the roots do not exist. There are always n roots for an n^{th} degree polynomial. They may be repeated and they may be complex. Computers are pretty good at finding roots approximately for reasonable size polynomials.

A good place to start is to plot the polynomial and check where it is zero. We can also simply try plugging in. We just start plugging in numbers $r = -2, -1, 0, 1, 2, \ldots$ and see if we get a hit (we can also try complex numbers). Even if we do not get a hit, we may get an indication of where the root is. For example, we plug $r = -2$ into our polynomial and get -15; we plug in $r = 0$ and get 3. That means there is a root between $r = -2$ and $r = 0$, because the sign changed. If we find one root, say r_1, then we know $(r - r_1)$ is a factor of our polynomial. Polynomial long division can then be used.

Another technique for guessing roots of polynomials is the Rational Roots Theorem, which says that any rational root of the polynomial must be of the form p/q where p divides the constant term of the polynomial and q divides the leading term, provided neither of them are zero. For more information on this see § B.1. In this case, we would know that p must divide 3, and q must divide 1. Therefore, the only possible options here are ± 1 and ± 3. These would be good places to start to look for rational roots.

A good strategy is to begin with $r = 0$, 1, or -1. These are easy to compute. Our polynomial has two such roots, $r_1 = -1$ and $r_2 = 1$. There should be 3 roots and the last root is reasonably easy to find. The constant term in a monic* polynomial such as this is the multiple of the negations of all the roots because $r^3 - 3r^2 - r + 3 = (r - r_1)(r - r_2)(r - r_3)$. So

$$3 = (-r_1)(-r_2)(-r_3) = (1)(-1)(-r_3) = r_3.$$

You should check that $r_3 = 3$ really is a root. Hence e^{-x}, e^x and e^{3x} are solutions to (2.11). They are linearly independent as can easily be checked, and there are 3 of them, which happens to be exactly the number we need. So the general solution is

$$y = C_1 e^{-x} + C_2 e^x + C_3 e^{3x}.$$

Another possible way to work out this general solution is by factoring the original polynomial. Since we want to solve

$$r^3 - 3r^2 - r + 3 = 0,$$

we can rewrite the polynomial as

$$r^2(r - 3) - 1(r - 3) = 0$$

which factors as

$$(r^2 - 1)(r - 3) = 0.$$

*The word monic means that the coefficient of the top degree r^d, in our case r^3, is 1.

Finally, using difference of two squares on the first factor gives

$$(r-1)(r+1)(r-3) = 0.$$

This gives roots of 1, -1, and 3, and so the same general solution as above.

Suppose we were given some initial conditions $y(0) = 1$, $y'(0) = 2$, and $y''(0) = 3$. Then

$$\begin{aligned} 1 = y(0) &= C_1 + C_2 + C_3, \\ 2 = y'(0) &= -C_1 + C_2 + 3C_3, \\ 3 = y''(0) &= C_1 + C_2 + 9C_3. \end{aligned}$$

It is possible to find the solution by high school algebra, but it would be a pain. The sensible way to solve a system of equations such as this is to use matrix algebra, see § 4.2 or Chapter 3. For now we note that the solution is $C_1 = -1/4$, $C_2 = 1$, and $C_3 = 1/4$. The specific solution to the ODE is

$$y = \frac{-1}{4}\, e^{-x} + e^x + \frac{1}{4}\, e^{3x}.$$

Next, suppose that we have real roots, but they are repeated. Let us say we have a root r repeated k times. In the spirit of the second order solution, and for the same reasons, we have the solutions

$$e^{rx}, \quad xe^{rx}, \quad x^2e^{rx}, \quad \ldots, \quad x^{k-1}e^{rx}.$$

We take a linear combination of these solutions to find the general solution.

Example 2.7.4: Solve

$$y^{(4)} - 3y''' + 3y'' - y' = 0.$$

Solution: We note that the characteristic equation is

$$r^4 - 3r^3 + 3r^2 - r = 0.$$

By inspection we note that $r^4 - 3r^3 + 3r^2 - r = r(r-1)^3$. Hence the roots given with multiplicity are $r = 0, 1, 1, 1$. Thus the general solution is

$$y = \underbrace{(C_1 + C_2x + C_3x^2)\, e^x}_{\text{terms coming from } r=1} + \underbrace{C_4}_{\text{from } r=0}.$$

Example 2.7.5: Find the general solution of

$$y''' + 2y'' - 5y' - 6y = 0$$

Solution: The characteristic equation for this example is

$$r^3 + 2r^2 - 5r - 6 = 0.$$

There is no convenient factoring by grouping or other quick formula to get to the roots here. The best hope we have is to try to guess the roots and see if we come up with anything. Once

we get one root, we'll be able to factor a term out and get down to a quadratic equation, where the quadratic formula will give us the other two roots.

The properties of polynomials tell us that all rational roots of this polynomial must be factors of $\frac{-6}{1}$ or -6. Thus, the options are ± 1, ± 2, and ± 3. At this point, the best bet is to start guessing and see if we can find one. Let's start with 1. Plugging this into the polynomial gives

$$1^3 + 2(1)^2 - 5(1) - 6 = -8 \neq 0.$$

Trying -1 next, we get

$$(-1)^3 + 2(-1)^2 - 5(-1) - 6 = -1 + 2 + 5 - 6 = 0.$$

Therefore, -1 is as root, and so $(r+1)$ is a factor of this polynomial.

We can then use synthetic (or long) division to see that

$$r^3 + 2r^2 - 5r - 6 = (r+1)(r^2 + r - 6).$$

For the quadratic, we can either use the quadratic formula, or just recognize that this factors as $(r-2)(r+3)$ to get that the characteric equation factors as

$$(r+1)(r-2)(r+3) = 0.$$

Therefore, the roots are -1, 2 and -3, so that the general solution to the differential equation is

$$y(x) = C_1 e^{-x} + C_2 e^{2x} + C_3 e^{-3x}.$$

For more information on synthetic division and finding roots of polynomials, see Appendix B.1.

The case of complex roots is similar to second order equations. Complex roots always come in pairs $r = \alpha \pm i\beta$. Suppose we have two such complex roots, each repeated k times. The corresponding solution is

$$\left(C_0 + C_1 x + \cdots + C_{k-1} x^{k-1}\right) e^{\alpha x} \cos(\beta x) + \left(D_0 + D_1 x + \cdots + D_{k-1} x^{k-1}\right) e^{\alpha x} \sin(\beta x).$$

where $C_0, \ldots, C_{k-1}$, $D_0, \ldots, D_{k-1}$ are arbitrary constants.

Example 2.7.6: Solve

$$y^{(4)} - 4y''' + 8y'' - 8y' + 4y = 0.$$

Solution: The characteristic equation is

$$\begin{aligned} r^4 - 4r^3 + 8r^2 - 8r + 4 &= 0, \\ \left(r^2 - 2r + 2\right)^2 &= 0, \\ \left(\left(r-1\right)^2 + 1\right)^2 &= 0. \end{aligned}$$

Hence the roots are $1 \pm i$, both with multiplicity 2. Hence the general solution to the ODE is

$$y = \left(C_1 + C_2 x\right) e^x \cos x + \left(C_3 + C_4 x\right) e^x \sin x.$$

The way we solved the characteristic equation above is really by guessing or by inspection. It is not so easy in general. We could also have asked a computer or an advanced calculator for the roots.

2.7.3 Non-Homogeneous Equations

Just like for second order equation, we can solve higher order non-homogeneous equations. The theory is the same; if we can find any single solution to the non-homogeneous problem, then the general solution of the non-homogeneous problem is this single solution plus the general solution to the corresponding homogeneous problem. The trick comes down to finding this single solution, and undetermined coefficients is the main method here.

In using undetermined coefficients, the guesses we want to make are the same as for second order equations. The only way it really gets more complicated is that now it is possible for any exponential or trigonometric function to be a solution to the homogeneous problem, and so more things will need to be multiplied by x in order to get the appropriate guess for the non-homogeneous solution.

Example 2.7.7: Find the general solution to

$$y''' + 2y'' - 5y' - 6y = 3e^{2x} + e^{4x}.$$

Solution: We found the general solution of the homogeneous problem in Example 2.7.5, which is

$$y(x) = C_1e^{-x} + C_2e^{2x} + C_3e^{-3x}.$$

Now, to solve the non-homogeneous problem, we use the method of undetermined coefficients. Since the non-homogeneous part of the equation has terms of the form e^{2x} and e^{4x}, we would want to guess

$$y_p(x) = Ae^{2x} + Be^{4x}.$$

However, e^{2x} solves the homogeneous problem, so we need to multiply it by x, making our actual guess become

$$y_p(x) = Axe^{2x} + Be^{4x}.$$

In order to plug this in, we need to take three derivatives of this guess, which are

$$\begin{aligned} y_p(x) &= Axe^{2x} + Be^{4x} \\ y_p'(x) &= Ae^{2x} + 2Axe^{2x} + 4Be^{4x} \\ y_p''(x) &= 4Ae^{2x} + 4Axe^{2x} + 16Be^{4x} \\ y_p'''(x) &= 12Ae^{2x} + 8Axe^{2x} + 64Be^{2x} \end{aligned}.$$

By putting this into the non-homogeneous equation we want to solve, we get

$$\begin{aligned} &(12Ae^{2x} + 8Axe^{2x} + 64Be^{2x}) + 2(4Ae^{2x} + 4Axe^{2x} + 16Be^{4x}) \\ &\quad - 5(Ae^{2x} + 2Axe^{2x} + 4Be^{4x}) - 6(Axe^{2x} + Be^{4x}) = 3e^{2x} + e^{4x}. \end{aligned}$$

Simplifying the left hand side of this expression gives

$$15Ae^{2x} + 70Be^{4x} = 3e^{2x} + e^{4x}.$$

To satisfy this equation, we want to set $A = \frac{1}{5}$ and $B = \frac{1}{70}$. Therefore, the general solution to the non-homogeneous problem is

$$y(x) = C_1e^{-x} + C_2e^{2x} + C_3e^{-3x} + \frac{1}{5}xe^{2x} + \frac{1}{70}e^{4x}.$$

Example 2.7.8: Determine the form of the guess using undetermined coefficients for finding a particular solution of the non-homogeneous problem

$$y^{(9)} + y^{(8)} - 2y^{(5)} - 2y^{(4)} + y' + y = e^x + 3e^{-x} + \sin(x) + 2x.$$

Solution: To determine the guess, we need to first find the solution to the homogeneous equations. The characteristic equation of the homogeneous equation is

$$r^9 + r^8 - 2r^5 - 2r^4 + r + 1 = 0.$$

We could use the root guessing method for this example, and all rational roots must be ± 1. However, that method is not great for polynomials that are of degree higher than around 3 or 4. So, we'll want to use some other technique to find all of the root.

If we start by grouping pairs of terms, we can rewrite this polynomial as

$$r^8(r+1) - 2r^4(r+1) + 1(r+1) = 0$$

so that it can be rewritten as

$$(r+1)(r^8 - 2r^4 + 1) = 0.$$

The second factor looks a lot like

$$(s-1)^2 = s^2 - 2s + 1$$

if we take $s = r^4$. Since

$$(r^4 - 1) = (r^2 + 1)(r^2 - 1) = (r^2 + 1)(r + 1)(r - 1)$$

using difference of squares twice. Thus, the entire characteristic equation can be written as

$$(r+1)(r^4-1)^2 = (r+1)[(r^2+1)(r+1)(r-1)]^2 = (r+1)^3(r-1)^2(r^2+1)^2.$$

Therefore, we have a triple root at -1, a double root at 1, and two copies of $(r^2 + 1)$, which has a root of i, corresponding to solutions $\sin(x)$ and $\cos(x)$. Putting all of this together, the general solution to the homogeneous equation is

$$y_c(x) = (C_1 + C_2x + C_3x^2)e^{-x} + (C_4 + C_5x)e^x + (C_6 + C_7x)\sin(x) + (C_8 + C_9x)\cos(x).$$

This has 9 unknown constants in it, which is expected from the ninth order equation.

Now, we need to figure out the appropriate guess for the non-homogeneous solution. Since the non-homogeneous part of the equation is $e^x + 3e^{-x} + \sin x + 2x$, the base guess would be of the form

$$Ae^x + Be^{-x} + C\sin x + D\cos x + Ex + F$$

because we always need to include both $\sin(x)$ and $\cos(x)$ whenever either of them appear. However, we need to factor in what terms show up in the homogeneous solution. For instance, the e^x term has a term with 1 and x in the homogeneous solution, we need to include the

next one up in our guess for the solution to the non-homogeneous problem. Taking this into account for all terms gives the desired guess as

$$y_p(x) = Ax^2e^x + Bx^3e^{-x} + Cx^2\sin(x) + Dx^2\cos(x) + Ex + F.$$

There is also an extension of variation of parameters to higher order equations. However, the fact that there are more terms in the solution means that the form of the expression is much more complicated that for second order, and is not worth looking into or trying to remember. The easier way to handle these situations using variation of parameters is by converting the higher order equation into a first order system and applying the methods there, which will be covered in § 4.1 and § 4.8 respectively.

2.7.4 Exercises

***Exercise* 2.7.2:** *Find the general solution for* $y''' - y'' + y' - y = 0$.

Exercise* 2.7.3: *Find the general solution of* $y^{(5)} - y^{(4)} = 0$.

***Exercise* 2.7.4:** *Find the general solution for* $y^{(4)} - 5y''' + 6y'' = 0$.

***Exercise* 2.7.5:** *Find the general solution for* $y''' + 2y'' + 2y' = 0$.

***Exercise* 2.7.6:** *Suppose the characteristic equation for an ODE is* $(r-1)^2(r-2)^2 = 0$.

a) *Find such a differential equation.*

b) *Find its general solution.*

***Exercise* 2.7.7:** *Suppose that a fourth order equation has a solution* $y = 2e^{4x}x\cos x$.

a) *Find such an equation.*

b) *Find the initial conditions that the given solution satisfies.*

Exercise* 2.7.8: *Suppose that the characteristic equation of a third order differential equation has roots* $\pm 2i$ *and* 3.

a) *What is the characteristic equation?*

b) *Find the corresponding differential equation.*

c) *Find the general solution.*

***Exercise* 2.7.9:** *Find the general solution for the equation of Exercise 2.7.7.*

Exercise* 2.7.10: *Find the general solution of*

$$y^{(4)} - y''' - 5y'' - 23y' - 20y = 0.$$

Exercise 2.7.11: *Find the general solution of*

$$y''' - 6y'' + 13y' - 10y = 4e^x + 5e^{3x} - 20.$$

Exercise 2.7.12: *Find the general solution of*

$$y''' - 3y' + 2y = 2e^x - e^{3x}.$$

Exercise 2.7.13: *Find the general solution of*

$$y''' + 2y'' + y' + 2y = 3\cos(x) + x.$$

Exercise 2.7.14: *Find the general solution of*

$$y^{(4)} + 2y'' + y = 4x\cos(x) - e^{3x} + 1$$

Hint: Remember, the guess needs to make sure that no terms in it solve the homogeneous equation.

Exercise 2.7.15: *Show that $y = \cos(2t)$ is a solution to $y^{(4)} + 2y''' + 9y'' + 8y' + 20y = 0$. This tells us something about the factorization of the characteristic polynomial of this DE. Factor the characteristic polynomial completely, and solve the DE.*

Exercise 2.7.16: *Consider*

$$y''' - y'' - 8y' + 12y = 0. \tag{2.12}$$

a) *Show that $y = e^{2t}$ is a solution of (2.12).*

b) *Find the general solution to (2.12).*

c) *Solve $y''' - y'' - 8y' + 12y = e^{2t}$.*

Exercise 2.7.17: *Let $f(x) = e^x - \cos x$, $g(x) = e^x + \cos x$, and $h(x) = \cos x$. Are $f(x)$, $g(x)$, and $h(x)$ linearly independent? If so, show it, if not, find a linear combination that works.*

Exercise 2.7.18: *Let $f(x) = 0$, $g(x) = \cos x$, and $h(x) = \sin x$. Are $f(x)$, $g(x)$, and $h(x)$ linearly independent? If so, show it, if not, find a linear combination that works.*

Exercise 2.7.19:* *Are e^x, e^{x+1}, e^{2x}, $\sin(x)$ linearly independent? If so, show it, if not find a linear combination that works.*

Exercise 2.7.20: *Are x, x^2, and x^4 linearly independent? If so, show it, if not, find a linear combination that works.*

Exercise 2.7.21: *Are e^x, xe^x, and x^2e^x linearly independent? If so, show it, if not, find a linear combination that works.*

Exercise 2.7.22:* *Are $\sin(x)$, x, $x\sin(x)$ linearly independent? If so, show it, if not find a linear combination that works.*

Exercise 2.7.23: *Show that $\{e^t, te^t, e^{-t}, te^{-t}\}$ is a linearly independent set.*

Exercise* 2.7.24: *Solve* $1001y''' + 3.2y'' + \pi y' - \sqrt{4}y = 0$, $y(0) = 0$, $y'(0) = 0$, $y''(0) = 0$.

***Exercise* 2.7.25:** *Could* $y = t^2 \cos t$ *be a solution of a homogeneous DE with constant real coefficients? If so, give the minimum possible order of such a DE, and state which functions must also be solutions. If not, explain why this is impossible.*

***Exercise* 2.7.26:** *Find a linear DE with constant coefficients whose general solution is*

$$y = c_1 e^{2t} + c_2 e^{-t} \cos(4t) + c_3 e^{-t} \sin(2t),$$

or explain why there is no such thing.

***Exercise* 2.7.27:** *Find an equation such that* $y = xe^{-2x} \sin(3x)$ *is a solution.*

Exercise* 2.7.28: *Find an equation of minimal order such that* $y = \cos(x)$, $y = \sin(x)$, $y = e^x$ *are solutions.*

***Exercise* 2.7.29:** *Find an equation of minimal order such that* $y = \cos(x)$, $y = \sin(2x)$, $y = e^{3x}$ *are solutions.*

***Exercise* 2.7.30:** *Find a homogeneous DE with general solution*

$$y = c_1 e^t + c_2 e^{-t} + c_3 \cos t + c_4 \sin t + c_5 t e^t + c_6 t e^{-t} + c_7 t \cos t + c_8 t \sin t.$$

Chapter 3

Linear algebra

3.1 Vectors, mappings, and matrices

Attribution: [JL], §A.1.

Learning Objectives

After this section, you will be able to:

- Express n-tuples of numbers as vectors,
- Perform operations on vectors, and
- Understand how linear maps on vectors give rise to matrices.

In real life, there is most often more than one variable. We wish to organize dealing with multiple variables in a consistent manner, and in particular organize dealing with linear equations and linear mappings, as those both rather useful and rather easy to handle. Mathematicians joke that "to an engineer every problem is linear, and everything is a matrix." And well, they (the engineers) are not wrong. Quite often, solving an engineering problem is figuring out the right finite-dimensional linear problem to solve, which is then solved with some matrix manipulation. Most importantly, linear problems are the ones that we know how to solve, and we have many tools to solve them. For engineers, mathematicians, physicists, and anybody in a technical field it is absolutely vital to learn linear algebra.

As motivation, suppose we wish to solve

$$\begin{aligned} x - y &= 2, \\ 2x + y &= 4, \end{aligned}$$

for x and y, that is, find numbers x and y such that the two equations are satisfied. Let us perhaps start by adding the equations together to find

$$x + 2x - y + y = 2 + 4, \qquad \text{or} \qquad 3x = 6.$$

In other words, $x = 2$. Once we have that, we plug in $x = 2$ into the first equation to find $2 - y = 2$, so $y = 0$. OK, that was easy. What is all this fuss about linear equations. Well,

try doing this if you have 5000 unknowns*. Also, we may have such equations not of just numbers, but of functions and derivatives of functions in differential equations. Clearly we need a more systematic way of doing things. A nice consequence of making things systematic and simpler to write down is that it becomes easier to have computers do the work for us. Computers are rather stupid, they do not think, but are very good at doing lots of repetitive tasks precisely, as long as we figure out a systematic way for them to perform the tasks.

3.1.1 Vectors and operations on vectors

Consider n real numbers as an n-tuple:

$$(x_1, x_2, \ldots, x_n).$$

The set of such n-tuples is the so-called *n-dimensional space*, often denoted by $\mathbb{R}^n$. Sometimes we call this the n-dimensional *euclidean space*†. In two dimensions, $\mathbb{R}^2$ is called the *cartesian plane*‡, and in three dimensions, it is the same "3-dimensional space" that is dealt with in multivariable calculus. Each such n-tuple represents a point in the n-dimensional space. For example, the point $(1, 2)$ in the plane $\mathbb{R}^2$ is one unit to the right and two units up from the origin.

When we do algebra with these n-tuples of numbers we call them *vectors*§. Mathematicians are keen on separating what is a vector and what is a point of the space or in the plane, and it turns out to be an important distinction, however, for the purposes of linear algebra we can think of everything being represented by a vector. A way to think of a vector, which is especially useful in calculus and differential equations, is an arrow. It is an object that has a *direction* and a *magnitude*. For example, the vector $(1, 2)$ is the arrow from the origin to the point $(1, 2)$ in the plane. The magnitude is the length of the arrow. See Figure 3.1 on the facing page. If we think of vectors as arrows, the arrow doesn't always have to start at the origin. If we do move it around, however, it should always keep the same direction and the same magnitude.

As vectors are arrows, when we want to give a name to a vector, we draw a little arrow above it:

$$\vec{x}$$

Another popular notation is $\mathbf{x}$, although we will use the little arrows. It may be easy to write a bold letter in a book, but it is not so easy to write it by hand on paper or on the board. Mathematicians often don't even write the arrows. A mathematician would write x and just remember that x is a vector and not a number. Just like you remember that Bob is your uncle, and you don't have to keep repeating "Uncle Bob" and you can just say "Bob." In this book, however, we will call Bob "Uncle Bob" and write vectors with the little arrows.

*One of the downsides of making everything look like a linear problem is that the number of variables tends to become huge.

†Named after the ancient Greek mathematician Euclid of Alexandria (around 300 BC), possibly the most famous of mathematicians; even small towns often have Euclid Street or Euclid Avenue.

‡Named after the French mathematician René Descartes (1596–1650). It is "cartesian" as his name in Latin is Renatus Cartesius.

§A common notation to distinguish vectors from points is to write $(1, 2)$ for the point and $\langle 1, 2 \rangle$ for the vector. We write both as $(1, 2)$.

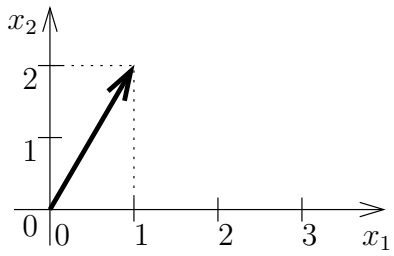


Figure 3.1: The vector $(1, 2)$ *drawn as an arrow from the origin to the point* $(1, 2)$.

The *magnitude* can be computed using Pythagorean theorem. The vector $(1, 2)$ drawn in the figure has magnitude $\sqrt{1^2 + 2^2} = \sqrt{5}$. The magnitude is denoted by $\|\vec{x}\|$, and, in any number of dimensions, it can be computed in the same way:

$$\|\vec{x}\| = \|(x_1, x_2, \ldots, x_n)\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

For reasons that will become clear in the next section, we often write vectors as so-called *column vectors*:

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Don't worry. It is just a different way of writing the same thing, and it will be useful later. For example, the vector $(1, 2)$ can be written as

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

The fact that we write arrows above vectors allows us to write several vectors $\vec{x}_1$, $\vec{x}_2$, etc., without confusing these with the components of some other vector $\vec{x}$.

So where is the *algebra* from *linear algebra*? Well, arrows can be added, subtracted, and multiplied by numbers. First we consider *addition*. If we have two arrows, we simply move along one, and then along the other. See Figure 3.2.

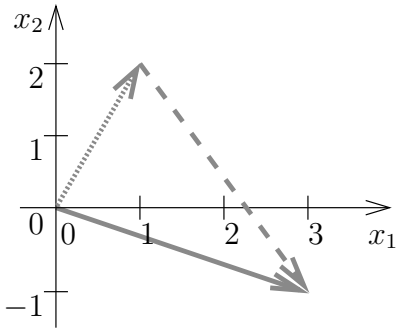


Figure 3.2: Adding the vectors $(1, 2)$*, drawn dotted, and* $(2, -3)$*, drawn dashed. The result,* $(3, -1)$*, is drawn as a solid arrow.*

It is rather easy to see what it does to the numbers that represent the vectors. Suppose we want to add $(1, 2)$ to $(2, -3)$ as in the figure. So we travel along $(1, 2)$ and then we travel along $(2, -3)$ in the sense of "tip-to-tail" addition that you may have seen in previous classes. What we did was travel one unit

right, two units up, and then we travelled two units right, and three units down (the negative three). That means that we ended up at $\big(1+2, 2+(-3)\big) = (3, -1)$. And that's how addition always works:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix}.$$

Subtracting is similar. What $\vec{x} - \vec{y}$ means visually is that we first travel along $\vec{x}$, and then we travel backwards along $\vec{y}$. See Figure 3.3. It is like adding $\vec{x} + (-\vec{y})$ where $-\vec{y}$ is the arrow we obtain by erasing the arrow head from one side and drawing it on the other side, that is, we reverse the direction. In terms of the numbers, we simply go backwards in both directions, so we negate both numbers. For example, if $\vec{y}$ is $(-2, 1)$, then $-\vec{y}$ is $(2, -1)$.

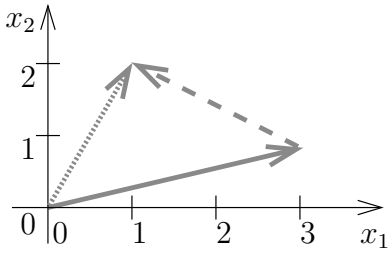


Figure 3.3: Subtraction, the vector $(1, 2)$, drawn dotted, minus $(-2, 1)$, drawn dashed. The result, $(3, 1)$, is drawn as a solid arrow.

Another intuitive thing to do to a vector is to *scale* it. We represent this by multiplication of a number with a vector. Because of this, when we wish to distinguish between vectors and numbers, we call the numbers *scalars*. For example, suppose we want to travel three times further. If the vector is $(1, 2)$, travelling 3 times further means going 3 units to the right and 6 units up, so we get the vector $(3, 6)$. We just multiply each number in the vector by 3. If α is a number, then

$$\alpha \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{bmatrix}.$$

Scaling (by a positive number) multiplies the magnitude and leaves direction untouched. The magnitude of $(1, 2)$ is $\sqrt{5}$. The magnitude of 3 times $(1, 2)$, that is, $(3, 6)$, is $3\sqrt{5}$.

When the scalar is negative, then when we multiply a vector by it, the vector is not only scaled, but it also switches direction. So multiplying $(1, 2)$ by -3 means we should go 3 times further but in the opposite direction, so 3 units to the left and 6 units down, or in other words, $(-3, -6)$. As we mentioned above, $-\vec{y}$ is a reverse of $\vec{y}$, and this is the same as $(-1)\vec{y}$.

In Figure 3.4 on the next page, you can see a couple of examples of what scaling a vector means visually.

We put all of these operations together to work out more complicated expressions. Let us compute a small example:

$$3\begin{bmatrix} 1 \\ 2 \end{bmatrix} + 2\begin{bmatrix} -4 \\ -1 \end{bmatrix} - 3\begin{bmatrix} -2 \\ 2 \end{bmatrix} = \begin{bmatrix} 3(1) + 2(-4) - 3(-2) \\ 3(2) + 2(-1) - 3(2) \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

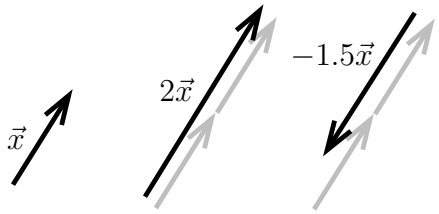


Figure 3.4: A vector $\vec{x}$, the vector $2\vec{x}$ (same direction, double the magnitude), and the vector $-1.5\vec{x}$ (opposite direction, 1.5 times the magnitude).

As we said a vector is a direction and a magnitude. Magnitude is easy to represent, it is just a number. The *direction* is usually given by a vector with magnitude one. We call such a vector a *unit vector*. That is, $\vec{u}$ is a unit vector when $\|\vec{u}\| = 1$. For example, the vectors $(1, 0)$, $(1/\sqrt{2}, 1/\sqrt{2})$, and $(0, -1)$ are all unit vectors.

To represent the direction of a vector $\vec{x}$, we need to find the unit vector in the same direction. To do so, we simply rescale $\vec{x}$ by the reciprocal of the magnitude, that is $\frac{1}{\|\vec{x}\|}\vec{x}$, or more concisely $\frac{\vec{x}}{\|\vec{x}\|}$.

For example, the unit vector in the direction of $(1, 2)$ is the vector

$$\frac{1}{\sqrt{1^2 + 2^2}}(1, 2) = \left(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}\right).$$

3.1.2 Matrices

The next object we need to define here is a *matrix*.

Definition 3.1.1

In general, an $m \times n$ matrix A is a rectangular array of mn numbers,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

An $m \times n$ matrix indicates that it will have m rows and n columns.

Matrices, just like vectors, are generally written with square brackets on the outside, although some books will use parentheses for this. The convention for notation is that matrices will be denoted by capital letters (A) and the individual *entries* of the matrix, the numbers that make it up, will be denoted using lowercase letters (a_{ij}) where the first number i indicates which row of the matrix we are talking about, and the second number j indicates

which column. For example, in the matrix

$$A = \begin{bmatrix} 1 & 4 & 0 \\ -2 & 3 & 1 \\ 2 & 0 & 5 \end{bmatrix},$$

we could talk about the entire matrix usint A, but would also have that $a_{21} = -2$ and $a_{33} = 5$.

Note that an $m \times 1$ matrix is just a column vector, so in terms of the basic structure, matrices are an extension of vectors. However, they can be used for so much more, as we will see in future sections.

Another way to view matrices is as a set of column vectors all laid out side-by-side. If we have $\vec{v}_1$, $\vec{v}_2$ and $\vec{v}_3$, three different four component vectors, we can form a 4×3 matrix B as

$$B = [\vec{v}_1 \mid \vec{v}_2 \mid \vec{v}_3]$$

that uses each of the given vectors as a column of the matrix. In this case, the vertical lines are used to indicate that this is actually a matrix, because each of the entries given there are vectors, not just individual numbers. If we wanted to write a 1×3 matrix this way, these vertical lines will not be included.

We will go into more properties of matrices and the operations we can perform on them in § 3.2. To conclude this section though, we will look at one other way that matrices come about, and that is as the representation of a linear map.

3.1.3 Linear mappings and matrices

A *vector-valued function* F is a rule that takes a vector $\vec{x}$ and returns another vector $\vec{y}$. For example, F could be a scaling that doubles the size of vectors:

$$F(\vec{x}) = 2\vec{x}.$$

For example,

$$F\left(\begin{bmatrix} 1 \\ 3 \end{bmatrix}\right) = 2\begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 6 \end{bmatrix}.$$

If F is a mapping that takes vectors in $\mathbb{R}^2$ to $\mathbb{R}^2$ (such as the above), we write

$$F\colon \mathbb{R}^2 \to \mathbb{R}^2.$$

The words *function* and *mapping* are used rather interchangeably, although more often than not, *mapping* is used when talking about a vector-valued function, and the word *function* is often used when the function is scalar-valued.

A beginning student of mathematics (and many a seasoned mathematician), that sees an expression such as

$$f(3x + 8y)$$

yearns to write

$$3f(x) + 8f(y).$$

After all, who hasn't wanted to write $\sqrt{x+y} = \sqrt{x} + \sqrt{y}$ or something like that at some point in their mathematical lives. Wouldn't life be simple if we could do that? Of course we can't always do that (for example, not with the square roots!) It turns out there are many functions where we can do exactly the above. Such functions are called *linear*.

A mapping $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is called *linear* if

$$F(\vec{x} + \vec{y}) = F(\vec{x}) + F(\vec{y}),$$

for any vectors $\vec{x}$ and $\vec{y}$, and also

$$F(\alpha\vec{x}) = \alpha F(\vec{x}),$$

for any scalar α. The F we defined above that doubles the size of all vectors is linear. Let us check:

$$F(\vec{x} + \vec{y}) = 2(\vec{x} + \vec{y}) = 2\vec{x} + 2\vec{y} = F(\vec{x}) + F(\vec{y}),$$

and also

$$F(\alpha\vec{x}) = 2\alpha\vec{x} = \alpha 2\vec{x} = \alpha F(\vec{x}).$$

We also call a linear function a *linear transformation*. If you want to be really fancy and impress your friends, you can call it a *linear operator*.

When a mapping is linear we often do not write the parentheses. We write simply

$$F\vec{x}$$

instead of $F(\vec{x})$. We do this because linearity means that the mapping F behaves like multiplying $\vec{x}$ by "something." That something is a matrix.

Now how does a matrix A relate to a linear mapping? Well a matrix tells you where certain special vectors go. Let's give a name to those certain vectors. The *standard basis vectors* of $\mathbb{R}^n$ are

$$\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \vec{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \vec{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \qquad \cdots, \qquad \vec{e}_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

For example, in $\mathbb{R}^3$ these vectors are

$$\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \vec{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \vec{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

You may recall from calculus of several variables that these are sometimes called $\vec{\imath}$, $\vec{\jmath}$, $\vec{k}$.

The reason these are called a *basis* is that every other vector can be written as a *linear combination* of them. For example, in $\mathbb{R}^3$ the vector $(4, 5, 6)$ can be written as

$$4\vec{e}_1 + 5\vec{e}_2 + 6\vec{e}_3 = 4 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 5 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 6 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}.$$

Keep this idea of linear combinations of vectors in mind; we'll see a lot more of it later.

So how does a matrix represent a linear mapping? Well, the columns of the matrix are the vectors where A as a linear mapping takes $\vec{e}_1$, $\vec{e}_2$, etc. For example, consider

$$M = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

As a linear mapping $M\colon \mathbb{R}^2 \to \mathbb{R}^2$ takes $\vec{e}_1 = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ to $\left[\begin{smallmatrix} 1 \\ 3 \end{smallmatrix}\right]$ and $\vec{e}_2 = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$ to $\left[\begin{smallmatrix} 2 \\ 4 \end{smallmatrix}\right]$. In other words,

$$M\vec{e}_1 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \qquad \text{and} \qquad M\vec{e}_2 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}.$$

More generally, if we have an $n \times m$ matrix A, that is we have n rows and m columns, then the mapping $A\colon \mathbb{R}^m \to \mathbb{R}^n$ takes $\vec{e}_j$ to the j^{th} column of A. For example,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \end{bmatrix}$$

represents a mapping from $\mathbb{R}^5$ to $\mathbb{R}^3$ that does

$$A\vec{e}_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix}, \qquad A\vec{e}_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix}, \qquad A\vec{e}_3 = \begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix}, \qquad A\vec{e}_4 = \begin{bmatrix} a_{14} \\ a_{24} \\ a_{34} \end{bmatrix}, \qquad A\vec{e}_5 = \begin{bmatrix} a_{15} \\ a_{25} \\ a_{35} \end{bmatrix}.$$

But what if I have another vector $\vec{x}$? Where does it go? Well we use linearity. First write the vector as a linear combination of the standard basis vectors:

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + x_5 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3 + x_4\vec{e}_4 + x_5\vec{e}_5.$$

Then

$$A\vec{x} = A(x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3 + x_4\vec{e}_4 + x_5\vec{e}_5) = x_1A\vec{e}_1 + x_2A\vec{e}_2 + x_3A\vec{e}_3 + x_4A\vec{e}_4 + x_5A\vec{e}_5.$$

If we know where A takes all the basis vectors, we know where it takes all vectors.

As an example, suppose M is the 2×2 matrix from above, and suppose we wish to find

$$M \begin{bmatrix} -2 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} -2 \\ 0.1 \end{bmatrix} = -2 \begin{bmatrix} 1 \\ 3 \end{bmatrix} + 0.1 \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} -1.8 \\ -5.6 \end{bmatrix}.$$

Every linear mapping from $\mathbb{R}^m$ to $\mathbb{R}^n$ can be represented by an $n \times m$ matrix. You just figure out where it takes the standard basis vectors. Conversely, every $n \times m$ matrix represents a linear mapping. Hence, we may think of matrices being linear mappings, and linear mappings being matrices.

Or can we? In this book we study mostly linear differential operators, and linear differential operators are linear mappings, although they are not acting on $\mathbb{R}^n$, but on an infinite-dimensional space of functions:

$$Lf = g$$

for a function f we get a function g, and L is linear in the sense that

$$L(f+h) = Lf + Lh, \qquad \text{and} \qquad L(\alpha f) = \alpha Lf.$$

for any number (scalars) α and all functions f and h.

So the answer is not really. But if we consider vectors in finite-dimensional spaces $\mathbb{R}^n$ then yes, every linear mapping is a matrix. We have mentioned at the beginning of this section, that we can "make everything a vector." That's not strictly true, but it is true approximately. Those "infinite-dimensional" spaces of functions can be approximated by a finite-dimensional space, and then linear operators are just matrices. So approximately, this is true. And as far as actual computations that we can do on a computer, we can work only with finitely many dimensions anyway. If you ask a computer or your calculator to plot a function, it samples the function at finitely many points and then connects the dots*. It does not actually give you infinitely many values. So the way that you have been using the computer or your calculator so far has already been a certain approximation of the space of functions by a finite-dimensional space.

3.1.4 Exercises

***Exercise* 3.1.1:** *On a piece of graph paper draw the vectors:*

a) $\begin{bmatrix} 2 \\ 5 \end{bmatrix}$ *b)* $\begin{bmatrix} -2 \\ -4 \end{bmatrix}$ *c)* $(3, -4)$

***Exercise* 3.1.2:** *On a piece of graph paper draw the vector* $(1, 2)$ *starting at (based at) the given point:*

a) based at $(0, 0)$ *b) based at* $(1, 2)$ *c) based at* $(0, -1)$

***Exercise* 3.1.3:** *On a piece of graph paper draw the following operations. Draw and label the vectors involved in the operations as well as the result:*

a) $\begin{bmatrix} 1 \\ -4 \end{bmatrix} + \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ *b)* $\begin{bmatrix} -3 \\ 2 \end{bmatrix} - \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ *c)* $3\begin{bmatrix} 2 \\ 1 \end{bmatrix}$

***Exercise* 3.1.4:** *Compute the magnitude of*

a) $\begin{bmatrix} 7 \\ 2 \end{bmatrix}$ *b)* $\begin{bmatrix} -2 \\ 3 \\ 1 \end{bmatrix}$ *c)* $(1, 3, -4)$

*In Matlab, you may have noticed that to plot a function, we take a vector of inputs, ask Matlab to compute the corresponding vector of values of the function, and then we ask it to plot the result.

Exercise 3.1.5:* *Compute the magnitude of*

a) $\begin{bmatrix}1\\3\end{bmatrix}$ *b)* $\begin{bmatrix}2\\3\\-1\end{bmatrix}$ *c)* $(-2,1,-2)$

Exercise 3.1.6: *Compute*

a) $\begin{bmatrix}2\\3\end{bmatrix}+\begin{bmatrix}7\\-8\end{bmatrix}$ *b)* $\begin{bmatrix}-2\\3\end{bmatrix}-\begin{bmatrix}6\\-4\end{bmatrix}$ *c)* $-\begin{bmatrix}-3\\2\end{bmatrix}$

d) $4\begin{bmatrix}-1\\5\end{bmatrix}$ *e)* $5\begin{bmatrix}1\\0\end{bmatrix}+9\begin{bmatrix}0\\1\end{bmatrix}$ *f)* $3\begin{bmatrix}1\\-8\end{bmatrix}-2\begin{bmatrix}3\\-1\end{bmatrix}$

Exercise 3.1.7:* *Compute*

a) $\begin{bmatrix}3\\1\end{bmatrix}+\begin{bmatrix}6\\-3\end{bmatrix}$ *b)* $\begin{bmatrix}-1\\2\end{bmatrix}-\begin{bmatrix}2\\-1\end{bmatrix}$ *c)* $-\begin{bmatrix}-5\\3\end{bmatrix}$

d) $2\begin{bmatrix}-2\\4\end{bmatrix}$ *e)* $3\begin{bmatrix}1\\0\end{bmatrix}+7\begin{bmatrix}0\\1\end{bmatrix}$ *f)* $2\begin{bmatrix}2\\-3\end{bmatrix}-6\begin{bmatrix}2\\-1\end{bmatrix}$

Exercise 3.1.8: *Find the unit vector in the direction of the given vector*

a) $\begin{bmatrix}1\\-3\end{bmatrix}$ *b)* $\begin{bmatrix}2\\1\\-1\end{bmatrix}$ *c)* $(3,1,-2)$

Exercise 3.1.9:* *Find the unit vector in the direction of the given vector*

a) $\begin{bmatrix}-1\\1\end{bmatrix}$ *b)* $\begin{bmatrix}1\\-1\\2\end{bmatrix}$ *c)* $(2,-5,2)$

Exercise 3.1.10: *If $\vec{x}=(1,2)$ and $\vec{y}$ are added together, we find $\vec{x}+\vec{y}=(0,2)$. What is $\vec{y}$?*

Exercise 3.1.11: *If $\vec{v}=(1,-4,3)$ and $\vec{w}=(-2,3,-1)$, compute $3\vec{v}-2\vec{w}$ and $4\vec{w}+\vec{v}$.*

Exercise 3.1.12: *Write $(1,2,3)$ as a linear combination of the standard basis vectors $\vec{e}_1$, $\vec{e}_2$, and $\vec{e}_3$.*

Exercise 3.1.13: *If the magnitude of $\vec{x}$ is 4, what is the magnitude of*

a) $0\vec{x}$ *b)* $3\vec{x}$ *c)* $-\vec{x}$ *d)* $-4\vec{x}$ *e)* $\vec{x}+\vec{x}$ *f)* $\vec{x}-\vec{x}$

Exercise 3.1.14:* *If the magnitude of $\vec{x}$ is 5, what is the magnitude of*

a) $4\vec{x}$ *b)* $-2\vec{x}$ *c)* $-4\vec{x}$

Exercise 3.1.15: *Suppose a linear mapping $F\colon \mathbb{R}^2 \to \mathbb{R}^2$ takes $(1,0)$ to $(2,-1)$ and it takes $(0,1)$ to $(3,3)$. Where does it take*

a) $(1,1)$ *b)* $(2,0)$ *c)* $(2,-1)$

***Exercise* 3.1.16:** *Suppose a linear mapping* $F\colon \mathbb{R}^3 \to \mathbb{R}^2$ *takes* $(1,0,0)$ *to* $(2,1)$ *and it takes* $(0,1,0)$ *to* $(3,4)$ *and it takes* $(0,0,1)$ *to* $(5,6)$*. Write down the matrix representing the mapping* F*.*

***Exercise* 3.1.17:** *Suppose that a mapping* $F\colon \mathbb{R}^2 \to \mathbb{R}^2$ *takes* $(1,0)$ *to* $(1,2)$*,* $(0,1)$ *to* $(3,4)$*, and it takes* $(1,1)$ *to* $(0,-1)$*. Explain why* F *is not linear.*

Exercise* 3.1.18: *Suppose a linear mapping* $F\colon \mathbb{R}^2 \to \mathbb{R}^2$ *takes* $(1,0)$ *to* $(1,-1)$ *and it takes* $(0,1)$ *to* $(2,0)$*. Where does it take*

a) $(1,1)$ *b)* $(0,2)$ *c)* $(1,-1)$

***Exercise* 3.1.19** (challenging)**:** *Let* P *represent the space of quadratic polynomials in* t*: a point* (a_0,a_1,a_2) *in* P *represents the polynomial* $a_0+a_1t+a_2t^2$*. Consider the derivative* $\frac{d}{dt}$ *as a mapping of* P *to* P*, and note that* $\frac{d}{dt}$ *is linear. Write down* $\frac{d}{dt}$ *as a* 3×3 *matrix.*

3.2 Matrix algebra

Attribution: [JL], §A.2.

Learning Objectives

After this section, you will be able to:

- Perform addition and multiplication operations on matrices,
- Compute inverses of 2×2 matrices, and
- Identify triangular, diagonal, and symmetric matrices.

3.2.1 One-by-one matrices

Let us motivate what we want to achieve with matrices. What do real-valued linear mappings look like? A linear function of real numbers that you have seen in calculus is of the form

$$f(x) = mx + b.$$

However, the properties of linear mappings discussed in the previous section are that

$$f(x+y) = f(x) + f(y) \qquad f(ax) = af(x).$$

Plugging in the definition from above gives that

$$\begin{aligned} f(x+y) &= m(x+y) + b = mx + my + b \\ f(ax) &= m(ax) + b = a(mx) + b \end{aligned}$$

and neither of these match up appropriately, since

$$\begin{aligned} f(x) + f(y) &= mx + b + my + b = mx + my + 2b \\ af(x) + a(mx+b) &= a(mx) + ab \end{aligned}.$$

In order for these to work, we need to have $b = 0$. Therefore, real-valued linear mappings of the real line, linear functions that eat numbers and spit out numbers, are just multiplications by a number.

Consider a mapping defined by multiplying by a number. Let's call this number α. The mapping then takes x to αx. What we can do is to *add* such mappings. If we have another mapping β, then

$$\alpha x + \beta x = (\alpha + \beta)x.$$

We get a new mapping $\alpha + \beta$ that multiplies x by, well, $\alpha + \beta$. If D is a mapping that doubles things, $Dx = 2x$, and T is a mapping that triples, $Tx = 3x$, then $D + T$ is a mapping that multiplies by 5, $(D + T)x = 5x$.

Similarly we can *compose* such mappings, that is, we could apply one and then the other. We take x, we run it through the first mapping α to get α times x, then we run αx through the second mapping β. In other words,

$$\beta(\alpha x) = (\beta\alpha)x.$$

We just multiply those two numbers. Using our doubling and tripling mappings, if we double and then triple, that is $T(Dx)$ then we obtain $3(2x) = 6x$. The composition TD is the mapping that multiplies by 6. For larger matrices, composition also ends up being a kind of multiplication.

3.2.2 Matrix addition and scalar multiplication

The mappings that multiply numbers by numbers are just 1×1 matrices. The number α above could be written as a matrix $[\alpha]$. So perhaps we would want to do the same things to all matrices that we did to those 1×1 matrices at the start of this section above. First, let us add matrices. If we have a matrix A and a matrix B that are of the same size, say $m \times n$, then they are mappings from $\mathbb{R}^n$ to $\mathbb{R}^m$. The mapping $A + B$ should also be a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$, and it should do the following to vectors:

$$(A + B)\vec{x} = A\vec{x} + B\vec{x}.$$

It turns out you just add the matrices element-wise: If the ij^{th} entry of A is a_{ij}, and the ij^{th} entry of B is b_{ij}, then the ij^{th} entry of $A + B$ is $a_{ij} + b_{ij}$. If

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \qquad \text{and} \qquad B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix},$$

then

$$A + B = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & a_{13} + b_{13} \\ a_{21} + b_{21} & a_{22} + b_{22} & a_{23} + b_{23} \end{bmatrix}.$$

Let us illustrate on a more concrete example:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} + \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & -1 \end{bmatrix} = \begin{bmatrix} 1+7 & 2+8 \\ 3+9 & 4+10 \\ 5+11 & 6-1 \end{bmatrix} = \begin{bmatrix} 8 & 10 \\ 12 & 14 \\ 16 & 5 \end{bmatrix}.$$

Let's check that this does the right thing to a vector. Let's use some of the vector algebra that we already know, and regroup things:

$$\begin{aligned} \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} + \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} &= \left(2 \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} \right) + \left(2 \begin{bmatrix} 7 \\ 9 \\ 11 \end{bmatrix} - \begin{bmatrix} 8 \\ 10 \\ -1 \end{bmatrix} \right) \\ &= 2 \left(\begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} + \begin{bmatrix} 7 \\ 9 \\ 11 \end{bmatrix} \right) - \left(\begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} + \begin{bmatrix} 8 \\ 10 \\ -1 \end{bmatrix} \right) \\ &= 2 \begin{bmatrix} 1+7 \\ 3+9 \\ 5+11 \end{bmatrix} - \begin{bmatrix} 2+8 \\ 4+10 \\ 6-1 \end{bmatrix} = 2 \begin{bmatrix} 8 \\ 12 \\ 16 \end{bmatrix} - \begin{bmatrix} 10 \\ 14 \\ 5 \end{bmatrix} \\ &= \begin{bmatrix} 8 & 10 \\ 12 & 14 \\ 16 & 5 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad \left(= \begin{bmatrix} 2(8) - 10 \\ 2(12) - 14 \\ 2(16) - 5 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \\ 27 \end{bmatrix} \right). \end{aligned}$$

If we replaced the numbers by letters that would constitute a proof! You'll notice that we didn't really have to even compute what the result is to convince ourselves that the two expressions were equal.

If the sizes of the matrices do not match, then addition is not defined. If A is 3×2 and B is 2×5, then we cannot add these matrices. We don't know what that could possibly mean.

It is also useful to have a matrix that when added to any other matrix does nothing. This is the zero matrix, the matrix of all zeros:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

We often denote the zero matrix by 0 without specifying size. We would then just write $A + 0$, where we just assume that 0 is the zero matrix of the same size as A.

There are really two things we can multiply matrices by. We can multiply matrices by scalars or we can multiply by other matrices. Let us first consider multiplication by scalars. For a matrix A and a scalar α we want αA to be the matrix that accomplishes

$$(\alpha A)\vec{x} = \alpha(A\vec{x}).$$

That is just scaling the result by α. If you think about it, scaling every term in A by α accomplishes just that: If

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}, \qquad \text{then} \qquad \alpha A = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \alpha a_{13} \\ \alpha a_{21} & \alpha a_{22} & \alpha a_{23} \end{bmatrix}.$$

For example,

$$2 \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{bmatrix}.$$

Let us list some properties of matrix addition and scalar multiplication. Denote by 0 the zero matrix, by α, β scalars, and by A, B, C matrices. Then:

$$\begin{aligned} A + 0 &= A = 0 + A, \\ A + B &= B + A, \\ (A + B) + C &= A + (B + C), \\ \alpha(A + B) &= \alpha A + \alpha B, \\ (\alpha + \beta)A &= \alpha A + \beta A. \end{aligned}$$

These rules should look very familiar.

3.2.3 Matrix multiplication

As we mentioned above, composition of linear mappings is also a multiplication of matrices. Suppose A is an $m \times n$ matrix, that is, A takes $\mathbb{R}^n$ to $\mathbb{R}^m$, and B is an $n \times p$ matrix, that is, B takes $\mathbb{R}^p$ to $\mathbb{R}^n$. The composition AB should work as follows

$$AB\vec{x} = A(B\vec{x}).$$

First, a vector $\vec{x}$ in $\mathbb{R}^p$ gets taken to the vector $B\vec{x}$ in $\mathbb{R}^n$. Then the mapping A takes it to the vector $A(B\vec{x})$ in $\mathbb{R}^m$. In other words, the composition AB should be an $m \times p$ matrix. In terms of sizes we should have

$$\text{“} \quad [m \times n]\,[n \times p] = [m \times p]. \quad \text{”}$$

Notice how the middle size must match.

OK, now we know what sizes of matrices we should be able to multiply, and what the product should be. Let us see how to actually compute matrix multiplication. We start with the so-called *dot product* (or *inner product*) of two vectors. Usually this is a row vector multiplied with a column vector of the same size. Dot product multiplies each pair of entries from the first and the second vector and sums these products. The result is a single number. For example,

$$\begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = a_1b_1 + a_2b_2 + a_3b_3.$$

And similarly for larger (or smaller) vectors. A dot product is really a product of two matrices: a $1 \times n$ matrix and an $n \times 1$ matrix resulting in a 1×1 matrix, that is, a number.

Armed with the dot product we define the *product of matrices*. First let us denote by $\text{row}_i(A)$ the i^{th} row of A and by $\text{column}_j(A)$ the j^{th} column of A. For an $m \times n$ matrix A and an $n \times p$ matrix B we can compute the product AB. The matrix AB is an $m \times p$ matrix whose ij^{th} entry is the dot product

$$\text{row}_i(A) \cdot \text{column}_j(B).$$

For example, given a 2×3 and a 3×2 matrix we should end up with a 2×2 matrix:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix}, \quad (3.1)$$

or with some numbers:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} -1 & 2 \\ -7 & 0 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 \cdot (-1) + 2 \cdot (-7) + 3 \cdot 1 & 1 \cdot 2 + 2 \cdot 0 + 3 \cdot (-1) \\ 4 \cdot (-1) + 5 \cdot (-7) + 6 \cdot 1 & 4 \cdot 2 + 5 \cdot 0 + 6 \cdot (-1) \end{bmatrix} = \begin{bmatrix} -12 & -1 \\ -33 & 2 \end{bmatrix}.$$

A useful consequence of the definition is that the evaluation $A\vec{x}$ for a matrix A and a (column) vector $\vec{x}$ is also matrix multiplication. That is really why we think of vectors as column vectors, or $n \times 1$ matrices. For example,

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 2 + 2 \cdot (-1) \\ 3 \cdot 2 + 4 \cdot (-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

If you look at the last section, that is precisely the last example we gave.

You should stare at the computation of multiplication of matrices AB and the previous definition of $A\vec{y}$ as a mapping for a moment. What we are doing with matrix multiplication

is applying the mapping A to the columns of B. This is usually written as follows. Suppose we write the $n \times p$ matrix $B = [\vec{b}_1 \ \vec{b}_2 \ \cdots \ \vec{b}_p]$, where $\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_p$ are the columns of B. Then for an $m \times n$ matrix A,

$$AB = A[\vec{b}_1 \ \vec{b}_2 \ \cdots \ \vec{b}_p] = [A\vec{b}_1 \ A\vec{b}_2 \ \cdots \ A\vec{b}_p].$$

The columns of the $m \times p$ matrix AB are the vectors $A\vec{b}_1, A\vec{b}_2, \ldots, A\vec{b}_p$. For example, in (3.1), the columns of

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}$$

are

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{12} \\ b_{22} \\ b_{32} \end{bmatrix}.$$

This is a very useful way to understand what matrix multiplication is. It should also make it easier to remember how to perform matrix multiplication.

3.2.4 Some rules of matrix algebra

For multiplication we want an analogue of a 1. That is, we desire a matrix that just leaves everything as it found it. This analogue is the so-called *identity matrix*. The identity matrix is a square matrix with 1s on the main diagonal and zeros everywhere else. It is usually denoted by I. For each size we have a different identity matrix and so sometimes we may denote the size as a subscript. For example, the I_3 would be the 3×3 identity matrix

$$I = I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Let us see how the matrix works on a smaller example,

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} \cdot 1 + a_{12} \cdot 0 & a_{11} \cdot 0 + a_{12} \cdot 1 \\ a_{21} \cdot 1 + a_{22} \cdot 0 & a_{21} \cdot 0 + a_{22} \cdot 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Multiplication by the identity from the left looks similar, and also does not touch anything.

We have the following rules for matrix multiplication. Suppose that A, B, C are matrices of the correct sizes so that the following make sense. Let α denote a scalar (number). Then

$$\begin{aligned} A(BC) &= (AB)C && \text{(associative law)}, \\ A(B+C) &= AB + AC && \text{(distributive law)}, \\ (B+C)A &= BA + CA && \text{(distributive law)}, \\ \alpha(AB) &= (\alpha A)B = A(\alpha B), \\ IA &= A = AI && \text{(identity)}. \end{aligned}$$

Example 3.2.1: Let us demonstrate a couple of these rules. For example, the associative law:

$$\underbrace{\begin{bmatrix}-3 & 3\\ 2 & -2\end{bmatrix}}_{A}\left(\underbrace{\begin{bmatrix}4 & 4\\ 1 & -3\end{bmatrix}}_{B}\underbrace{\begin{bmatrix}-1 & 4\\ 5 & 2\end{bmatrix}}_{C}\right) = \underbrace{\begin{bmatrix}-3 & 3\\ 2 & -2\end{bmatrix}}_{A}\underbrace{\begin{bmatrix}16 & 24\\ -16 & -2\end{bmatrix}}_{BC} = \underbrace{\begin{bmatrix}-96 & -78\\ 64 & 52\end{bmatrix}}_{A(BC)},$$

and

$$\left(\underbrace{\begin{bmatrix}-3 & 3\\ 2 & -2\end{bmatrix}}_{A}\underbrace{\begin{bmatrix}4 & 4\\ 1 & -3\end{bmatrix}}_{B}\right)\underbrace{\begin{bmatrix}-1 & 4\\ 5 & 2\end{bmatrix}}_{C} = \underbrace{\begin{bmatrix}-9 & -21\\ 6 & 14\end{bmatrix}}_{AB}\underbrace{\begin{bmatrix}-1 & 4\\ 5 & 2\end{bmatrix}}_{C} = \underbrace{\begin{bmatrix}-96 & -78\\ 64 & 52\end{bmatrix}}_{(AB)C}.$$

Or how about multiplication by scalars:

$$10\left(\underbrace{\begin{bmatrix}-3 & 3\\ 2 & -2\end{bmatrix}}_{A}\underbrace{\begin{bmatrix}4 & 4\\ 1 & -3\end{bmatrix}}_{B}\right) = 10\underbrace{\begin{bmatrix}-9 & -21\\ 6 & 14\end{bmatrix}}_{AB} = \underbrace{\begin{bmatrix}-90 & -210\\ 60 & 140\end{bmatrix}}_{10(AB)},$$

$$\left(10\underbrace{\begin{bmatrix}-3 & 3\\ 2 & -2\end{bmatrix}}_{A}\right)\underbrace{\begin{bmatrix}4 & 4\\ 1 & -3\end{bmatrix}}_{B} = \underbrace{\begin{bmatrix}-30 & 30\\ 20 & -20\end{bmatrix}}_{10A}\underbrace{\begin{bmatrix}4 & 4\\ 1 & -3\end{bmatrix}}_{B} = \underbrace{\begin{bmatrix}-90 & -210\\ 60 & 140\end{bmatrix}}_{(10A)B},$$

and

$$\underbrace{\begin{bmatrix}-3 & 3\\ 2 & -2\end{bmatrix}}_{A}\left(10\underbrace{\begin{bmatrix}4 & 4\\ 1 & -3\end{bmatrix}}_{B}\right) = \underbrace{\begin{bmatrix}-3 & 3\\ 2 & -2\end{bmatrix}}_{A}\underbrace{\begin{bmatrix}40 & 40\\ 10 & -30\end{bmatrix}}_{10B} = \underbrace{\begin{bmatrix}-90 & -210\\ 60 & 140\end{bmatrix}}_{A(10B)}.$$

A multiplication rule you have used since primary school on numbers is quite conspicuously missing for matrices. That is, matrix multiplication is not commutative. Firstly, just because AB makes sense, it may be that BA is not even defined. For example, if A is 2×3, and B is 3×4, the we can multiply AB but not BA.

Even if AB and BA are both defined, does not mean that they are equal. For example, take $A = \left[\begin{smallmatrix}1 & 1\\ 1 & 1\end{smallmatrix}\right]$ and $B = \left[\begin{smallmatrix}1 & 0\\ 0 & 2\end{smallmatrix}\right]$:

$$AB = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix} = \begin{bmatrix}1 & 2\\ 1 & 2\end{bmatrix} \qquad \neq \qquad \begin{bmatrix}1 & 1\\ 2 & 2\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix} = BA.$$

3.2.5 Inverse

A couple of other algebra rules you know for numbers do not quite work on matrices:

(i) $AB = AC$ does not necessarily imply $B = C$, even if A is not 0.

(ii) $AB = 0$ does not necessarily mean that $A = 0$ or $B = 0$.

For example:

$$\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix} = \begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix} = \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}\begin{bmatrix}0 & 2\\ 0 & 0\end{bmatrix}.$$

To make these rules hold, we do not just need one of the matrices to not be zero, we would need to "divide" by a matrix. This is where the *matrix inverse* comes in.

Definition 3.2.1

Suppose that A and B are $n \times n$ matrices such that

$$AB = I = BA.$$

Then we call B the inverse of A and we denote B by A^{-1}.
If the inverse of A exists, then we say A is *invertible*. If A is not invertible, we say A is *singular*.

Perhaps not surprisingly, $(A^{-1})^{-1} = A$, since if the inverse of A is B, then the inverse of B is A.

If $A = [a]$ is a 1×1 matrix, then A^{-1} is $a^{-1} = \frac{1}{a}$. That is where the notation comes from. The computation is not nearly as simple when A is larger.

The proper formulation of the cancellation rule is:

$$\textit{If } A \textit{ is invertible, then } AB = AC \textit{ implies } B = C.$$

The computation is what you would do in regular algebra with numbers, but you have to be careful never to commute matrices:

$$\begin{aligned} AB &= AC, \\ A^{-1}AB &= A^{-1}AC, \\ IB &= IC, \\ B &= C. \end{aligned}$$

And similarly for cancellation on the right:

$$\textit{If } A \textit{ is invertible, then } BA = CA \textit{ implies } B = C.$$

The rule says, among other things, that the inverse of a matrix is unique if it exists: If $AB = I = AC$, then A is invertible and $B = C$.

We will see later how to compute an inverse of a matrix in general. For now, let us note that there is a simple formula for the inverse of a 2×2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

For example:

$$\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix}^{-1} = \frac{1}{1 \cdot 4 - 1 \cdot 2} \begin{bmatrix} 4 & -1 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -1/2 \\ -1 & 1/2 \end{bmatrix}.$$

Let's try it:

$$\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 2 & -1/2 \\ -1 & 1/2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 & -1/2 \\ -1 & 1/2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Just as we cannot divide by every number, not every matrix is invertible. In the case of matrices however we may have singular matrices that are not zero. For example,

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$$

is a singular matrix. But didn't we just give a formula for an inverse? Let us try it:

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}^{-1} = \frac{1}{1 \cdot 2 - 1 \cdot 2} \begin{bmatrix} 2 & -1 \\ -2 & 1 \end{bmatrix} = ?$$

We get into a bit of trouble; we are trying to divide by zero.

So a 2×2 matrix A is invertible whenever

$$ad - bc \neq 0$$

and otherwise it is singular. The expression $ad - bc$ is called the *determinant* and we will look at it more carefully in a later section. There is a similar expression for a square matrix of any size.

3.2.6 Special types of matrices

A simple (and surprisingly useful) type of a square matrix is a so-called *diagonal matrix*. It is a matrix whose entries are all zero except those on the main diagonal from top left to bottom right. For example a 4×4 diagonal matrix is of the form

$$\begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & d_3 & 0 \\ 0 & 0 & 0 & d_4 \end{bmatrix}.$$

Such matrices have nice properties when we multiply by them. If we multiply them by a vector, they multiply the k^{th} entry by d_k. For example,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 \cdot 4 \\ 2 \cdot 5 \\ 3 \cdot 6 \end{bmatrix} = \begin{bmatrix} 4 \\ 10 \\ 18 \end{bmatrix}.$$

Similarly, when they multiply another matrix from the left, they multiply the k^{th} row by d_k. For example,

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 2 \\ 3 & 3 & 3 \\ -1 & -1 & -1 \end{bmatrix}.$$

On the other hand, multiplying on the right, they multiply the columns:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 3 & -1 \\ 2 & 3 & -1 \\ 2 & 3 & -1 \end{bmatrix}.$$

And it is really easy to multiply two diagonal matrices together:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 2 & 0 & 0 \\ 0 & 2 \cdot 3 & 0 \\ 0 & 0 & 3 \cdot (-1) \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & -3 \end{bmatrix}.$$

For this last reason, they are easy to invert, you simply invert each diagonal element:

$$\begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix}^{-1} = \begin{bmatrix} d_1^{-1} & 0 & 0 \\ 0 & d_2^{-1} & 0 \\ 0 & 0 & d_3^{-1} \end{bmatrix}.$$

Let us check an example

$$\underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}^{-1}}_{A^{-1}} \underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}}_{A^{-1}} \underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{I}.$$

It is no wonder that the way we solve many problems in linear algebra (and in differential equations) is to try to reduce the problem to the case of diagonal matrices.

Another type of matrix that has similarly nice properties are *triangular* matrices. A matrix is *upper triangular* if all of the entries below the diagonal are zero. For a 3×3 matrix, an upper triangular matrix looks like

$$\begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix}$$

where the $*$ can be any number. Similarly, a *lower triangular* matrix is one where all of the entries above the diagonal are zero, or, for a 3×3 matrix, something that looks like

$$\begin{bmatrix} * & 0 & 0 \\ * & * & 0 \\ * & * & * \end{bmatrix}.$$

A matrix that is both upper and lower triangular is diagonal, because only the entries on the diagonal can be non-zero.

3.2.7 Transpose

Vectors do not always have to be column vectors, that is just a convention. Swapping rows and columns is from time to time needed. The operation that swaps rows and columns is the so-called *transpose*. The transpose of A is denoted by A^T. Example:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

So transpose takes an $m \times n$ matrix to an $n \times m$ matrix.

A key fact about the transpose is that if the product AB makes sense then $B^T A^T$ also makes sense, at least from the point of view of sizes. In fact, we get precisely the transpose of AB. That is:

$$(AB)^T = B^T A^T.$$

For example,

$$\left(\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & -2 \end{bmatrix} \right)^T = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

It is left to the reader to verify that computing the matrix product on the left and then transposing is the same as computing the matrix product on the right.

If we have a column vector $\vec{x}$ to which we apply a matrix A and we transpose the result, then the row vector $\vec{x}^T$ applies to A^T from the left:

$$(A\vec{x})^T = \vec{x}^T A^T.$$

Another place where transpose is useful is when we wish to apply the dot product* to two column vectors:

$$\vec{x} \cdot \vec{y} = \vec{y}^T \vec{x}.$$

That is the way that one often writes the dot product in software.

We say a matrix A is *symmetric* if $A = A^T$. For example,

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$

is a symmetric matrix. Notice that a symmetric matrix is always square, that is, $n \times n$. Symmetric matrices have many nice properties†, and come up quite often in applications.

To end the section, we notice how $A\vec{x}$ can be written more succintly. Suppose

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \qquad \text{and} \qquad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Then

$$A\vec{x} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \end{bmatrix}.$$

For example,

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 2 + 2 \cdot (-1) \\ 3 \cdot 2 + 4 \cdot (-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

In other words, you take a row of the matrix, you multiply them by the entries in your vector, you add things up, and that's the corresponding entry in the resulting vector.

*As a side note, mathematicians write $\vec{y}^T \vec{x}$ and physicists write $\vec{x}^T \vec{y}$. Shhh... don't tell anyone, but the physicists are probably right on this.

†Although so far we have not learned enough about matrices to really appreciate them.

3.2.8 Exercises

***Exercise* 3.2.1:** *Add the following matrices*

a) $\begin{bmatrix} -1 & 2 & 2 \\ 5 & 8 & -1 \end{bmatrix} + \begin{bmatrix} 3 & 2 & 3 \\ 8 & 3 & 5 \end{bmatrix}$

b) $\begin{bmatrix} 1 & 2 & 4 \\ 2 & 3 & 1 \\ 0 & 5 & 1 \end{bmatrix} + \begin{bmatrix} 2 & -8 & -3 \\ 3 & 1 & 0 \\ 6 & -4 & 1 \end{bmatrix}$

Exercise* 3.2.2: *Add the following matrices*

a) $\begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & -1 \end{bmatrix} + \begin{bmatrix} 5 & 3 & 4 \\ 1 & 2 & 5 \end{bmatrix}$

b) $\begin{bmatrix} 6 & -2 & 3 \\ 7 & 3 & 3 \\ 8 & -1 & 2 \end{bmatrix} + \begin{bmatrix} -1 & -1 & -3 \\ 6 & 7 & 3 \\ -9 & 4 & -1 \end{bmatrix}$

***Exercise* 3.2.3:** *Compute*

a) $3 \begin{bmatrix} 0 & 3 \\ -2 & 2 \end{bmatrix} + 6 \begin{bmatrix} 1 & 5 \\ -1 & 5 \end{bmatrix}$

b) $2 \begin{bmatrix} -3 & 1 \\ 2 & 2 \end{bmatrix} - 3 \begin{bmatrix} 2 & -1 \\ 3 & 2 \end{bmatrix}$

Exercise* 3.2.4: *Compute*

a) $2 \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + 3 \begin{bmatrix} -1 & 3 \\ 1 & 2 \end{bmatrix}$

b) $3 \begin{bmatrix} 2 & -1 \\ 1 & 3 \end{bmatrix} - 2 \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}$

***Exercise* 3.2.5:** *Multiply the following matrices*

a) $\begin{bmatrix} -1 & 2 \\ 3 & 1 \\ 5 & 8 \end{bmatrix} \begin{bmatrix} 3 & -1 & 3 & 1 \\ 8 & 3 & 2 & -3 \end{bmatrix}$

b) $\begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 1 \\ 1 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 3 & 1 & 7 \\ 1 & 2 & 3 & -1 \\ 1 & -1 & 3 & 0 \end{bmatrix}$

c) $\begin{bmatrix} 4 & 1 & 6 & 3 \\ 5 & 6 & 5 & 0 \\ 4 & 6 & 6 & 0 \end{bmatrix} \begin{bmatrix} 2 & 5 \\ 1 & 2 \\ 3 & 5 \\ 5 & 6 \end{bmatrix}$

d) $\begin{bmatrix} 1 & 1 & 4 \\ 0 & 5 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 1 & 0 \\ 6 & 4 \end{bmatrix}$

Exercise* 3.2.6: *Multiply the following matrices*

a) $\begin{bmatrix} 2 & 1 & 4 \\ 3 & 4 & 4 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 6 & 3 \\ 3 & 5 \end{bmatrix}$

b) $\begin{bmatrix} 0 & 3 & 3 \\ 2 & -2 & 1 \\ 3 & 5 & -2 \end{bmatrix} \begin{bmatrix} 6 & 6 & 2 \\ 4 & 6 & 0 \\ 2 & 0 & 4 \end{bmatrix}$

c) $\begin{bmatrix} 3 & 4 & 1 \\ 2 & -1 & 0 \\ 4 & -1 & 5 \end{bmatrix} \begin{bmatrix} 0 & 2 & 5 & 0 \\ 2 & 0 & 5 & 2 \\ 3 & 6 & 1 & 6 \end{bmatrix}$

d) $\begin{bmatrix} -2 & -2 \\ 5 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 0 & 3 \\ 1 & 3 \end{bmatrix}$

Exercise 3.2.7:

a) *How must the dimensions of two matrices line up in order to multiply them together? If they can be multiplied, what is the dimension of the product?*

b) *If A is a 3×2 matrix and the product AB is a 3×4 matrix, then what are the dimensions of B?*

c) *If A is a 5×3 matrix, is it possible to find a matrix B so that the product AB is a 4×3 matrix? What about a matrix C so that the product CA is a 4×3 matrix?*

Exercise 3.2.8: *Assume that A is a 3×4 matrix.*

a) *What must the dimensions of B be in order for the product AB to be defined?*

b) *What must the dimensions of B be in order for the product BA to be defined?*

c) *What about if we want to compute ABA or BAB?*

Exercise 3.2.9: *Complete Exercise 3.2.8 but with A being a 2×2 matrix.*

Exercise 3.2.10: *Compute the inverse of the given matrices*

a) $\begin{bmatrix} -3 \end{bmatrix}$ b) $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ c) $\begin{bmatrix} 1 & 4 \\ 1 & 3 \end{bmatrix}$ d) $\begin{bmatrix} 2 & 2 \\ 1 & 4 \end{bmatrix}$

Exercise 3.2.11:* *Compute the inverse of the given matrices*

a) $\begin{bmatrix} 2 \end{bmatrix}$ b) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ c) $\begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix}$ d) $\begin{bmatrix} 4 & 2 \\ 4 & 4 \end{bmatrix}$

Exercise 3.2.12: *Compute the inverse of the given matrices*

a) $\begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix}$ b) $\begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ c) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & -5 \end{bmatrix}$

Exercise 3.2.13:* *Compute the inverse of the given matrices*

a) $\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$ b) $\begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & -1 \end{bmatrix}$ c) $\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix}$

Exercise 3.2.14: *Consider the matrices*

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} \quad C = \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}.$$

a) *Compute the products AB and AC.*

b) *Verify that for these matrices $AB = AC$, but $B \neq C$.*

***Exercise* 3.2.15:** *Consider the matrices*

$$A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Verify that AB and BA are both equal to zero, but neither of the matrices A and B are zero.

***Exercise* 3.2.16:**

a) *Let A be a 3×4 matrix. What dimension does the vector $\vec{v}$ need to be in order for the product $A\vec{v}$ to be defined? If this product is defined, what is the dimension of the product $A\vec{v}$?*

b) *Let B be a 3×3 matrix. What dimension does the vector $\vec{v}$ need to be in order for the product $B\vec{v}$ to be defined? If this product is defined, what is the dimension of the product $B\vec{v}$?*

3.3 Elimination

Attribution: [JL], §A.3.

Learning Objectives

After this section, you will be able to:

- Write a system of linear equations in matrix form,
- Use row reduction to put a matrix into row echlon form or reduced row echelon form, and
- Determine whether a system of linear equations has no solution, one solution, or infinitely many solutions.

3.3.1 Linear systems of equations

One application of matrices is to solve systems of linear equations*. Consider the following system of linear equations

$$\begin{aligned} 2x_1 + 2x_2 + 2x_3 &= 2, \\ x_1 + \ x_2 + 3x_3 &= 5, \\ x_1 + 4x_2 + \ x_3 &= 10. \end{aligned} \tag{3.2}$$

There is a systematic procedure called *elimination* to solve such a system. In this procedure, we attempt to eliminate each variable from all but one equation. We want to end up with equations such as $x_3 = 2$, where we can just read off the answer.

We write a system of linear equations as a matrix equation:

$$A\vec{x} = \vec{b}.$$

The system (3.2) is written as

$$\underbrace{\begin{bmatrix} 2 & 2 & 2 \\ 1 & 1 & 3 \\ 1 & 4 & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{\vec{x}} = \underbrace{\begin{bmatrix} 2 \\ 5 \\ 10 \end{bmatrix}}_{\vec{b}}.$$

If we knew the inverse of A, then we would be done; we would simply solve the equation:

$$\vec{x} = A^{-1}A\vec{x} = A^{-1}\vec{b}.$$

Well, but that is part of the problem, we do not know how to compute the inverse for matrices bigger than 2×2. We will see later that to compute the inverse we are really solving $A\vec{x} = \vec{b}$ for several different $\vec{b}$. In other words, we will need to do elimination to find A^{-1}. In addition, we may wish to solve $A\vec{x} = \vec{b}$ even if A is not invertible, or perhaps not even square.

*Although perhaps we have this backwards, quite often we solve a linear system of equations to find out something about matrices, rather than vice versa.

Let us return to the equations themselves and see how we can manipulate them. There are a few operations we can perform on the equations that do not change the solution. First, perhaps an operation that may seem stupid, we can swap two equations in (3.2):

$$\begin{aligned} x_1 + \ x_2 + 3x_3 &= 5, \\ 2x_1 + 2x_2 + 2x_3 &= 2, \\ x_1 + 4x_2 + \ x_3 &= 10. \end{aligned}$$

Clearly these new equations have the same solutions x_1, x_2, x_3. A second operation is that we can multiply an equation by a nonzero number. For example, we multiply the third equation in (3.2) by 3:

$$\begin{aligned} 2x_1 + \ 2x_2 + 2x_3 &= 2, \\ x_1 + \ x_2 + 3x_3 &= 5, \\ 3x_1 + 12x_2 + 3x_3 &= 30. \end{aligned}$$

Finally we can add a multiple of one equation to another equation. For example, we add 3 times the third equation in (3.2) to the second equation:

$$\begin{aligned} 2x_1 + \ & 2x_2 + \ & 2x_3 &= 2, \\ (1+3)x_1 + \ & (1+12)x_2 + \ & (3+3)x_3 &= 5+30, \\ x_1 + \ & 4x_2 + \ & x_3 &= 10. \end{aligned}$$

The same x_1, x_2, x_3 should still be solutions to the new equations. These were just examples; we did not get any closer to the solution. We must to do these three operations in some more logical manner, but it turns out these three operations suffice to solve every linear equation.

The first thing is to write the equations in a more compact manner. Given

$$A\vec{x} = \vec{b},$$

we write down the so-called *augmented matrix*

$$[A \mid \vec{b}],$$

where the vertical line is just a marker for us to know where the "right-hand side" of the equation starts. For example, for the system (3.2) the augmented matrix is

$$\left[\begin{array}{ccc|c} 2 & 2 & 2 & 2 \\ 1 & 1 & 3 & 5 \\ 1 & 4 & 1 & 10 \end{array}\right].$$

The entire process of elimination, which we will describe, is often applied to any sort of matrix, not just an augmented matrix. Simply think of the matrix as the 3×4 matrix

$$\begin{bmatrix} 2 & 2 & 2 & 2 \\ 1 & 1 & 3 & 5 \\ 1 & 4 & 1 & 10 \end{bmatrix}.$$

3.3.2 Row echelon form and elementary operations

We apply the three operations above to the matrix. We call these the *elementary operations* or *elementary row operations*.

Definition 3.3.1

The elementary row operations on a matrix are:

(i) Swap two rows.

(ii) Multiply a row by a nonzero number.

(iii) Add a multiple of one row to another row.

Note that these are the same three operations that we could do with equations to try to solve them earliner in this section. We run these operations until we get into a state where it is easy to read off the answer, or until we get into a contradiction indicating no solution.

More specifically, we run the operations until we obtain the so-called *row echelon form*. Let us call the first (from the left) nonzero entry in each row the *leading entry*. A matrix is in *row echelon form* if the following conditions are satisfied:

(i) The leading entry in any row is strictly to the right of the leading entry of the row above.

(ii) Any zero rows are below all the nonzero rows.

(iii) All leading entries are 1.

A matrix is in *reduced row echelon form* if furthermore the following condition is satisfied.

(iv) All the entries above a leading entry are zero.

Example 3.3.1: The following matrices are in row echelon form. The leading entries are marked:

$$\begin{bmatrix} \boxed{1} & 2 & 9 & 3 \\ 0 & 0 & \boxed{1} & 5 \\ 0 & 0 & 0 & \boxed{1} \end{bmatrix} \qquad \begin{bmatrix} \boxed{1} & -1 & -3 \\ 0 & \boxed{1} & 5 \\ 0 & 0 & \boxed{1} \end{bmatrix} \qquad \begin{bmatrix} \boxed{1} & 2 & 1 \\ 0 & \boxed{1} & 2 \\ 0 & 0 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & \boxed{1} & -5 & 2 \\ 0 & 0 & 0 & \boxed{1} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Note that the definition applies to matrices of any size. None of the matrices above are in *reduced* row echelon form. For example, in the first matrix none of the entries above the second and third leading entries are zero; they are 9, 3, and 5.

The following matrices are in reduced row echelon form. The leading entries are marked:

$$\begin{bmatrix} \boxed{1} & 3 & 0 & 8 \\ 0 & 0 & \boxed{1} & 6 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \begin{bmatrix} \boxed{1} & 0 & 2 & 0 \\ 0 & \boxed{1} & 3 & 0 \\ 0 & 0 & 0 & \boxed{1} \end{bmatrix} \qquad \begin{bmatrix} \boxed{1} & 0 & 3 \\ 0 & \boxed{1} & -2 \\ 0 & 0 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & \boxed{1} & 2 & 0 \\ 0 & 0 & 0 & \boxed{1} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The procedure we will describe to find a reduced row echelon form of a matrix is called *Gauss–Jordan elimination*. The first part of it, which obtains a row echelon form, is called *Gaussian elimination* or *row reduction*. For some problems, a row echelon form is sufficient, and it is a bit less work to only do this first part.

To attain the row echelon form we work *systematically*. We go column by column, starting at the first column. We find topmost entry in the first column that is not zero, and we call it the *pivot*. If there is no nonzero entry we move to the next column. We swap rows to put the row with the pivot as the first row. We divide the first row by the pivot to make the pivot entry be a 1. Now look at all the rows below and subtract the correct multiple of the pivot row so that all the entries below the pivot become zero.

After this procedure we forget that we had a first row (it is now fixed), and we forget about the column with the pivot and all the preceding zero columns. Below the pivot row, all the entries in these columns are just zero. Then we focus on the smaller matrix and we repeat the steps above.

It is best shown by example, so let us go back to the example from the beginning of the section. We keep the vertical line in the matrix, even though the procedure works on any matrix, not just an augmented matrix. We start with the first column and we locate the pivot, in this case the first entry of the first column.

$$\left[\begin{array}{ccc|c} \boxed{2} & 2 & 2 & 2 \\ 1 & 1 & 3 & 5 \\ 1 & 4 & 1 & 10 \end{array}\right]$$

We multiply the first row by $1/2$.

$$\left[\begin{array}{ccc|c} \boxed{1} & 1 & 1 & 1 \\ 1 & 1 & 3 & 5 \\ 1 & 4 & 1 & 10 \end{array}\right]$$

We subtract the first row from the second and third row (two elementary operations).

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 4 \\ 0 & 3 & 0 & 9 \end{array}\right]$$

We are done with the first column and the first row for now. We almost pretend the matrix doesn't have the first column and the first row.

$$\left[\begin{array}{ccc|c} * & * & * & * \\ * & 0 & 2 & 4 \\ * & 3 & 0 & 9 \end{array}\right]$$

OK, look at the second column, and notice that now the pivot is in the third row.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 4 \\ 0 & \boxed{3} & 0 & 9 \end{array}\right]$$

We swap rows.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & \boxed{3} & 0 & 9 \\ 0 & 0 & 2 & 4 \end{array}\right]$$

And we divide the pivot row by 3.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & \boxed{1} & 0 & 3 \\ 0 & 0 & 2 & 4 \end{array}\right]$$

We do not need to subtract anything as everything below the pivot is already zero. We move on, we again start ignoring the second row and second column and focus on

$$\left[\begin{array}{ccc|c} * & * & * & * \\ * & * & * & * \\ * & * & 2 & 4 \end{array}\right].$$

We find the pivot, then divide that row by 2:

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & \boxed{2} & 4 \end{array}\right] \qquad \rightarrow \qquad \left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 \end{array}\right].$$

The matrix is now in row echelon form.

The equation corresponding to the last row is $x_3 = 2$. We know x_3 and we could substitute it into the first two equations to get equations for x_1 and x_2. Then we could do the same thing with x_2, until we solve for all 3 variables. This procedure is called *backsubstitution* and we can achieve it via elementary operations. We start from the lowest pivot (leading entry in the row echelon form) and subtract the right multiple from the row above to make all the entries above this pivot zero. Then we move to the next pivot and so on. After we are done, we will have a matrix in reduced row echelon form.

We continue our example. Subtract the last row from the first to get

$$\left[\begin{array}{ccc|c} 1 & 1 & 0 & -1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 \end{array}\right].$$

The entry above the pivot in the second row is already zero. So we move onto the next pivot, the one in the second row. We subtract this row from the top row to get

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & -4 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 \end{array}\right].$$

The matrix is in reduced row echelon form.

If we now write down the equations for x_1, x_2, x_3, we find

$$x_1 = -4, \qquad x_2 = 3, \qquad x_3 = 2.$$

In other words, we have solved the system.

Example 3.3.2: Solve the following system of equations using row reduction:

$$\begin{aligned} -x_1 + x_2 + 3x_3 &= 7 \\ -3x_1 \quad + \quad x_3 &= -5 \\ -2x_1 - x_2 &= -4 \end{aligned}$$

Solution: In order to solve this problem, we need to set up the augmented matrix for this system, which is

$$\left[\begin{array}{ccc|c} -1 & 1 & 3 & 7 \\ -3 & 0 & 1 & -5 \\ -2 & -1 & 0 & -4 \end{array}\right]$$

To carry out the process, we need to get a 1 in the top left corner, then work from there. We multiply the first row by -1 to get

$$\left[\begin{array}{ccc|c} 1 & -1 & -3 & -7 \\ -3 & 0 & 1 & -5 \\ -2 & -1 & 0 & -4 \end{array}\right].$$

Next, we want to use row 1 to cancel out the -3 and -2 in column 1. To do this, we add three copies of row 1 to row 2, and two copies of row 1 to row 3 to get the augmented matrix

$$\left[\begin{array}{ccc|c} 1 & -1 & -3 & -7 \\ 0 & -3 & -8 & -26 \\ 0 & -3 & -6 & -18 \end{array}\right].$$

Normally, the next step would be to divide the second row by -3 in order to put a 1 in that pivot spot. However, since both the second and third rows have a -3 in the second column, we can combine these two rows directly without dividing by -3 first. We subtract row 2 from row 3 to get

$$\left[\begin{array}{ccc|c} 1 & -1 & -3 & -7 \\ 0 & -3 & -8 & -26 \\ 0 & 0 & 2 & 8 \end{array}\right]$$

and we can now use this to solve the system. The bottom row says that $2x_3 = 8$, so that $x_3 = 4$. The second row says that $-3x_2 - 8x_3 = -26$, since $x_3 = 4$, we have that $-3x_2 = -26 + 32 = 6$, so $x_2 = -2$. Finally, the first row of the augmented matrix says that $x_1 - x_2 - 3x_3 = -7$. Plugging in our values for x_2 and x_3 gives $x_1 = -7 - 2 + 12 = 3$. Therefore, the solution is

$$x_1 = 3 \quad x_2 = -2 \quad x_3 = 4.$$

3.3.3 Non-unique solutions and inconsistent systems

It is possible that the solution of a linear system of equations is not unique, or that no solution exists. Suppose for a moment that the row echelon form we found was

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{array}\right].$$

Then the last row gives the equation $0x_1 + 0x_2 + 0x_3 = 1$, or $0 = 1$. That is impossible and the equations are *inconsistent*. There is no solution to $A\vec{x} = \vec{b}$.

On the other hand, if we find a row echelon form

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{array}\right],$$

then there is no issue with finding solutions. In fact, we will find way too many. Let us continue with backsubstitution (subtracting 3 times the third row from the first) to find the reduced row echelon form and let's mark the pivots.

$$\left[\begin{array}{ccc|c} \boxed{1} & 2 & 0 & -5 \\ 0 & 0 & \boxed{1} & 3 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

The last row is all zeros; it just says $0 = 0$ and we ignore it. The two remaining equations are

$$x_1 + 2x_2 = -5, \qquad x_3 = 3.$$

Let us solve for the variables that corresponded to the pivots, that is x_1 and x_3 as there was a pivot in the first column and in the third column:

$$\begin{aligned} x_1 &= -2x_2 - 5, \\ x_3 &= 3. \end{aligned}$$

The variable x_2 can be anything you wish and we still get a solution. The x_2 is called a *free variable*. There are infinitely many solutions, one for every choice of x_2. For example, if we pick $x_2 = 0$, then $x_1 = -5$, and $x_3 = 3$ give a solution. But we also get a solution by picking say $x_2 = 1$, in which case $x_1 = -9$ and $x_3 = 3$, or by picking $x_2 = -5$ in which case $x_1 = 5$ and $x_3 = 3$.

The general idea is that if any row has all zeros in the columns corresponding to the variables, but a nonzero entry in the column corresponding to the right-hand side $\vec{b}$, then the system is inconsistent and has no solutions. In other words, the system is inconsistent if you find a pivot on the right side of the vertical line drawn in the augmented matrix. Otherwise, the system is consistent, and at least one solution exists.

If the system is consistent:

(i) If every column corresponding to a variable has a pivot element, then the solution is unique.

(ii) If there are columns corresponding to variables with no pivot, then those are *free variables* that can be chosen arbitrarily, and there are infinitely many solutions.

Another way to interpret this idea of free variables is that at the beginning, before you look at the system of equations, all of the variables can be anything, and there are no constraints on them. The equations then give us constraints on these variables, because they give us

rules that the variables must satisfy. When we have a row of the augmented matrix that becomes all zeros, it means that the equation that was there is redundant and doesn't add any constraints to the equations. This may result in an *underdetermined* system, which will likely have free variables.

Example 3.3.3: Solve the following two systems of equations, or determine that no solution exists, using row reduction:

$$\begin{aligned} x_1 - \ x_2 - 3x_3 &= -3 \\ -x_1 - 2x_2 + 4x_3 &= 6 \\ x_1 + 5x_2 - 5x_3 &= -9 \end{aligned} \qquad \begin{aligned} x_1 - \ x_2 - 3x_3 &= -3 \\ -x_1 - 2x_2 + 4x_3 &= 6 \\ x_1 + 5x_2 - 5x_3 &= 1 \end{aligned}$$

Solution: For the first of these systems, we will set up the augmented matrix and proceed through the process like normal. The augmented matrix is

$$\left[\begin{array}{ccc|c} 1 & -1 & -3 & -3 \\ -1 & -2 & 4 & 6 \\ 1 & 5 & -5 & -9 \end{array}\right].$$

Since we already have a 1 in the top-left corner of this matrix, we can use it to cancel the entries in the rest of column 1. We add one copy of row 1 to row 2, and subtract row 1 from row 3 to get the next augmented form matrix as

$$\left[\begin{array}{ccc|c} 1 & -1 & -3 & -3 \\ 0 & -3 & 1 & 3 \\ 0 & 6 & -2 & -6 \end{array}\right].$$

Looking at the matrix here, we see that row 3 is -2 times row 2. Therefore, if we add two copies of row 2 to row 3, we get the augmented matrix

$$\left[\begin{array}{ccc|c} 1 & -1 & -3 & -3 \\ 0 & -3 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

Therefore, we have a situation where there are only two pivot columns, and the last row is all zeros. Since there are three variables and column 3 is not a pivot column, we can take x_3 as a free variable. If we do that, the second equation tells us that $-3x_2 + x_3 = 3$, or, since we are taking x_3 as a free variable, we can write $x_2 = -1 + 1/3 x_3$. We can then take the first equation, which says that $x_1 - x_2 - 3x_3 = 3$ or, by rearranging

$$x_1 = 3 + x_2 + 3x_3 = 3 + \left(-1 + \frac{1}{3}x_3\right) + 3x_3 = 2 + \frac{10}{3}x_3.$$

This means that for any value of t, our solution is determined by

$$\begin{aligned} x_1 &= 2 + \frac{10}{3}t \\ x_2 &= -1 + \frac{1}{3}t \\ x_3 &= t \end{aligned}$$

The use of t here is just to separate it from the variable x_3. For example, we could pick $t = 3$, in which case we would get $x_1 = 12$, $x_2 = 0$, $x_3 = 3$.

For the second version of the problem, we again set up the augmented matrix

$$\left[\begin{array}{ccc|c} 1 & -1 & -3 & -3 \\ -1 & -2 & 4 & 6 \\ 1 & 5 & -5 & 1 \end{array}\right].$$

Since the left side matrix part is the same as the previous version, the process of row reducing the matrix is identical to what was done previously. When we carry out this process we get the augmented matrix

$$\left[\begin{array}{ccc|c} 1 & -1 & -3 & -3 \\ 0 & -3 & 1 & 3 \\ 0 & 0 & 0 & 10 \end{array}\right].$$

In this case, we see that the last row corresponds to the equation $0 = 10$ so these equations are inconsistent and do not have a solution. ⌟

The point of the above example is to illustrate the fact that whether or not a system is inconsistent or has free variables in the solution depends on the right-hand side of the equation, even if the left-hand side has the same coefficients. We'll see more about why this is in § 3.5.

When $\vec{b} = \vec{0}$, we have a so-called *homogeneous matrix equation*

$$A\vec{x} = \vec{0}.$$

There is no need to write an augmented matrix in this case. As the elementary operations do not do anything to a zero column, it always stays a zero column. Moreover, $A\vec{x} = \vec{0}$ always has at least one solution, namely $\vec{x} = \vec{0}$. Such a system is always consistent. It may have other solutions: If you find any free variables, then you get infinitely many solutions.

The set of solutions of $A\vec{x} = \vec{0}$ comes up quite often so people give it a name. It is called the *nullspace* or the *kernel* of A. One place where the kernel comes up is invertibility of a square matrix A. If the kernel of A contains a nonzero vector, then it contains infinitely many vectors (there was a free variable). But then it is impossible to invert $\vec{0}$, since infinitely many vectors go to $\vec{0}$, so there is no unique vector that A takes to $\vec{0}$. So if the kernel is nontrivial, that is, if there are any nonzero vectors, in other words, if there are any free variables, or in yet other words, if the row echelon form of A has columns without pivots, then A is not invertible. We will return to this idea later.

3.3.4 Exercises

***Exercise* 3.3.1:** *Compute the reduced row echelon form for the following matrices:*

a) $\begin{bmatrix} 1 & 3 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ *b)* $\begin{bmatrix} 3 & 3 \\ 6 & -3 \end{bmatrix}$ *c)* $\begin{bmatrix} 3 & 6 \\ -2 & -3 \end{bmatrix}$ *d)* $\begin{bmatrix} 6 & 6 & 7 & 7 \\ 1 & 1 & 0 & 1 \end{bmatrix}$

e) $\begin{bmatrix} 9 & 3 & 0 & 2 \\ 8 & 6 & 3 & 6 \\ 7 & 9 & 7 & 9 \end{bmatrix}$ *f)* $\begin{bmatrix} 2 & 1 & 3 & -3 \\ 6 & 0 & 0 & -1 \\ -2 & 4 & 4 & 3 \end{bmatrix}$ *g)* $\begin{bmatrix} 6 & 6 & 5 \\ 0 & -2 & 2 \\ 6 & 5 & 6 \end{bmatrix}$ *h)* $\begin{bmatrix} 0 & 2 & 0 & -1 \\ 6 & 6 & -3 & 3 \\ 6 & 2 & -3 & 5 \end{bmatrix}$

Exercise* 3.3.2: *Compute the reduced row echelon form for the following matrices:*

a) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ *b)* $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ *c)* $\begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix}$ *d)* $\begin{bmatrix} 1 & -3 & 1 \\ 4 & 6 & -2 \\ -2 & 6 & -2 \end{bmatrix}$

e) $\begin{bmatrix} 2 & 2 & 5 & 2 \\ 1 & -2 & 4 & -1 \\ 0 & 3 & 1 & -2 \end{bmatrix}$ *f)* $\begin{bmatrix} -2 & 6 & 4 & 3 \\ 6 & 0 & -3 & 0 \\ 4 & 2 & -1 & 1 \end{bmatrix}$ *g)* $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ *h)* $\begin{bmatrix} 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 5 \end{bmatrix}$

***Exercise* 3.3.3:** *Solve (find all solutions), or show no solution exists*

a)
$$\begin{aligned} 4x_1 + 3x_2 &= -2 \\ -x_1 + x_2 &= 4 \end{aligned}$$

b)
$$\begin{aligned} x_1 + 5x_2 + 3x_3 &= 7 \\ 8x_1 + 7x_2 + 8x_3 &= 8 \\ 4x_1 + 8x_2 + 6x_3 &= 4 \end{aligned}$$

c)
$$\begin{aligned} 4x_1 + 8x_2 + 2x_3 &= 3 \\ -x_1 - 2x_2 + 3x_3 &= 1 \\ 4x_1 + 8x_2 &= 2 \end{aligned}$$

d)
$$\begin{aligned} x + 2y + 3z &= 4 \\ 2x - y + 3z &= 1 \\ 3x + y + 6z &= 6 \end{aligned}$$

Exercise* 3.3.4: *Solve (find all solutions), or show no solution exists*

a)
$$\begin{aligned} 4x_1 + 3x_2 &= -1 \\ 5x_1 + 6x_2 &= 4 \end{aligned}$$

b)
$$\begin{aligned} 5x + 6y + 5z &= 7 \\ 6x + 8y + 6z &= -1 \\ 5x + 2y + 5z &= 2 \end{aligned}$$

c)
$$\begin{aligned} a + b + c &= -1 \\ a + 5b + 6c &= -1 \\ -2a + 5b + 6c &= 8 \end{aligned}$$

d)
$$\begin{aligned} -2x_1 + 2x_2 + 8x_3 &= 6 \\ x_2 + x_3 &= 2 \\ x_1 + 4x_2 + x_3 &= 7 \end{aligned}$$

Exercise* 3.3.5: *Solve the system of equations*

$$\begin{aligned} -4x_2 + x_3 + 2x_4 &= 16 \\ 2x_1 + 2x_2 - 4x_3 - 3x_4 &= 1 \\ x_1 + x_2 + 2x_3 + 3x_4 &= 6 \\ 2x_1 - 2x_3 + 4x_4 &= 24 \end{aligned}$$

or determine that no solution exists.

Exercise* 3.3.6: *Solve the system of equations*

$$\begin{aligned} 3x_2 + 3x_3 + 2x_4 &= 4 \\ 4x_1 + 4x_2 + 2x_3 - 4x_4 &= -26 \\ x_1 - 3x_2 - 2x_3 + 2x_4 &= 1 \\ 3x_1 + 3x_2 + 3x_3 - x_4 &= -14 \end{aligned}$$

or determine that no solution exists.

Exercise* 3.3.7*: *Solve the system of equations*

$$\begin{aligned} 2x_1 + x_2 - x_3 + 4x_4 &= 11 \\ x_1 + 4x_2 - 4x_3 - x_4 &= -7 \\ -2x_1 - 3x_2 + 2x_3 + x_4 &= 11 \\ 3x_1 + x_3 + 4x_4 &= 3 \end{aligned}$$

or determine that no solution exists.

Exercise* 3.3.8*: *Solve the system of equations*

$$\begin{aligned} x_1 - x_3 - 4x_4 &= -3 \\ x_1 + x_2 + x_4 &= 0 \\ x_1 + 3x_2 + 3x_3 - 4x_4 &= -28 \\ 6x_1 + 3x_2 - 4x_3 + 6x_4 &= 25 \end{aligned}$$

or determine that no solution exists.

Exercise* 3.3.9*: *Assume that you are solving a three component linear system of equations via row reduction of an augmented matrix and reach the matrix*

$$\left[\begin{array}{ccc|c} 1 & 0 & 3 & 4 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{array}\right].$$

What does this mean about the solution to this system of equations?

Exercise* 3.3.10*: *Assume that you are solving a three component linear system of equations via row reduction of an augmented matrix and reach the matrix*

$$\left[\begin{array}{ccc|c} 1 & 1 & 3 & 6 \\ 0 & 1 & 2 & 4 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

What does this mean about the solution to this system of equations?

***Exercise* 3.3.11:** *Assume that you are solving a four component linear system of equations via row reduction of an augmented matrix and reach the matrix*

$$\left[\begin{array}{cccc|c} 1 & 2 & 3 & 5 & 1 \\ 0 & 2 & 1 & 4 & 2 \\ 0 & 1 & 0 & 3 & 0 \\ 0 & 3 & 2 & -1 & 1 \end{array}\right].$$

What is the next step in reducing this matrix? Carry out the rest of this problem to solve the corresponding system of equations.

Exercise* 3.3.12*: *Assume that someone else has provided you the solution to an augmented matrix reduction for solving a system of equations given below*

$$\begin{bmatrix} 1 & 1 & 2 & 4 \\ 0 & 1 & 3 & 5 \\ 1 & 2 & 4 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 2 & 4 \\ 0 & 1 & 3 & 5 \\ 0 & 1 & 2 & -5 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 2 & 4 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -9 \end{bmatrix}.$$

Is this work correct? If so, what does this say about the solution(s) to the system? If not, correct the work to solve the system.

3.4 Linear independence, rank, and dimension

Attribution: [JL], §A.4.

Learning Objectives

After this section, you will be able to:

- Determine if a set of vectors is linearly independent,
- Compute the rank of a matrix,
- Find a maximal linearly independent subset of a set of vectors, and
- Compute a basis of a subspace and the dimension of that subspace.

3.4.1 Linear independence and rank

As we saw in the § 3.3, it is possible to have a set of equations that is redundant; that is, at least one of the equations does not give us any more information or constraints on the variables. In a lot of cases, this either led to inconsistent systems or free variables. We would like to have a better way to talk about this idea, both in terms of systems of equations and matrices in general. The concept we want is that of linear independence. The same concept is useful for differential equations, for example in Chapter 2.

Definition 3.4.1

Given row or column vectors $\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_n$, a *linear combination* is an expression of the form

$$\alpha_1\vec{y}_1 + \alpha_2\vec{y}_2 + \cdots + \alpha_n\vec{y}_n,$$

where $\alpha_1, \alpha_2, \ldots, \alpha_n$ are all scalars.

For example, $3\vec{y}_1 + \vec{y}_2 - 5\vec{y}_3$ is a linear combination of $\vec{y}_1$, $\vec{y}_2$, and $\vec{y}_3$.
We have seen linear combinations before. The expression

$$A\vec{x}$$

is a linear combination of the columns of A, while

$$\vec{x}^T A = (A^T\vec{x})^T$$

is a linear combination of the rows of A.

The way linear combinations come up in our study of differential equations is similar to the following computation. Suppose that $\vec{x}_1$, $\vec{x}_2$, …, $\vec{x}_n$ are solutions to $A\vec{x}_1 = \vec{0}$, $A\vec{x}_2 = \vec{0}$, …, $A\vec{x}_n = \vec{0}$. Then the linear combination

$$\vec{y} = \alpha_1\vec{x}_1 + \alpha_2\vec{x}_2 + \cdots + \alpha_n\vec{x}_n$$

is a solution to $A\vec{y} = \vec{0}$:

$$A\vec{y} = A(\alpha_1\vec{x}_1 + \alpha_2\vec{x}_2 + \cdots + \alpha_n\vec{x}_n) = \\ = \alpha_1 A\vec{x}_1 + \alpha_2 A\vec{x}_2 + \cdots + \alpha_n A\vec{x}_n = \alpha_1\vec{0} + \alpha_2\vec{0} + \cdots + \alpha_n\vec{0} = \vec{0}.$$

We have seen this computation before in the sense of solutions to homogeneous second order equations. We used $L[y]$ to represent a second order linear differential equation, and showed that if we knew that functions y_1 and y_2 solved

$$L[y_1] = 0 \qquad L[y_2] = 0$$

then $L[c_1y_1 + c_2y_2] = 0$ for any constants c_1 and c_2. We did this by showing that

$$L[c_1y_1 + c_2y_2] = c_1L[y_1] + C_2L[y_2]$$

which mirrors the expression computed above.

Our original question was about when equations are redundant. That is answered by the following definition.

Definition 3.4.2

We say the vectors $\vec{y}_1$, $\vec{y}_2$, ..., $\vec{y}_n$ are *linearly independent* if the only way to pick α_1, $\alpha_2, ..., \alpha_n$ to satisfy

$$\alpha_1\vec{x}_1 + \alpha_2\vec{x}_2 + \cdots + \alpha_n\vec{x}_n = \vec{0}$$

is $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$. Otherwise, we say the vectors are *linearly dependent*.

If the equations (or their coefficients, as we will see later) are linearly dependent, then they are redundant equations, and not all of them are necessary to define the same solution to the equation. If they are linearly independent, then they are all required.

Example 3.4.1: The vectors $\left[\begin{smallmatrix}1\\2\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ are linearly independent.

Solution: Let's try:

$$\alpha_1\begin{bmatrix}1\\2\end{bmatrix} + \alpha_2\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}\alpha_1\\2\alpha_1 + \alpha_2\end{bmatrix} = \vec{0} = \begin{bmatrix}0\\0\end{bmatrix}.$$

So $\alpha_1 = 0$, and then it is clear that $\alpha_2 = 0$ as well. In other words, the vectors are linearly independent. ⌟

If a set of vectors is linearly dependent, that is, we have an expression of the form

$$\alpha_1\vec{x}_1 + \alpha_2\vec{x}_2 + \cdots + \alpha_n\vec{x}_n = 0$$

with some of the α_j's are nonzero, then we can solve for one vector in terms of the others. Suppose $\alpha_1 \neq 0$. Since $\alpha_1\vec{x}_1 + \alpha_2\vec{x}_2 + \cdots + \alpha_n\vec{x}_n = \vec{0}$, then

$$\vec{x}_1 = \frac{-\alpha_2}{\alpha_1}\vec{x}_2 - \frac{-\alpha_3}{\alpha_1}\vec{x}_3 + \cdots + \frac{-\alpha_n}{\alpha_1}\vec{x}_n.$$

For example,

$$2\begin{bmatrix}1\\2\\3\end{bmatrix} - 4\begin{bmatrix}1\\1\\1\end{bmatrix} + 2\begin{bmatrix}1\\0\\-1\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix},$$

and so

$$\begin{bmatrix}1\\2\\3\end{bmatrix} = 2\begin{bmatrix}1\\1\\1\end{bmatrix} - \begin{bmatrix}1\\0\\-1\end{bmatrix}.$$

You may have noticed that solving for those α_j's is just solving linear equations, and so you may not be surprised that to check if a set of vectors is linearly independent we use row reduction.

Given a set of vectors, we may not be interested in just finding if they are linearly independent or not, we may be interested in finding a linearly independent subset. Or perhaps we may want to find some other vectors that give the same linear combinations and are linearly independent. The way to figure this out is to form a matrix out of our vectors. If we have row vectors we consider them as rows of a matrix. If we have column vectors we consider them columns of a matrix.

Definition 3.4.3

Given a matrix A, the maximal number of linearly independent rows is called the *rank* of A, and we write "rank A" for the rank.

For example,

$$\operatorname{rank}\begin{bmatrix}1 & 1 & 1\\2 & 2 & 2\\-1 & -1 & -1\end{bmatrix} = 1.$$

The second and third row are multiples of the first one. We cannot choose more than one row and still have a linearly independent set. But what is

$$\operatorname{rank}\begin{bmatrix}1 & 2 & 3\\4 & 5 & 6\\7 & 8 & 9\end{bmatrix} \quad = \quad ?$$

That seems to be a tougher question to answer. The first two rows are linearly independent, so the rank is at least two. If we would set up the equations for the α_1, α_2, and α_3, we would find a system with infinitely many solutions. One solution is

$$\begin{bmatrix}1 & 2 & 3\end{bmatrix} - 2\begin{bmatrix}4 & 5 & 6\end{bmatrix} + \begin{bmatrix}7 & 8 & 9\end{bmatrix} = \begin{bmatrix}0 & 0 & 0\end{bmatrix}.$$

So the set of all three rows is linearly dependent, the rank cannot be 3. Therefore the rank is 2.

But how can we do this in a more systematic way? We find the row echelon form!

$$\text{Row echelon form of} \quad \begin{bmatrix}1 & 2 & 3\\4 & 5 & 6\\7 & 8 & 9\end{bmatrix} \quad \text{is} \quad \begin{bmatrix}1 & 2 & 3\\0 & 1 & 2\\0 & 0 & 0\end{bmatrix}.$$

The elementary row operations do not change the set of linear combinations of the rows (that was one of the main reasons for defining them as they were). In other words, the span of the rows of the A is the same as the span of the rows of the row echelon form of A. In particular, the number of linearly independent rows is the same. And in the row echelon form, all nonzero rows are linearly independent. This is not hard to see. Consider the two nonzero rows in the example above. Suppose we tried to solve for the α_1 and α_2 in

$$\alpha_1 \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} + \alpha_2 \begin{bmatrix} 0 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}.$$

Since the first column of the row echelon matrix has zeros except in the first row means that $\alpha_1 = 0$. For the same reason, α_2 is zero. We only have two nonzero rows, and they are linearly independent, so the rank of the matrix is 2. This also tells us that if we were trying to solve the system of equations

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= a \\ 4x_1 + 5x_2 + 6x_3 &= b \\ 7x_1 + 8x_2 + 9x_3 &= c \end{aligned}$$

we would get that one row of the reduced augmented matrix has all zeros on the left side, and so this system either has a free variable or is inconsistent, because only two equations here are relevant. We will see more examples of the rank of a matrix once we have more terminology to talk about it.

3.4.2 Subspaces and span

Now, let's consider a different scenario. Assume that we find two vectors that solve $A\vec{x} = 0$. What other vectors also solve this equation? In our discussion of linear combinations, we saw that if $\vec{x}_1$ and $\vec{x}_2$ solve $A\vec{x} = 0$, then so does $A(\alpha_1\vec{x}_1 + \alpha_2\vec{x}_2)$ for any constants α_1 and α_2. Thus, all linear combinations will also solve the equation. This leads to the definition of the span of a set of vectors.

Definition 3.4.4

The set of all linear combinations of a set of vectors is called their *span*.

$$\text{span}\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\} = \{\text{Set of all linear combinations of } \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}.$$

Thus, if two vectors solve a homogeneous equation, so does everything in the span of those two vectors. The span of a collection of vectors is an example of a subspace, which is a common object in linear algebra. We say that a set S of vectors in $\mathbb{R}^n$ is a *subspace* if whenever $\vec{x}$ and $\vec{y}$ are members of S and α is a scalar, then

$$\vec{x} + \vec{y}, \qquad \text{and} \qquad \alpha\vec{x}$$

are also members of S. That is, we can add and multiply by scalars and we still land in S. So every linear combination of vectors of S is still in S. That is really what a subspace is. It is a subset where we can take linear combinations and still end up being in the subset.

Example 3.4.2: If we let $S = \mathbb{R}^n$, then this S is a subspace of $\mathbb{R}^n$. Adding any two vectors in $\mathbb{R}^n$ gets a vector in $\mathbb{R}^n$, and so does multiplying by scalars.

The set $S' = \{\vec{0}\}$, that is, the set of the zero vector by itself, is also a subspace of $\mathbb{R}^n$. There is only one vector in this subspace, so we only need to check for that one vector, and everything checks out: $\vec{0} + \vec{0} = \vec{0}$ and $\alpha\vec{0} = \vec{0}$.

The set S'' of all the vectors of the form (a, a) for any real number a, such as $(1, 1)$, $(3, 3)$, or $(-0.5, -0.5)$ is a subspace of $\mathbb{R}^2$. Adding two such vectors, say $(1, 1) + (3, 3) = (4, 4)$ again gets a vector of the same form, and so does multiplying by a scalar, say $8(1, 1) = (8, 8)$.

We can apply these ideas to the vectors that live inside a matrix. The span of the rows of a matrix A is called the *row space*. The row space of A and the row space of the row echelon form of A are the same, because reducing the matrix A to its row echelon form involves taking linear combinations, which will preserve the span. In the example,

$$\text{row space of } \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \text{span}\left\{ \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}, \begin{bmatrix} 4 & 5 & 6 \end{bmatrix}, \begin{bmatrix} 7 & 8 & 9 \end{bmatrix} \right\}$$
$$= \text{span}\left\{ \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 2 \end{bmatrix} \right\}.$$

Similarly to row space, the span of columns is called the *column space*.

$$\text{column space of } \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \text{span}\left\{ \begin{bmatrix} 1 \\ 4 \\ 7 \end{bmatrix}, \begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix}, \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix} \right\}.$$

So it may also be good to find the number of linearly independent columns of A. One way to do that is to find the number of linearly independent rows of A^T. It is a tremendously useful fact that the number of linearly independent columns is always the same as the number of linearly independent rows:

Theorem 3.4.1

$\operatorname{rank} A = \operatorname{rank} A^T$

In particular, to find a set of linearly independent columns we need to look at where the pivots were. If you recall above, when solving $A\vec{x} = \vec{0}$ the key was finding the pivots, any non-pivot columns corresponded to free variables. That means we can solve for the non-pivot columns in terms of the pivot columns. Let's see an example.

Example 3.4.3: Find the linearly independent columns of the matrix

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix}.$$

Solution: We find a pivot and reduce the rows below:

$$\begin{bmatrix} \boxed{1} & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix} \to \begin{bmatrix} \boxed{1} & 2 & 3 & 4 \\ 0 & 0 & -1 & -2 \\ 3 & 6 & 7 & 8 \end{bmatrix} \to \begin{bmatrix} \boxed{1} & 2 & 3 & 4 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & -2 & -4 \end{bmatrix}.$$

We find the next pivot, make it one, and rinse and repeat:

$$\begin{bmatrix} \boxed{1} & 2 & 3 & 4 \\ 0 & 0 & \boxed{-1} & -2 \\ 0 & 0 & -2 & -4 \end{bmatrix} \to \begin{bmatrix} \boxed{1} & 2 & 3 & 4 \\ 0 & 0 & \boxed{1} & 2 \\ 0 & 0 & -2 & -4 \end{bmatrix} \to \begin{bmatrix} \boxed{1} & 2 & 3 & 4 \\ 0 & 0 & \boxed{1} & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The final matrix is the row echelon form of the matrix. Consider the pivots that we marked. The pivot columns are the first and the third column. All other columns correspond to free variables when solving $A\vec{x} = \vec{0}$, so all other columns can be solved in terms of the first and the third column. In other words

$$\text{column space of } \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix} = \text{span} \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}, \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}, \begin{bmatrix} 4 \\ 6 \\ 8 \end{bmatrix} \right\} = \text{span} \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix} \right\}.$$

We could perhaps use another pair of columns to get the same span, but the first and the third are guaranteed to work because they are pivot columns. The discussion above could be expanded into a proof of the theorem if we wanted. As each nonzero row in the row echelon form contains a pivot, then the rank is the number of pivots, which is the same as the maximal number of linearly independent columns.

In the previous example, this means that only the first and third colums are "important" in the sense of generating the full column space as a span. We would like to have a way to talk about what these first and third columns do.

Definition 3.4.5

Let S be a subspace of a vector space. The set $\{\vec{v}_1, \vec{v}_2, ..., \vec{v}_k\}$ is a *spanning set* for the subspace S if each of these vectors are in S and the span of $\{\vec{v}_1, \vec{v}_2, ..., \vec{v}_k\}$ is equal to S.

In the context of the previous example, for the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix}$$

we know that

$$\text{column space of } \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix} = \text{span} \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}, \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}, \begin{bmatrix} 4 \\ 6 \\ 8 \end{bmatrix} \right\} = \text{span} \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix} \right\}.$$

This means that both

$$\left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}, \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}, \begin{bmatrix} 4 \\ 6 \\ 8 \end{bmatrix} \right\} \quad \text{and} \quad \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix} \right\}$$

are spanning sets for this column space.

The idea also works in reverse. Suppose we have a bunch of column vectors and we just need to find a linearly independent set. For example, suppose we started with the vectors

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \vec{v}_2 = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}, \quad \vec{v}_3 = \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}, \quad \vec{v}_4 = \begin{bmatrix} 4 \\ 6 \\ 8 \end{bmatrix}.$$

These vectors are not linearly independent as we saw above. In particular, the span $\vec{v}_1$ and $\vec{v}_3$ is the same as the span of all four of the vectors. So $\vec{v}_2$ and $\vec{v}_4$ can both be written as linear combinations of $\vec{v}_1$ and $\vec{v}_3$. A common thing that comes up in practice is that one gets a set of vectors whose span is the set of solutions of some problem. But perhaps we get way too many vectors, we want to simplify. For example above, all vectors in the span of $\vec{v}_1, \vec{v}_2, \vec{v}_3, \vec{v}_4$ can be written $\alpha_1\vec{v}_1 + \alpha_2\vec{v}_2 + \alpha_3\vec{v}_3 + \alpha_4\vec{v}_4$ for some numbers $\alpha_1, \alpha_2, \alpha_3, \alpha_4$. But it is also true that every such vector can be written as $a\vec{v}_1 + b\vec{v}_3$ for two numbers a and b. And one has to admit, that looks much simpler. Moreover, these numbers a and b are unique. More on that later in this section.

To find this linearly independent set we simply take our vectors and form the matrix $[\vec{v}_1\ \vec{v}_2\ \vec{v}_3\ \vec{v}_4]$, that is, the matrix

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix}.$$

We crank up the row-reduction machine, feed this matrix into it, and find the pivot columns and pick those. In this case, $\vec{v}_1$ and $\vec{v}_3$.

3.4.3 Basis and dimension

At this point, we have talked about subspaces, and two other properties of sets of vectors: linear independence and being a spanning set for a subspace. In some sense, these two properties are in opposition to each other. If I add more vectors to a set, I am more likely to become a spanning set (because I have more options for adding to get other vectors), but less likely to be independent (because there are more possibilities for a linear combination to be zero). Similarly, the reverse is true; removing vectors means the set is more likely to be linearly independent, but less likely to span a given subspace. The question then becomes if there is a sweet spot where both things are true, and that leads to the definition of a basis.

Definition 3.4.6

If S is a subspace and we can find k linearly independent vectors in S

$$\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k,$$

such that every other vector in S is a linear combination of $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k$, then the set $\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k\}$ is called a *basis* of S. In other words, S is the span of $\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k\}$. We say that S has *dimension* k, and we write

$$\dim S = k.$$

The next theorem illustrates the main properties and classification of a basis of a vector space.

Theorem 3.4.2

If $S \subset \mathbb{R}^n$ is a subspace and S is not the trivial subspace $\{\vec{0}\}$, then there exists a unique positive integer k (the dimension) and a (not unique) basis $\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k\}$, such that every $\vec{w}$ in S can be uniquely represented by

$$\vec{w} = \alpha_1\vec{v}_1 + \alpha_2\vec{v}_2 + \cdots + \alpha_k\vec{v}_k,$$

for some scalars α_1, α_2, $\ldots$, α_k.

We should reiterate that while k is unique (a subspace cannot have two different dimensions), the set of basis vectors is not at all unique. There are lots of different bases for any given subspace. Finding just the right basis for a subspace is a large part of what one does in linear algebra. In fact, that is what we spend a lot of time on in linear differential equations, although at first glance it may not seem like that is what we are doing.

Example 3.4.4: The standard basis

$$\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n,$$

is a basis of $\mathbb{R}^n$ (hence the name). So as expected

$$\dim \mathbb{R}^n = n.$$

On the other hand the subspace $\{\vec{0}\}$ is of dimension 0.

The subspace S'' from a previous example, that is, the set of vectors (a, a) is of dimension 1. One possible basis is simply $\{(1, 1)\}$, the single vector $(1, 1)$: every vector in S'' can be represented by $a(1, 1) = (a, a)$. Similarly another possible basis would be $\{(-1, -1)\}$. Then the vector (a, a) would be represented as $(-a)(-1, -1)$. In this case, the subspace S'' has many different bases, two of which are $\{(1, 1)\}$ and $\{(-1, -1)\}$, and the vector (a, a) has a different representation (different constant) for the different bases.

Row and column spaces of a matrix are also examples of subspaces, as they are given as the span of vectors. We can use what we know about rank, row spaces, and column spaces from the previous section to find a basis.

Example 3.4.5: Earlier, we considered the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix}.$$

Using row reduction to find the pivot columns, we found

$$\text{column space of } A \left(\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix} \right) = \text{span} \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix} \right\}.$$

What we did was we found the basis of the column space. The basis has two elements, and so the column space of A is two dimensional. Notice that the rank of A is two.

We would have followed the same procedure if we wanted to find the basis of the subspace X spanned by

$$\begin{bmatrix}1\\2\\3\end{bmatrix}, \begin{bmatrix}2\\4\\6\end{bmatrix}, \begin{bmatrix}3\\5\\7\end{bmatrix}, \begin{bmatrix}4\\6\\8\end{bmatrix}.$$

We would have simply formed the matrix A with these vectors as columns and repeated the computation above. The subspace X is then the column space of A.

Example 3.4.6: Consider the matrix

$$L = \begin{bmatrix}1 & 2 & 0 & 0 & 3\\0 & 0 & 1 & 0 & 4\\0 & 0 & 0 & 1 & 5\end{bmatrix}$$

Conveniently, the matrix is in reduced row echelon form. The matrix is of rank 3. The column space is the span of the pivot columns, because the pivot columns always form a basis for the column space of a matrix. It is the 3-dimensional space

$$\text{column space of } L = \text{span}\left\{\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}\right\} = \mathbb{R}^3.$$

The row space is the 3-dimensional space

$$\text{row space of } L = \text{span}\left\{\begin{bmatrix}1 & 2 & 0 & 0 & 3\end{bmatrix}, \begin{bmatrix}0 & 0 & 1 & 0 & 4\end{bmatrix}, \begin{bmatrix}0 & 0 & 0 & 1 & 5\end{bmatrix}\right\}.$$

As these vectors have 5 components, we think of the row space of L as a subspace of $\mathbb{R}^5$.

The way the dimensions worked out in the examples is not an accident. Since the number of vectors that we needed to take was always the same as the number of pivots, and the number of pivots is the rank, we get the following result.

Theorem 3.4.3 (Rank)

The dimension of the column space and the dimension of the row space of a matrix A are both equal to the rank of A.

3.4.4 Exercises

***Exercise* 3.4.1:** *Compute the rank of the given matrices*

a) $\begin{bmatrix}6 & 3 & 5\\1 & 4 & 1\\7 & 7 & 6\end{bmatrix}$ *b)* $\begin{bmatrix}5 & -2 & -1\\3 & 0 & 6\\2 & 4 & 5\end{bmatrix}$ *c)* $\begin{bmatrix}1 & 2 & 3\\-1 & -2 & -3\\2 & 4 & 6\end{bmatrix}$

Exercise* 3.4.2: *Compute the rank of the given matrices*

a) $\begin{bmatrix} 7 & -1 & 6 \\ 7 & 7 & 7 \\ 7 & 6 & 2 \end{bmatrix}$ b) $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 2 & 2 & 2 \end{bmatrix}$ c) $\begin{bmatrix} 0 & 3 & -1 \\ 6 & 3 & 1 \\ 4 & 7 & -1 \end{bmatrix}$

Exercise 3.4.3: *For the matrices in Exercise 3.4.1, find a linearly independent set of row vectors that span the row space (they don't need to be rows of the matrix).*

Exercise 3.4.4: *For the matrices in Exercise 3.4.1, find a linearly independent set of columns that span the column space. That is, find the pivot columns of the matrices.*

Exercise 3.4.5:* *For the matrices in Exercise 3.4.2, find a linearly independent set of row vectors that span the row space (they don't need to be rows of the matrix).*

Exercise 3.4.6:* *For the matrices in Exercise 3.4.2, find a linearly independent set of columns that span the column space. That is, find the pivot columns of the matrices.*

Exercise 3.4.7: *Compute the rank of the matrix*

$$\begin{bmatrix} 10 & -2 & 11 & -7 \\ -5 & -2 & -5 & 5 \\ 1 & 0 & -4 & -4 \\ 1 & 2 & 2 & -1 \end{bmatrix}$$

Exercise 3.4.8: *Compute the rank of the matrix*

$$\begin{bmatrix} 4 & -2 & 0 & -4 \\ 3 & -5 & 2 & 0 \\ 1 & -2 & 0 & 1 \\ -1 & 1 & 3 & -3 \end{bmatrix}$$

Exercise 3.4.9: *Find a linearly independent subset of the following vectors that has the same span.*

$$\begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ -2 \\ -4 \end{bmatrix}, \quad \begin{bmatrix} -2 \\ 4 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} -1 \\ 3 \\ -2 \end{bmatrix}$$

Exercise 3.4.10:* *Find a linearly independent subset of the following vectors that has the same span.*

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ 1 \\ -5 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 3 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} -3 \\ 2 \\ 4 \end{bmatrix}$$

Exercise 3.4.11: *For the following sets of vectors, determine if the set is linearly independent. Then find a basis for the subspace spanned by the vectors, and find the dimension of the subspace.*

a) $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix}$ b) $\begin{bmatrix} 1 \\ 0 \\ 5 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}$ c) $\begin{bmatrix} -4 \\ -3 \\ 5 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix}$

d) $\begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ 2 \end{bmatrix}$ e) $\begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \end{bmatrix}$ f) $\begin{bmatrix} 3 \\ 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \\ -4 \end{bmatrix}, \begin{bmatrix} -5 \\ -5 \\ -2 \end{bmatrix}$

Exercise* 3.4.12: *For the following sets of vectors, determine if the set is linearly independent. Then find a basis for the subspace spanned by the vectors, and find the dimension of the subspace.*

a) $\begin{bmatrix}1\\2\end{bmatrix}, \quad \begin{bmatrix}1\\1\end{bmatrix}$

b) $\begin{bmatrix}1\\1\\1\end{bmatrix}, \quad \begin{bmatrix}2\\2\\2\end{bmatrix}, \quad \begin{bmatrix}1\\1\\2\end{bmatrix}$

c) $\begin{bmatrix}5\\3\\1\end{bmatrix}, \quad \begin{bmatrix}5\\-1\\5\end{bmatrix}, \quad \begin{bmatrix}-1\\3\\-4\end{bmatrix}$

d) $\begin{bmatrix}2\\2\\4\end{bmatrix}, \quad \begin{bmatrix}2\\2\\3\end{bmatrix}, \quad \begin{bmatrix}4\\4\\-3\end{bmatrix}$

e) $\begin{bmatrix}1\\0\end{bmatrix}, \quad \begin{bmatrix}2\\0\end{bmatrix}, \quad \begin{bmatrix}3\\0\end{bmatrix}$

f) $\begin{bmatrix}1\\0\\0\end{bmatrix}, \quad \begin{bmatrix}2\\0\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\\2\end{bmatrix}$

***Exercise* 3.4.13:** *Suppose that X is the set of all the vectors of $\mathbb{R}^3$ whose third component is zero. Is X a subspace? And if so, find a basis and the dimension.*

Exercise* 3.4.14: *Consider a set of 3 component vectors.*

a) How can it be shown if these vectors are linearly independent?

b) Can a set of 4 of these 3 component vectors be linearly independent? Explain your answer.

c) Can a set of 2 of these 3 component vectors be linearly independent? Explain.

d) How would it be shown if these vectors make up a spanning set for all 3 component vectors?

e) Can 4 vectors be a spanning set? Explain.

f) Can 2 vectors be a spanning set? Explain.

Exercise* 3.4.15: *Consider the vectors*

$$\vec{v}_1 = \begin{bmatrix}4\\2\\-1\end{bmatrix} \quad \vec{v}_2 = \begin{bmatrix}3\\5\\1\end{bmatrix} \qquad \begin{bmatrix}1\\-1\\-1\end{bmatrix}.$$

Let A be the matrix with these vectors as columns and $\vec{b}$ the vector $[1\ 0\ 0]$.

a) Compute the rank of A to determine how many of these vectors are linearly independent.

b) Determine if $\vec{b}$ is in the span of the given vectors by using row reduction to try to solve $A\vec{x} = \vec{b}$.

c) Look at the columns of the row-reduced form of A. Is $\vec{b}$ in the span of those vectors?

d) What do these last two parts tell you about the span of the columns of a matrix, and the span of the columns of the row-reduced matrix?

e) Now, build a matrix D with these vectors as rows. Row-reduce this matrix to get a matrix D_2.

f) Is $\vec{b}$ in the span of the rows of D_2? You can't check this in using the matrix form; instead, just brute force it based on the form of D_2. What does this potentially say about the span of the rows of D and the rows of D_2?

***Exercise* 3.4.16:** *Complete Exercise 3.4.15 with*

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix} \quad \vec{v}_2 = \begin{bmatrix} -6 \\ 2 \\ 3 \\ -1 \end{bmatrix} \quad \begin{bmatrix} -13 \\ 3 \\ 1 \\ 1 \end{bmatrix} \quad \vec{v}_4 \begin{bmatrix} 11 & -1 \\ -5 \\ -1 \end{bmatrix} \quad \vec{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

3.5 Determinant

Attribution: [JL], §A.6.

Learning Objectives

After this section, you will be able to:

- Compute the determinant of a 2×2 matrix,
- Use cofactor expansion to compute the determinant of larger matrices, and
- Use the determinant to make statements about invertibility or rank of a matrix, and linear independence of the columns of that matrix.

For square matrices we define a useful quantity called the *determinant*. We define the determinant of a 1×1 matrix as the value of its only entry

$$\det\left(\begin{bmatrix} a \end{bmatrix}\right) \stackrel{\text{def}}{=} a.$$

For a 2×2 matrix we define

$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) \stackrel{\text{def}}{=} ad - bc.$$

Before defining the determinant for larger matrices, we note the meaning of the determinant. An $n \times n$ matrix gives a mapping of the n-dimensional euclidean space $\mathbb{R}^n$ to itself. In particular, a 2×2 matrix A is a mapping of the plane to itself. The determinant of A is the factor by which the area of objects changes. If we take the unit square (square of side 1) in the plane, then A takes the square to a parallelogram of area $|\det(A)|$. The sign of $\det(A)$ denotes a change of orientation (negative if the axes get flipped). For example, let

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

Then $\det(A) = 1 + 1 = 2$. Let us see where A sends the unit square with vertices $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$. The point $(0,0)$ gets sent to $(0,0)$.

$$\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

The image of the square is another square with vertices $(0,0)$, $(1,-1)$, $(1,1)$, and $(2,0)$. The image square has a side of length $\sqrt{2}$ and is therefore of area 2. See Figure 3.5 on the following page.

In general the image of a square is going to be a parallelogram. In high school geometry, you may have seen a formula for computing the area of a parallelogram with vertices $(0,0)$, (a,c), (b,d) and $(a+b, c+d)$. The area is

$$\left|\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right)\right| = |ad - bc|.$$

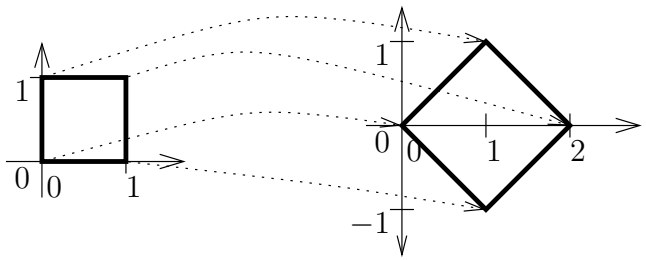


Figure 3.5: Image of the unit quare via the mapping A.

The vertical lines above mean absolute value. The matrix $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ carries the unit square to the given parallelogram.

There are a number of ways to define the determinant for an $n \times n$ matrix. Let us use the so-called *cofactor expansion*. We define A_{ij} as the matrix A with the i^{th} row and the j^{th} column deleted. For example,

$$\text{If} \qquad A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \qquad \text{then} \qquad A_{12} = \begin{bmatrix} 4 & 6 \\ 7 & 9 \end{bmatrix} \qquad \text{and} \qquad A_{23} = \begin{bmatrix} 1 & 2 \\ 7 & 8 \end{bmatrix}.$$

We now define the determinant recursively

$$\det(A) \stackrel{\text{def}}{=} \sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det(A_{1j}),$$

or in other words

$$\det(A) = a_{11}\det(A_{11}) - a_{12}\det(A_{12}) + a_{13}\det(A_{13}) - \cdots \begin{cases} +a_{1n}\det(A_{1n}) & \text{if } n \text{ is odd,} \\ -a_{1n}\det(A_{1n}) & \text{if } n \text{ even.} \end{cases}$$

For a 3×3 matrix, we get $\det(A) = a_{11}\det(A_{11}) - a_{12}\det(A_{12}) + a_{13}\det(A_{13})$. For example,

$$\begin{aligned} \det\left(\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}\right) &= 1 \cdot \det\left(\begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix}\right) - 2 \cdot \det\left(\begin{bmatrix} 4 & 6 \\ 7 & 9 \end{bmatrix}\right) + 3 \cdot \det\left(\begin{bmatrix} 4 & 5 \\ 7 & 8 \end{bmatrix}\right) \\ &= 1(5 \cdot 9 - 6 \cdot 8) - 2(4 \cdot 9 - 6 \cdot 7) + 3(4 \cdot 8 - 5 \cdot 7) = 0. \end{aligned}$$

It turns out that we did not have to necessarily use the first row. That is for any i,

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij}).$$

It is sometimes useful to use a row other than the first. In the following example it is more convenient to expand along the second row. Notice that for the second row we are starting

with a negative sign.

$$\det\left(\begin{bmatrix}1 & 2 & 3\\ 0 & 5 & 0\\ 7 & 8 & 9\end{bmatrix}\right) = -0\cdot\det\left(\begin{bmatrix}2 & 3\\ 8 & 9\end{bmatrix}\right) + 5\cdot\det\left(\begin{bmatrix}1 & 3\\ 7 & 9\end{bmatrix}\right) - 0\cdot\det\left(\begin{bmatrix}1 & 2\\ 7 & 8\end{bmatrix}\right)$$
$$= 0 + 5(1\cdot 9 - 3\cdot 7) + 0 = -60.$$

Let us check if it is really the same as expanding along the first row,

$$\det\left(\begin{bmatrix}1 & 2 & 3\\ 0 & 5 & 0\\ 7 & 8 & 9\end{bmatrix}\right) = 1\cdot\det\left(\begin{bmatrix}5 & 0\\ 8 & 9\end{bmatrix}\right) - 2\cdot\det\left(\begin{bmatrix}0 & 0\\ 7 & 9\end{bmatrix}\right) + 3\cdot\det\left(\begin{bmatrix}0 & 5\\ 7 & 8\end{bmatrix}\right)$$
$$= 1(5\cdot 9 - 0\cdot 8) - 2(0\cdot 9 - 0\cdot 7) + 3(0\cdot 8 - 5\cdot 7) = -60.$$

In computing the determinant, we alternately add and subtract the determinants of the submatrices A_{ij} multiplied by a_{ij} for a fixed i and all j. The numbers $(-1)^{i+j}\det(A_{ij})$ are called *cofactors* of the matrix. And that is why this method of computing the determinant is called the *cofactor expansion*.

Similarly we do not need to expand along a row, we can expand along a column. For any j

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij}\det(A_{ij}).$$

A related fact is that

$$\det(A) = \det(A^T).$$

Recall that a matrix is *upper triangular* if all elements below the main diagonal are 0. For example,

$$\begin{bmatrix}1 & 2 & 3\\ 0 & 5 & 6\\ 0 & 0 & 9\end{bmatrix}$$

is upper triangular. Similarly a *lower triangular* matrix is one where everything above the diagonal is zero. For example,

$$\begin{bmatrix}1 & 0 & 0\\ 4 & 5 & 0\\ 7 & 8 & 9\end{bmatrix}.$$

The determinant for triangular matrices is very simple to compute. Consider the lower triangular matrix. If we expand along the first row, we find that the determinant is 1 times the determinant of the lower triangular matrix $\left[\begin{smallmatrix}5 & 0\\ 8 & 9\end{smallmatrix}\right]$. So the deteriminant is just the product of the diagonal entries:

$$\det\left(\begin{bmatrix}1 & 0 & 0\\ 4 & 5 & 0\\ 7 & 8 & 9\end{bmatrix}\right) = 1\cdot 5\cdot 9 = 45.$$

Similarly for upper triangular matrices

$$\det\left(\begin{bmatrix}1 & 2 & 3\\ 0 & 5 & 6\\ 0 & 0 & 9\end{bmatrix}\right) = 1 \cdot 5 \cdot 9 = 45.$$

In general, if A is triangular, then

$$\det(A) = a_{11}a_{22}\cdots a_{nn}.$$

If A is diagonal, then it is also triangular (upper and lower), so same formula applies. For example,

$$\det\left(\begin{bmatrix}2 & 0 & 0\\ 0 & 3 & 0\\ 0 & 0 & 5\end{bmatrix}\right) = 2 \cdot 3 \cdot 5 = 30.$$

In particular, the identity matrix I is diagonal, and the diagonal entries are all 1. Thus,

$$\det(I) = 1.$$

Another way that we can compute determinants is by using row reduction. Since the row echelon form is a diagonal matrix, this will make it easy to compute the determinant using the product of the diagonal entries. However, we need to know how the determinant is affected by elementary row operations.

Theorem 3.5.1 (Properties of the Determinant)

Let A be a square $n \times n$ matrix.

1. If B obtained from A by interchanging two rows (or two columns) of A, then $\det(B) = -\det(A)$.
2. If B is obtained from A by multiplying a row of column by the number r, then $\det(B) = r\det(A)$.
3. If B is obtained from A by multiplying a row (or column) by a non-zero number r and adding the result to another row, then $\det(B) = \det(A)$.

Proof. The proof of each of these facts comes from the cofactor expansion of the determinant.

1. Assume that B is obtained by interchanging the first and second row of A. We will use cofactor expansion along the first row to find the determinant of A, and the second row for the determinant of B. We get that

$$\det(A) = \sum_{j=1}^{n}(-1)^{1+j}a_{1j}\det(A_{1j})$$

and

$$\det(B) = \sum_{j=1}^{n}(-1)^{2+j}b_{2j}\det(B_{2j}).$$

However, since the second row of B is the first row of A, we know that $b_{2j} = a_{1j}$ for all j. In addition, this swap means that we also have that $B_{2j} = A_{1j}$ for each of the cofactors in this expansion. All of these cofactor matrices are made up of the second through last rows of A, with the appropriate columns removed at each step.

Therefore, the only difference between these two formulas is that the A formula starts with $(-1)^{1+j}$ and the B formula starts with $(-1)^{2+j}$. Thus, $\det(B)$ will have an additional factor of -1 in it, giving the desired result.

The exact same process works for swapping any two adjacent rows of the matrix, giving that this also provides a -1 in the computation of the determinant. For non-adjacent rows, we use the fact that to any swap of non-adjacent rows of a matrix requires an *odd* number of adjacent row swaps. For example, if we want to swap rows 1 and 3, we can swap row 1 with row 2, then row 2 with row 3, and finally swap row 1 with row 2 again. This will put the first row in the third spot and the third row up in the first slot. Since each of these adjacent switches adds a minus sign, doing an odd number of switches still results in adding a single minus sign to the computation of the determinant.

2. Assume that we want to multiply the kth row of A by the number r to get B. We use cofactor expansion along this same kth row to find the determinant of each matrix. We get that

$$\det(A) = \sum_{j=1}^{n} (-1)^{k+j} a_{kj} \det(A_{kj})$$

and

$$\det(B) = \sum_{j=1}^{n} (-1)^{k+j} b_{kj} \det(B_{kj}) = \sum_{j=1}^{n} (-1)^{k+j} r a_{kj} \det(B_{kj}).$$

However, the minor B_{kj} ignores the kth row of the matrix B, so the minors are identical to those of A. Thus, we have that

$$\det(B) = \sum_{j=1}^{n} (-1)^{k+j} r a_{kj} \det(B_{kj}) = r \sum_{j=1}^{n} (-1)^{k+j} a_{kj} \det(A_{kj}) = r \det(A).$$

3. Assume that B is formed by adding r copies of the kth row of A to the ith row. Since the ith row is the one being changed, we will use cofactor expansion there to compute each determinant. We get that

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij}$$

and

$$\det(B) = \sum_{j=1}^{n} (-1)^{i+j} b_{ij} \det(B_{ij}) = \sum (-1)^{i+j} (a_{ij} + r a_{kj}) \det(A_{ij})$$

where we have replaced the minors of B by the minors of A because they ignore the ith row, which is the only thing that has changed. We can now split the determinant

of B into two parts

$$\sum(-1)^{i+j}(a_{ij}+ra_{kj})\det(A_{ij}) = \sum(-1)^{i+j}a_{ij}\det(A_{ij}) + \sum(-1)^{i+j}ra_{kj}\det(A_{ij}).$$

The first of these is the determinant of the matrix A. The second is the determinant of a new matrix that we will call C. C is the same as the matrix A, except that we have replaced the ith row of A by r times the kth row of A. Thus, the ith row of this matrix C is a multiple of the kth row. This means that the rows of C are not linearly independent. By Theorem 3.5.4 coming up later (don't worry, it does not depend on this result), this tells us that the determinant of C is zero. Therefore

$$\det(B) = \det(A) + \det(C) = \det(A)$$

so this operation does not change the determinant of the matrix.

□

These correspond to the three elementary row operations that we use to row reduce matrices. In order to use this to compute determinants, we need to keep track of each of these operations and how the determinant changes at each step.

Example 3.5.1: Compute the determinant of the matrix

$$\begin{bmatrix} -4 & -2 & 4 \\ -3 & -3 & 2 \\ -2 & -3 & 1 \end{bmatrix}$$

using row reduction.

Solution: We will go through the process of row reduction to find the determinant. We need to keep track of each time that we swap rows (to add a minus sign) and that we multiply a row by a constant (to factor in that constant). Throughout this process, we will use A to refer to the initial matrix

$$A = \begin{bmatrix} -4 & -2 & 4 \\ -3 & -3 & 2 \\ -2 & -3 & 1 \end{bmatrix}$$

and M will refer to wherever we are in the process. So we will start by dividing the first row of the matrix by -4

$$\begin{bmatrix} -4 & -2 & 4 \\ -3 & -3 & 2 \\ -2 & -3 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 1/2 & -1 \\ -3 & -3 & 2 \\ -2 & -3 & 1 \end{bmatrix}.$$

Since we divided by -4, Theorem 3.5.1 tells us that

$$\det(M) = -\frac{1}{4}\det(A).$$

The next step of row reduction will be to use the 1 in the top left to cancel out the -3 and -2 below it. Part (c) in Theorem 3.5.1 says that this doesn't change the determinant. Therefore, the row reduction gives

$$\begin{bmatrix} 1 & 1/2 & -1 \\ -3 & -3 & 2 \\ -2 & -3 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 1/2 & -1 \\ 0 & -3/2 & -1 \\ 0 & -2 & -1 \end{bmatrix}$$

and we still have that

$$\det(M) = -\frac{1}{4}\det(A).$$

Next, we will multiply row 2 by $-\frac{2}{3}$, which gives

$$\begin{bmatrix} 1 & 1/2 & -1 \\ 0 & -3/2 & -1 \\ 0 & -2 & -1 \end{bmatrix} \to \begin{bmatrix} 1 & 1/2 & -1 \\ 0 & 1 & 2/3 \\ 0 & -2 & -1 \end{bmatrix}.$$

Adding this in to our previous steps using Theorem 3.5.1, we get that

$$\det(M) = \left(-\frac{2}{3}\right)\left(-\frac{1}{4}\right)\det(A).$$

Finally, we add two copies of row 2 to row 3, which does not change the determinant and gives the matrix

$$\begin{bmatrix} 1 & 1/2 & -1 \\ 0 & 1 & 2/3 \\ 0 & -2 & -1 \end{bmatrix} \to \begin{bmatrix} 1 & 1/2 & -1 \\ 0 & 1 & 2/3 \\ 0 & 0 & 1/3 \end{bmatrix}$$

with

$$\det(M) = \left(-\frac{2}{3}\right)\left(-\frac{1}{4}\right)\det(A).$$

We can rearrange this expression to say that

$$\det(A) = 6\det(M)$$

and we can easily compute that $\det(M) = \frac{1}{3}$ by multiplying the diagonal entries. Thus, we have that $\det(A) = 2$. ⌟

***Exercise* 3.5.1:** *Compute* $\det(A)$ *using cofactor expansion and show that you get the same answer.*

The determinant is telling you how geometric objects scale. If B doubles the sizes of geometric objects and A triples them, then AB (which applies B to an object and then it applies A) should make size go up by a factor of 6. This is true in general:

Theorem 3.5.2

$$\det(AB) = \det(A)\det(B).$$

This property is one of the most useful, and it is employed often to actually compute determinants. A particularly interesting consequence is to note what it means for existence of inverses. Take A and B to be inverses, that is $AB = I$. Then

$$\det(A)\det(B) = \det(AB) = \det(I) = 1.$$

Neither $\det(A)$ nor $\det(B)$ can be zero. This fact is an extremely useful property of the determinant, and one which is used often in this book:

Theorem 3.5.3

An $n \times n$ matrix A is invertible if and only if $\det(A) \neq 0$.

In fact, $\det(A^{-1})\det(A) = 1$ says that

$$\det(A^{-1}) = \frac{1}{\det(A)}.$$

So we know what the determinant of A^{-1} is without computing A^{-1}.

Let us return to the formula for the inverse of a 2×2 matrix:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Notice the determinant of the matrix $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ in the denominator of the fraction. The formula only works if the determinant is nonzero, otherwise we are dividing by zero.

A common notation for the determinant is a pair of vertical lines:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = \det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right).$$

Personally, I find this notation confusing as vertical lines usually mean a positive quantity, while determinants can be negative. Also think about how to write the absolute value of a determinant. This notation is not used in this book.

With this discussion of determinants complete, we can now state a major theorem from linear algebra that will help us here and when we get back to solving differential equations using this linear algebra. In a full course on linear algebra, this theorem would be covered in full detail, including all of the proofs. For this introduction, we give some idea as to why everything is true here, but not all of the details.

Note: This is an example of an *equivalence* theorem, which is fairly common in mathematics. It means that if any one of the statements are true, then we know that all of the others are true as well. It means it's harder to prove, but once we have such a theorem, it is very powerful in how we can use it going forward.

Theorem 3.5.4

Let A be an $n \times n$ matrix. The following are equivalent:

(a) A is invertible.

(b) $\det(A) \neq 0$.

(c) There is a unique solution to $A\vec{x} = \vec{b}$ for every vector $\vec{b}$.

(d) The only solution to $A\vec{x} = \vec{0}$ is $\vec{x} = \vec{0}$.

(e) The reduced row echelon form of A is I_n, the identity matrix.

(f) The rank of A is n.

(g) The rows of A are linearly independent.

(h) The columns of A are linearly independent.

Proof. Why is all of this true? For (a) and (b), we have Theorem 3.5.3 to say that they are equivalent. For (c), if A is invertible, then the unique solution to $A\vec{x} = \vec{b}$ is $\vec{x} = A^{-1}\vec{b}$. If we take $\vec{b} = \vec{0}$ here, we get (d), that the solution is $\vec{x} = A^{-1}\vec{0} = \vec{0}$. This means that reducing the system of equations $A\vec{x} = 0$ gives $x_1 = 0$, $x_2 = 0$, ..., $x_n = 0$, which means the reduced row echelon form of A is just the identity matrix, which is (e). This has n pivot rows, so that the rank of A is n. Finally, this means that the dimension of the column space and row space is both n, and since there are n of these vectors, it means they are all linearly independent. □

This is a massive theorem that forms most of the backbone of linear algebra. We will only be using a few parts of it later, but since we have seen all of the components, it is nice to see them all put together into one complete statement.

3.5.1 Exercises

***Exercise* 3.5.2:** *Compute the determinant of the following matrices:*

a) $\begin{bmatrix} 3 \end{bmatrix}$ *b)* $\begin{bmatrix} 1 & 3 \\ 2 & 1 \end{bmatrix}$ *c)* $\begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix}$ *d)* $\begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix}$

e) $\begin{bmatrix} 2 & 1 & 0 \\ -2 & 7 & -3 \\ 0 & 2 & 0 \end{bmatrix}$ *f)* $\begin{bmatrix} 2 & 1 & 3 \\ 8 & 6 & 3 \\ 7 & 9 & 7 \end{bmatrix}$ *g)* $\begin{bmatrix} 0 & 2 & 5 & 7 \\ 0 & 0 & 2 & -3 \\ 3 & 4 & 5 & 7 \\ 0 & 0 & 2 & 4 \end{bmatrix}$ *h)* $\begin{bmatrix} 0 & 1 & 2 & 0 \\ 1 & 1 & -1 & 2 \\ 1 & 1 & 2 & 1 \\ 2 & -1 & -2 & 3 \end{bmatrix}$

Exercise* 3.5.3: *Compute the determinant of the following matrices:*

a) $\begin{bmatrix}-2\end{bmatrix}$ b) $\begin{bmatrix}2 & -2\\ 1 & 3\end{bmatrix}$ c) $\begin{bmatrix}2 & 2\\ 2 & 2\end{bmatrix}$ d) $\begin{bmatrix}2 & 9 & -11\\ 0 & -1 & 5\\ 0 & 0 & 3\end{bmatrix}$

e) $\begin{bmatrix}2 & 1 & 0\\ -2 & 7 & 3\\ 1 & 1 & 0\end{bmatrix}$ f) $\begin{bmatrix}5 & 1 & 3\\ 4 & 1 & 1\\ 4 & 5 & 1\end{bmatrix}$ g) $\begin{bmatrix}3 & 2 & 5 & 7\\ 0 & 0 & 2 & 0\\ 0 & 4 & 5 & 0\\ 2 & 1 & 2 & 4\end{bmatrix}$ h) $\begin{bmatrix}0 & 2 & 1 & 0\\ 1 & 2 & -3 & 4\\ 5 & 6 & -7 & 8\\ 1 & 2 & 3 & -2\end{bmatrix}$

***Exercise* 3.5.4:** *For which x are the following matrices singular (not invertible).*

a) $\begin{bmatrix}2 & 3\\ 2 & x\end{bmatrix}$ b) $\begin{bmatrix}2 & x\\ 1 & 2\end{bmatrix}$ c) $\begin{bmatrix}x & 1\\ 4 & x\end{bmatrix}$ d) $\begin{bmatrix}x & 0 & 1\\ 1 & 4 & 2\\ 1 & 6 & 2\end{bmatrix}$

Exercise* 3.5.5: *For which x are the following matrices singular (not invertible).*

a) $\begin{bmatrix}1 & 3\\ 1 & x\end{bmatrix}$ b) $\begin{bmatrix}3 & x\\ 1 & 3\end{bmatrix}$ c) $\begin{bmatrix}x & 3\\ 3 & x\end{bmatrix}$ d) $\begin{bmatrix}x & 1 & 0\\ 1 & 4 & 0\\ 1 & 6 & 2\end{bmatrix}$

Exercise* 3.5.6: *Consider the matrix*

$$A = \begin{bmatrix}0 & -1 & 0\\ -5 & -4 & -5\\ 2 & 3 & 4\end{bmatrix}.$$

a) *Compute the determinant of A using cofactor expansion along row 1.*

b) *Compute the determinant of A using cofactor expansion along column 2.*

c) *Compute the determinant using row reduction.*

Exercise* 3.5.7: *Consider the matrix*

$$A = \begin{bmatrix}-1 & 0 & -3\\ 1 & 2 & 1\\ 3 & 3 & 3\end{bmatrix}.$$

a) *Compute the determinant of A using cofactor expansion along row 1.*

b) *Compute the determinant of A using cofactor expansion along column 3.*

c) *Compute the determinant using row reduction.*

Exercise* 3.5.8: *Consider the matrix*

$$A = \begin{bmatrix}-2 & 0 & 1 & 0\\ 0 & -1 & 1 & -2\\ -5 & 3 & 1 & 3\\ -3 & 4 & 1 & 3\end{bmatrix}.$$

a) Compute the determinant of A using cofactor expansion along row 1.

b) Compute the determinant of A using cofactor expansion along column 4.

c) Compute the determinant using row reduction.

Exercise 3.5.9: *Is the matrix A below invertible? How do you know?*

$$A = \begin{bmatrix} 4 & 0 & 3 & 1 \\ 2 & 1 & -2 & 0 \\ 0 & 0 & 1 & -3 \\ 3 & 2 & 1 & -5 \end{bmatrix}$$

Exercise 3.5.10:* *Compute the rank of the matrix A below.*

$$A = \begin{bmatrix} 0 & -3 & 2 & 4 \\ -5 & -4 & -5 & -1 \\ 1 & 4 & -3 & -5 \\ -2 & -3 & -2 & 1 \end{bmatrix}$$

What does this tell you about the invertibility of A? How about the solutions to $A\vec{x} = \vec{0}$?

Exercise 3.5.11:* *Compute the rank of the matrix A below.*

$$A = \begin{bmatrix} 3 & -5 & 5 \\ 2 & -3 & 3 \\ 4 & 0 & -1 \end{bmatrix}$$

What does this tell you about the invertibility of A? How about the solutions to $A\vec{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$?

Exercise 3.5.12:* *Compute the determinant of the matrix*

$$A = \begin{bmatrix} 5 & 4 & 3 \\ -4 & -3 & -4 \\ -5 & -5 & 4 \end{bmatrix}$$

using row reduction. What does this say about the solutions to $A\vec{x} = 0$?

Exercise 3.5.13:* *Compute the determinant of the matrix*

$$A = \begin{bmatrix} -5 & -3 & -5 & -1 \\ 4 & 0 & -5 & 4 \\ 0 & -2 & -1 & -2 \\ -1 & -5 & -4 & -4 \end{bmatrix}$$

using row reduction. What does this say about the columns of A?

Exercise* 3.5.14: *Compute the determinant of the matrix*

$$A = \begin{bmatrix} 4 & 1 & -3 & 0 \\ -1 & 4 & 2 & -2 \\ -1 & -3 & 3 & 2 \\ -5 & -4 & 1 & 1 \end{bmatrix}$$

using row reduction. What does this say about the solutions to $A\vec{x} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 1 \end{bmatrix}$.

***Exercise* 3.5.15:** *Compute*

$$\det\left(\begin{bmatrix} 2 & 1 & 2 & 3 \\ 0 & 8 & 6 & 5 \\ 0 & 0 & 3 & 9 \\ 0 & 0 & 0 & 1 \end{bmatrix}^{-1}\right)$$

without computing the inverse.

Exercise* 3.5.16: *Compute*

$$\det\left(\begin{bmatrix} 3 & 4 & 7 & 12 \\ 0 & -1 & 9 & -8 \\ 0 & 0 & -2 & 4 \\ 0 & 0 & 0 & 2 \end{bmatrix}^{-1}\right)$$

without computing the inverse.

***Exercise* 3.5.17:** *Suppose*

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 7 & \pi & 1 & 0 \\ 2^8 & 5 & -99 & 1 \end{bmatrix} \qquad \text{and} \qquad U = \begin{bmatrix} 5 & 9 & 1 & -\sin(1) \\ 0 & 1 & 88 & -1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Let $A = LU$. *Compute* $\det(A)$ *in a simple way, without computing what is* A. *Hint: First read off* $\det(L)$ *and* $\det(U)$.

***Exercise* 3.5.18:** *Consider the linear mapping from* $\mathbb{R}^2$ *to* $\mathbb{R}^2$ *given by the matrix* $A = \begin{bmatrix} 1 & x \\ 2 & 1 \end{bmatrix}$ *for some number* x. *You wish to make* A *such that it doubles the area of every geometric figure. What are the possibilities for* x *(there are two answers).*

Exercise* 3.5.19** (challenging):*** *Find all the* x *that make the matrix inverse*

$$\begin{bmatrix} 1 & 2 \\ 1 & x \end{bmatrix}^{-1}$$

have only integer entries (no fractions). Note that there are two answers.

***Exercise* 3.5.20:** *Suppose* A *and* S *are* $n \times n$ *matrices, and* S *is invertible. Suppose that* $\det(A) = 3$. *Compute* $\det(S^{-1}AS)$ *and* $\det(SAS^{-1})$. *Justify your answer using the theorems in this section.*

***Exercise* 3.5.21:** *Let* A *be an* $n \times n$ *matrix such that* $\det(A) = 1$. *Compute* $\det(xA)$ *given a number* x. *Hint: First try computing* $\det(xI)$, *then note that* $xA = (xI)A$.

3.6 Eigenvalues and Eigenvectors

Learning Objectives

After this section, you will be able to:

- Find the eigenvalues and eigenvectors of a matrix,
- Use complex numbers to find eigenvalues and eigenvectors if necessary, and
- Identify the algebraic and geometric multiplicity of an eigenvalue to determine if it is defective.

Consider the matrix

$$A = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix}.$$

We can compute a few operations with this matrix. For instance

$$A \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

and

$$A \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 3 \end{bmatrix}.$$

This last computation is fairly interesting, because the result we get is the same as 3 times the original vector. However, the matrix A does not multiply every vector by 3, as seen in the first example and the fact that

$$A \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

so A actually preserves this vector, multiplying it by 1. So, these vectors, $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$, and numbers, 3 and 1, are somehow special for this matrix A. With this information, we want to define these vectors as *eigenvectors* and numbers as *eigenvalues* of the matrix A.

Definition 3.6.1

For a square matrix A, we say that non-zero vector $\vec{v}$ is an *eigenvector* of the matrix A if there exists a number λ so that

$$A\vec{v} = \lambda\vec{v}.$$

In this case, we say that λ is an *eigenvalue* of A and it is the *corresponding eigenvalue* for the eigenvector $\vec{v}$.

Thus, we can say that, for the matrix

$$A = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix},$$

we see that $\begin{bmatrix}2\\1\end{bmatrix}$ is an eigenvector with corresponding eigenvalue 3, and that $\begin{bmatrix}4\\3\end{bmatrix}$ is an eigenvector with corresponding eigenvalue 1.

Why are these important? It turns out that these eigenvalues and eigenvectors characterize the behavior of the matrix A. For example, if we wanted to figure out what happens when A is applied to the vector $\begin{bmatrix}6\\4\end{bmatrix}$, we can figure this out as

$$\begin{aligned} A\begin{bmatrix}6\\4\end{bmatrix} &= A\left(\begin{bmatrix}4\\3\end{bmatrix}+\begin{bmatrix}2\\1\end{bmatrix}\right) \\ &= A\begin{bmatrix}4\\3\end{bmatrix}+A\begin{bmatrix}2\\1\end{bmatrix} \\ &= \begin{bmatrix}4\\3\end{bmatrix}+3\begin{bmatrix}2\\1\end{bmatrix}=\begin{bmatrix}10\\6\end{bmatrix} \end{aligned}$$

In addition, eigenvectors determine directions in which multiplying by the matrix A behaves just like scalar multiplication. This idea will be very important for our understanding of systems of differential equations, because we have already seen how to solve a scalar first order equation way back in § 0.1 and § 1.3.

3.6.1 Finding Eigenvalues and Eigenvectors

Since eigenvalues and eigenvectors are so important, we want to know how to find them. To do this, we are looking for a number λ and a non-zero vector $\vec{v}$ so that

$$A\vec{v}=\lambda\vec{v}.$$

We can rewrite this as

$$A\vec{v}-\lambda\vec{v}=0$$

or, using the identity matrix,

$$(A-\lambda I)\vec{v}=0.$$

This means that we are looking for a non-zero solution to a homogeneous vector equation of the form $B\vec{v}=0$. This is where all of our linear algebra theory comes into play.

Theorem 3.5.4 tells us that, combining parts (b) and (d), that there is a non-zero solution to $(A-\lambda I)\vec{v}=0$ if and only if the determinant of the matrix $A-\lambda I$ is zero. Therefore, we can compute this determinant, find the values of λ so that $\det(A-\lambda I)=0$, and these will give us our eigenvalues. Let's see an example of what this looks like.

Example 3.6.1: Compute $\det(A-\lambda I)$ for the matrix

$$A=\begin{bmatrix}7&-8\\3&-3\end{bmatrix}.$$

Solution: For this matrix, we have that

$$A-\lambda I=\begin{bmatrix}7&-8\\3&-3\end{bmatrix}-\lambda\begin{bmatrix}1&0\\0&1\end{bmatrix}=\begin{bmatrix}7-\lambda&-8\\3&-3-\lambda\end{bmatrix}.$$

Thus

$$\begin{aligned}\det(A-\lambda I) &= \det\left(\begin{bmatrix}7-\lambda & -8\\ 3 & -3-\lambda\end{bmatrix}\right)\\ &= (7-\lambda)(-3-\lambda)-(-8)(3) = \lambda^2+3\lambda-7\lambda-21+24\\ &= \lambda^2-4\lambda+3\end{aligned}$$

If we were looking for eigenvalues here, we could then set this equal to zero, getting that

$$0 = \lambda^2-4\lambda+3 = (\lambda-1)(\lambda-3)$$

so that the eigenvalues are 1 and 3. ⌟

In this case, we saw that computing $\det(A-\lambda I)$ for this case, we ended up with a quadratic polynomial, so it was easy to find the eigenvalues. Thankfully, no matter the size of the matrix, we will always get a polynomial here.

Definition 3.6.2

For a matrix A, the expression $\det(A-\lambda I)$ is called the *characteristic polynomial* of the matrix. It will always be a polynomial, and for A an $n\times n$ matrix, it will be a degree n polynomial.

This explains why we got a quadratic polynomial for the 2×2 matrix A. Therefore, for a matrix A, the roots of the characteristic polynomial are the eigenvalues of A.

Once we have the eigenvalues, we can use them to find the eigenvectors. As with how we started this discussion, we are looking for a non-zero vector $\vec{v}$ so that

$$(A-\lambda I)\vec{v} = 0,$$

and we know the value of λ. Therefore, we can set up a system of equations that correponds to

$$(A-\lambda I)\vec{v} = 0$$

and solve it for the components of the eigenvector.

Example 3.6.2: Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix}7 & -8\\ 3 & -3\end{bmatrix}.$$

Solution: The previous example shows that the eigenvalues for this matrix are 1 and 3. For the eigenvalue 1, we want to find a non-zero solution to $(A-I)\vec{v}=0$, which means we want to solve for

$$(A-I)\vec{v} = \begin{bmatrix}7-1 & -8\\ 3 & -3-1\end{bmatrix}\vec{v} = \begin{bmatrix}6 & -8\\ 3 & -4\end{bmatrix}\vec{v} = 0.$$

Writing the vector $\vec{v}$ as $\begin{bmatrix}v_1\\ v_2\end{bmatrix}$, this system of equations becomes

$$\begin{aligned}6v_1-8v_2 &= 0\\ 3v_1-4v_2 &= 0\end{aligned}.$$

Since the second equation is two times the first one, these equations are redundant, so we only need to satisfy $3v_1 - 4v_2 = 0$. We can do this by choosing $v_1 = 4$ and $v_2 = 3$, which gives that for $\lambda = 1$, a corresponding eigenvector is $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$.

We can follow the same process for the eigenvalue 3. For this, we want to find a non-zero solution to $(A - 3I)\vec{v} = 0$, which means that we want to solve

$$(A - 3I)\vec{v} = \begin{bmatrix} 7-3 & -8 \\ 3 & -3-3 \end{bmatrix} \vec{v} = \begin{bmatrix} 4 & -8 \\ 3 & -6 \end{bmatrix} \vec{v} = 0.$$

Writing the vector $\vec{v}$ as $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, we get the two equations

$$\begin{aligned} 4v_1 - 8v_2 &= 0 \\ 3v_1 - 6v_2 &= 0 \end{aligned}.$$

As before, these two equations are the same, since they are both a multiple of $v_1 - 2v_2 = 0$. Therefore, we just need to find a solution to that previous equation, which can be done with $v_1 = 2$ and $v_2 = 1$. Therefore, an eigenvector for eigenvalue 3 is $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$.

This example illustrates the standard process that is always used to find eigenvalues and eigenvectors of matrices: find the characteristic polynomial, get the roots of this polynomial, and use each of these eigenvalues to set up a system of equations for the components of each eigenvector. In addition, the equations that we get from this system will always be redundant if we have found the eigenvalue correctly. Since $\det(A - \lambda I) = 0$, we know that the rows of the matrix $A - \lambda I$ are not linearly independent, and so the row-echelon form of $A - \lambda I$ must have a zero row in it. This process works for any size matrix, but it becomes harder to find the roots of this polynomial when it is higher degree.

Example 3.6.3: Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 1 & 6 & 0 \\ 9 & -4 & 10 \\ 2 & -6 & 3 \end{bmatrix}.$$

Solution: We start by hunting for eigenvalues by taking the determinant of $A - \lambda I$, which will require the cofactor expansion in order to solve.

$$\begin{aligned}
\det(A-\lambda I) &= \det\left(\begin{bmatrix} 1-\lambda & 6 & 0 \\ 9 & -4-\lambda & 10 \\ 2 & -6 & 3-\lambda \end{bmatrix}\right) \\
&= (1-\lambda)\det\left(\begin{bmatrix} -4-\lambda & 10 \\ -6 & 3-\lambda \end{bmatrix}\right) - 6\det\left(\begin{bmatrix} 9 & 10 \\ 2 & 3-\lambda \end{bmatrix}\right) \\
&= (1-\lambda)((-4-\lambda)(3-\lambda)+60) - 6(9(3-\lambda)-20) \\
&= (1-\lambda)(\lambda^2+4\lambda-3\lambda-12+60) - 6(27-9\lambda-20) \\
&= (1-\lambda)(\lambda^2+\lambda+48) - 42 + 54\lambda \\
&= \lambda^2+\lambda+48-\lambda^3-\lambda^2-48\lambda-42+54\lambda \\
&= -\lambda^3+7\lambda+6
\end{aligned}$$

We need to look for the roots of this polynomial. There's no nice way to factor this right away, so we need to start guessing roots. We know that the root must be a factor of 6. If we try $\lambda = 1$, we get

$$-1+7+6 = 12 \neq 0$$

so that one doesn't work. Plugging in $\lambda = -1$, we get

$$-(-1)^3 - 7 + 6 = 1 - 7 + 6 = 0$$

so this is a root, meaing that $\lambda + 1$ is a factor of the characteristic polynomial. We can then use polynomial long division to get that

$$-\lambda^3+7\lambda+6 = (\lambda+1)(-\lambda^2+\lambda+6) = -(\lambda+1)(\lambda^2-\lambda-6)$$

and the quadratic term here factors as $(\lambda-3)(\lambda+2)$. Thus, the characteristic polynomial of this matrix is

$$(\lambda+1)(\lambda-3)(\lambda+2)$$

so the eigenvalues are -1, 3, and -2.

For the eigenvalue -1, the eigenvector must satisfy

$$(A+I)\vec{v} = \vec{0}$$

which we can write as

$$\begin{bmatrix} 2 & 6 & 0 \\ 9 & -3 & 10 \\ 2 & -6 & 4 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \vec{0}.$$

To solve this, we row-reduce the coefficient matrix.

$$\begin{bmatrix} 2 & 6 & 0 \\ 9 & -3 & 10 \\ 2 & -6 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 3 & 0 \\ 9 & -3 & 10 \\ 2 & -6 & 4 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} 1 & 3 & 0 \\ 0 & -30 & 10 \\ 0 & -12 & 4 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} 1 & 3 & 0 \\ 0 & -3 & 1 \\ 0 & -12 & 4 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} 1 & 3 & 0 \\ 0 & -3 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Therefore, the eigenvector must satsify $v_1 + 3v_2 = 0$ and $-3v_2 + v_3 = 0$. We need to pick any non-zero set of numbers that solves these equations. For example, we could pick $v_2 = 1$ to get that we need $v_1 = -3$ and $v_3 = 3$. This gives an eigenvector of

$$\begin{bmatrix} -3 \\ 1 \\ 3 \end{bmatrix}.$$

For the eigenvalue 3, the eigenvector must satisfy

$$\begin{bmatrix} -2 & 6 & 0 \\ 9 & -7 & 10 \\ 2 & -6 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \vec{0}.$$

Row reduction gives

$$\begin{bmatrix} -2 & 6 & 0 \\ 9 & -7 & 10 \\ 2 & -6 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -3 & 0 \\ 9 & -7 & 10 \\ 2 & -6 & 0 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} 1 & -3 & 0 \\ 0 & 20 & 10 \\ 0 & 0 & 0 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} 1 & -3 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

which means that the eigenvector must satisfy $v_1 - 3v_2 = 0$ and $2v_2 + v_3 = 0$. Again, choosing $v_2 = 1$ gives that we want $v_1 = 3$ and $v_3 = -2$. Therefore, a corresponding eigenvector here is

$$\begin{bmatrix} 3 \\ 1 \\ -2 \end{bmatrix}.$$

For the eigenvalue -2, the eigenvector must satisfy

$$\begin{bmatrix} 3 & 6 & 0 \\ 9 & -2 & 10 \\ 2 & -6 & 5 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \vec{0}$$

where we can row reduce the coefficient matrix.

$$\begin{bmatrix} 3 & 6 & 0 \\ 9 & -2 & 10 \\ 2 & -6 & 5 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 0 \\ 9 & -2 & 10 \\ 2 & -6 & 5 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} 1 & 2 & 0 \\ 0 & -20 & 10 \\ 0 & -10 & 5 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} 1 & 2 & 0 \\ 0 & -2 & 1 \\ 0 & -10 & 5 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} 1 & 2 & 0 \\ 0 & -2 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Therefore, the eigenvector must satisfy $v_1 + 2v_2 = 0$ and $-2v_2 + v_3 = 0$. Picking $v_2 = 1$ again gives that we want $v_1 = -2$ and $v_3 = 2$. Therefore, an eigenvector with eigenvalue -2 is

$$\begin{bmatrix} -2 \\ 1 \\ 2 \end{bmatrix}.$$

3.6.2 Real Eigenvalues

Since eigenvalues come from finding the roots of a polynomial, there are a few different situations that can arise in terms of these eigenvalues. If we take a quadratic polynomial, there are three options for the two roots.

- Two real and different roots,
- Two complex roots in a conjugate pair, or
- One double (repeated) root.

The same is true for eigenvalues, they are either all real and distinct, there are some that appear in complex conjugate pairs, or there are some repeated eigenvalues. The easiest of these cases is when the characteristic polynomial has all real and distinct eigenvalues.

In this case, we get a very nice result. We know that for each eigenvalue, there will always be at least one eigenvector, otherwise it wouldn't be an eigenvalue. If the matrix A is an

$n \times n$ matrix, then the characteristic polynomial is a degree n polynomial, which will have n distinct roots by our assumption. Each of these will have a corresponding eigenvector, giving us n eigenvectors as well. A more involved result tells us that eigenvectors for different eigenvalues are always linearly independent. Therefore, we get n vectors in $\mathbb{R}^n$, that are linearly independent, and so they are a basis. This gives the following result.

Theorem 3.6.1

Let A be an $n \times n$ matrix. Assume that the characteristic polynomial of A has all real and distinct roots, namely that

$$\det(A - \lambda I) = (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n)$$

for $\lambda_1, ..., \lambda_n$ the distinct real eigenvalues. Then there exist vectors $\vec{v}_1$, ..., $\vec{v}_n$ such that $\vec{v}_i$ is an eigenvector for eigenvalue λ_i and $\{\vec{v}_1, ..., \vec{v}_n\}$ form a basis of $\mathbb{R}^n$.

To reference, look at the previous example. We found three distinct real eigenvalues of -1, 3, and -2. For these eigenvalues, we had eigenvectors

$$-1 \to \begin{bmatrix} -3 \\ 1 \\ 3 \end{bmatrix} \quad 3 \to \begin{bmatrix} 3 \\ 1 \\ -2 \end{bmatrix} \quad -2 \to \begin{bmatrix} -2 \\ 1 \\ 2 \end{bmatrix}.$$

These three vectors are linearly independent (check this!) and since they are three component vectors, the space has dimension 3, and so 3 linearly independent vectors must make up a basis. This is useful to know for now, but will be critical when we want to use this information to solve systems of differential equations later.

3.6.3 Complex Eigenvalues

When the matrix has complex eigenvalues, the process is very similar to before. However, the eigenvector will necessarily also be complex, that is, some of the components of this vector will be complex numbers. Let's illustrate this with an example.

Example 3.6.4: Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 3 & -8 \\ 5 & -9 \end{bmatrix}.$$

Solution: We first look for the eigenvalues using the characteristic polynomial of A.

$$\begin{aligned} \det(A - \lambda I) &= \det\left(\begin{bmatrix} 3-\lambda & -8 \\ 5 & -9-\lambda \end{bmatrix}\right) \\ &= (3-\lambda)(-9-\lambda) + 40 \\ &= \lambda^2 + 9\lambda - 3\lambda - 27 + 40 \\ &= \lambda^2 + 6\lambda + 13 \end{aligned}.$$

This quadratic does not factor, so we use the quadratic formula to find that

$$\lambda = \frac{-6 \pm \sqrt{6^2 - 4 \cdot 13}}{2} = \frac{-6 \pm \sqrt{-16}}{2} = -3 \pm 2i$$

so that we have complex eigenvalues.

We now look for the eigenvectors in the same way as in the real case. If we take the eigenvalue $-3 + 2i$, then such an eigenvector must satisfy

$$(A - (-3 + 2i)I)\vec{v} = \vec{0}.$$

This means that

$$\begin{bmatrix} 3 - (-3 + 2i) & -8 \\ 5 & -9 - (-3 + 2i) \end{bmatrix} \vec{v} = \begin{bmatrix} 6 - 2i & -8 \\ 5 & -6 - 2i \end{bmatrix} \vec{v} = \vec{0}.$$

These two equations should be redundant, and to verify that, we will multiply the top row by $6 + 2i$ in row reduction to get

$$\begin{bmatrix} 6 - 2i & -8 \\ 5 & -6 - 2i \end{bmatrix} \to \begin{bmatrix} (6 - 2i)(6 + 2i) & -8(6 + 2i) \\ 5 & -6 - 2i \end{bmatrix}$$
$$\to \begin{bmatrix} 40 & -48 - 16i \\ 5 & -6 - 2i \end{bmatrix}$$

and from this, we can see that the top row is 8 times the bottom one, so they are redundant. Thus, an eigenvector must satisfy

$$5v_1 - (6 + 2i)v_2 = 0$$

and we can pick any non-zero numbers that satisfy this. One simple way to do this is by switching the coefficients, so that $v_1 = 6 + 2i$ and $v_2 = 5$. Therefore, an eigenvector that we get is

$$\begin{bmatrix} 6 + 2i \\ 5 \end{bmatrix}.$$

Now, we can take the other eigenvalue, $-3 - 2i$. The process is the same, so that the vector must satisfy

$$\begin{bmatrix} 3 - (-3 - 2i) & -8 \\ 5 & -9 - (-3 - 2i) \end{bmatrix} \vec{v} = \begin{bmatrix} 6 + 2i & -8 \\ 5 & -6 + 2i \end{bmatrix} \vec{v} = \vec{0}.$$

To check redundancy again, we multiply the top row by $6 - 2i$ to get

$$\begin{bmatrix} 6 + 2i & -8 \\ 5 & -6 + 2i \end{bmatrix} \to \begin{bmatrix} (6 + 2i)(6 - 2i) & -8(6 - 2i) \\ 5 & -6 + 2i \end{bmatrix}$$
$$\to \begin{bmatrix} 40 & -48 + 16i \\ 5 & -6 + 2i \end{bmatrix}$$

and again, the first equation is 8 times the second one. Thus, the eigenvector will need to satisfy

$$5v_1 - (6-2i)v_2 = 0$$

which can be done by picking $v_1 = 6-2i$ and $v_2 = 5$, giving an eigenvector of

$$\begin{bmatrix} 6-2i \\ 5 \end{bmatrix}.$$

□

The process here is the same as it was in the real case, except that now all of the equations are complex equations. In particular, the "redundancy" that we expect to see between the equations will likely be via a complex multiple. The easiest way to verify that these equations are redundant is by multiplying the first entry in each row by its complex conjugate. This is because, if we have the complex number $a+bi$, multiplying this by $a-bi$ gives

$$(a+bi)(a-bi) = a^2 + abi - abi - b^2i^2 = a^2 + b^2$$

which is now a real number. This will make it easier to compare the two equations to make sure that they are redundant, and that the eigenvalue was found correctly.

Example 3.6.5: Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 1 & 9 & 6 \\ 0 & 1 & 6 \\ 0 & -3 & -5 \end{bmatrix}.$$

Solution: We first look for eigenvalues, like always. We get these by computing

$$\det(A - \lambda I) = \det\left(\begin{bmatrix} 1-\lambda & 9 & 6 \\ 0 & 1-\lambda & 6 \\ 0 & -3 & -5-\lambda \end{bmatrix}\right).$$

We will compute this by cofactor expansion along the second row.

$$\begin{aligned} \det\left(\begin{bmatrix} 1-\lambda & 9 & 6 \\ 0 & 1-\lambda & 6 \\ 0 & -3 & -5-\lambda \end{bmatrix}\right) &= (-1)^{2+2}(1-\lambda)\det\left(\begin{bmatrix} 1-\lambda & 6 \\ 0 & -5-\lambda \end{bmatrix}\right) \\ &\quad + (-1)^{2+3}6\det\left(\begin{bmatrix} 1-\lambda & 9 \\ 0 & -3 \end{bmatrix}\right) \\ &= (1-\lambda)(1-\lambda)(-5-\lambda) - 6(1-\lambda)(-3) \\ &= (1-\lambda)((1-\lambda)(-5-\lambda)+18) \\ &= (1-\lambda)(\lambda^2 + 4\lambda + 13) \end{aligned}$$

so that one eigenvalue is at $\lambda = 1$. For the other two, we use the quadratic formula to obtain

$$\lambda = \frac{-4 \pm \sqrt{16 - 4\cdot 13}}{2} = \frac{-4 \pm \sqrt{-36}}{2} = -2 \pm 3i.$$

Thus, we have one real eigenvalue and two complex eigenvalues.

For $\lambda = 1$, we know that the eigenvector must satisfy

$$\begin{bmatrix} 0 & 9 & 6 \\ 0 & 0 & 6 \\ 0 & -3 & -6 \end{bmatrix} \vec{v} = \vec{0}.$$

Row reduction will reduce this matrix to

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

(*Check this!*) so that the eigenvector in this case is

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

For the eigenvalue $-2 + 3i$, we get that the eigenvector must satisfy

$$\begin{bmatrix} 3-3i & 9 & 6 \\ 0 & 3-3i & 6 \\ 0 & -3 & -3-3i \end{bmatrix} \vec{v} = \vec{0}.$$

We now want to row reduce the coefficient matrix. To do so, we start by dividing the first row by 3 then multiplying by $1 + i$.

$$\begin{bmatrix} 3-3i & 9 & 6 \\ 0 & 3-3i & 6 \\ 0 & -3 & -3-3i \end{bmatrix} \to \begin{bmatrix} 1-i & 3 & 2 \\ 0 & 3-3i & 6 \\ 0 & -3 & -3-3i \end{bmatrix}$$
$$\to \begin{bmatrix} (1-i)(1+i) & 3(1+i) & 2(1+i) \\ 0 & 3-3i & 6 \\ 0 & -3 & -3-3i \end{bmatrix}.$$
$$\to \begin{bmatrix} 2 & 3+3i & 2+2i \\ 0 & 3-3i & 6 \\ 0 & -3 & -3-3i \end{bmatrix}$$

We could divide the first row by 2 to get to a 1 in the top-right entry, but we'll wait on that in order to avoid fractions. To row reduce the rest of the matrix, we will divide each of the remaining two rows by 3, and then multiply the second by $1 + i$, just like we did to the first row.

$$\begin{bmatrix} 2 & 3+3i & 2+2i \\ 0 & 3-3i & 6 \\ 0 & -3 & -3-3i \end{bmatrix} \to \begin{bmatrix} 2 & 3+3i & 2+2i \\ 0 & 1-i & 2 \\ 0 & -1 & -1-i \end{bmatrix}$$
$$\to \begin{bmatrix} 2 & 3+3i & 2+2i \\ 0 & 2 & 2+2i \\ 0 & -1 & -1-i \end{bmatrix}$$
$$\to \begin{bmatrix} 2 & 3+3i & 2+2i \\ 0 & 1 & 1+i \\ 0 & -1 & -1-i \end{bmatrix}$$

which illlustrates that the last two rows are redundant. Thus, the reduced form of the matrix that we have (which is not quite a row echelon form, but it is enough to back-solve for the eigenvector) is

$$\begin{bmatrix} 2 & 3+3i & 2+2i \\ 0 & 1 & 1+i \\ 0 & 0 & 0 \end{bmatrix}.$$

This means that the eigenvector $\vec{v}$ must satisfy

$$2v_1 + (3+3i)v_2 + (2+2i)v_3 = 0 \qquad v_2 + (1+i)v_3 = 0.$$

We can satisfy the second of these equations by choosing $v_2 = 1+i$ and $v_3 = -1$. Plugging these values into the first equation gives that

$$\begin{aligned} 0 &= 2v_1 + (3+3i)v_2 + (2+2i)v_3 \\ &= 2v_1 + (3+3i)(1+i) + (2+2i)(-1) \\ &= 2v_1 + 3 + 3i + 3i - 3 - 2 - 2i \\ &= 2v_1 - 2 + 4i \end{aligned}$$

Therefore, we need to take $v_1 = 1-2i$, giving that the eigenvector is

$$\begin{bmatrix} 1-2i \\ 1+i \\ -1 \end{bmatrix}.$$

A very similar computation following the same set of steps (or just using the remark below) for the eigenvalue $-2-3i$ gives that this corresponding eigenvector is

$$\begin{bmatrix} 1+2i \\ 1-i \\ -1 \end{bmatrix}.$$

One fact that comes out of those examples is that the eigenvectors for conjugate eigenvalues are also complex conjugates. This comes from the fact that A is a real matrix, which means that if

$$A\vec{v} = \lambda\vec{v}$$

and we take the complex conjugate of both sides, we get that

$$A\bar{\vec{v}} = \bar{A}\bar{\vec{v}} = \bar{\lambda}\bar{\vec{v}} = \bar{\lambda}\bar{\vec{v}}$$

so that $\bar{\vec{v}}$ is an eigenvector for $\bar{\lambda}$. This means that, when solving these types of problems, we only need to find one of the complex eigenvectors and can get the other by taking the complex conjugate.

3.6.4 Repeated Eigenvalues

Distinct and complex eigenvalues all work out nicely and in pretty much the same manner. For repeated eigenvalues, the issues get more significant.

Example 3.6.6: Find the eigenvalues and eigenvectors of the matrices

$$A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} \qquad B = \begin{bmatrix} 4 & -1 \\ 1 & 2 \end{bmatrix}.$$

Solution: For the matrix A, we can compute the characteristic polynomial

$$\det(A - \lambda I) = \det\left(\begin{bmatrix} 3-\lambda & 0 \\ 0 & 3-\lambda \end{bmatrix}\right) = (3-\lambda)(3-\lambda)$$

Therefore, we have a double root at 3 for this matrix. Therefore, the only eigenvalue we get is 3. When we look to find the eigenvectors, we get

$$A - 3I = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

so that this matrix multiplied by *any* vector is zero. Therefore, we can use both $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ as eigenvectors.

On the other hand, the matrix B has a characteristic polynomial

$$\begin{aligned} \det(B - \lambda I) &= \det\left(\begin{bmatrix} 4-\lambda & -1 \\ 1 & 2-\lambda \end{bmatrix}\right) \\ &= (4-\lambda)(2-\lambda) - (-1)(1) = \lambda^2 - 6\lambda + 8 + 1 \\ &= \lambda^2 - 6\lambda + 9 = (\lambda - 3)^2 \end{aligned}$$

so again, we have a double root at 3. However, when we go to find the eigenvectors, we get that

$$B - 3I = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$$

which gives that an eigenvector must satisfy $v_1 - v_2 = 0$ so $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ works. ⌟

There is a big difference between these two examples. Both had the same characteristic polynomial of $(\lambda - 3)^2$, but for A, we could find two linearly independent eigenvectors, but

for B, we could only find 1. This seems like it might be a problem, since we would like to get to two eigenvectors like we did for both of the previous two cases. This leads us to define the following for A and $n \times n$ matrix and r an eigenvalue of A.

Definition 3.6.3

- The *algebraic multiplicity* of r is the power of $(\lambda - r)$ in the characteristic polynomial if A.
- The *geometric multiplicity* of r is the number of linearly independent eigenvectors of A with eigenvalue r.
- The *defect* of r is the difference between the algebraic multiplicity and the geometric multiplicity of r.
- We say that an eigenvalue is *defective* if the defect is at least 1.

For the previous example, the algebraic multiplicity of 3 for both A and B was 2, but the geometric multiplictiy of 3 for A is 2, and for B is it only 1. Therefore A has a defect of 0 and B has a defect of 1, so 3 is a defective eigenvalue for matrix B.

In terms of these multiplicities, there are two facts that are known to be true.

1. If r is an eigenvalue, then both the algebraic and geometric multiplicity are at least 1.
2. The algebraic multiplicity of any eigenvalue is always greater than or equal to the geometric multiplicity.

This tells us that in the case of real and distinct eigenvalues, every eigenvalue has multiplicity 1. Since the geometric multiplicity is also 1, this means that none of these eigenvalues are defective. This was great, because it let us get to n eigenvectors for an $n \times n$ matrix, and these generated a basis of $\mathbb{R}^n$.

Why is a defective eigenvalue a problem? When we go solve differential equations using the method in Chapter 4, having a 'full set' of eigenvectors, or n eigenvectors for an $n \times n$ matrix, will be very important. When we have a defective eigenvalue, we can't get there. Since the degree of the characteristic polynomial is n, the only way we get to n eigenvectors is if every eigenvalue has a number of linearly independent eigenvectors equal to the algebraic multiplicity, which means they are not defective.

So how can we fix this? Well, there's not really much we can do in the way of finding more eigenvectors, because they don't exist. The replacement that we have is, in linear algebra contexts, called a *generalized eigenvector*. We will see this idea come back up in § 4.6 in a more natural way. The rest of this section contains a more detailed definition of generalized eigenvectors. You are welcome to skip this part on a first reading and come back after you are more comfortable with eigenvalues and eigenvectors, or when the material comes back around again in § 4.6.

If r is an defective eigenvalue of the matrix A with eigenvector $\vec{v}$, a *generalized eigenvector* of A is a vector $\vec{w}$ so that $(A - rI)\vec{w} = \vec{v}$. This is the same as the normal eigenvector equation with $\vec{v}$ on the right-hand side instead of $\vec{0}$. Since $(A - rI)\vec{v} = \vec{0}$, this also means that

$$(A - rI)^2\vec{w} = 0.$$

More generally, a generalized eigenvector is a vector $\vec{w}$ where there is a power $k \geq 1$ so that

$$(A - rI)^k \vec{w} = 0 \qquad \text{but} \qquad (A - rI)^{k-1} \vec{w} \neq 0.$$

It might seems strange where this comes from, but we will see why this formula makes more sense once we try to solve differential equations using matrices in § 4.6.

Example 3.6.7: Find a generalized eigenvector of eigenvalue 3 for the matrix

$$B = \begin{bmatrix} 4 & -1 \\ 1 & 2 \end{bmatrix}.$$

Solution: Previously, we found that $\vec{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is an eigenvector for B with eigenvalue 3. To find a generalized eigenvector, we need a vector $\vec{w}$ so that

$$(B - 3I)\vec{w} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Plugging in the matrix for $B - 3I$ gives that we need

$$\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Both of the rows of this matrix becomes the equation

$$w_1 - w_2 = 1.$$

There are many values of w_1 and w_2 that make this work. We can pick $w_1 = 1$ and $w_2 = 0$. This will give a generalized eigenvector of $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$. We could also pick $w_1 = 3$ and $w_2 = 2$, to get a generalized eigenvector as $\begin{bmatrix} 3 \\ 2 \end{bmatrix}$. Any of these choices work as a generalized eigenvector. ⌟

Example 3.6.8: Find the eigenvalues and eigenvectors (and generalized eigenvectors if needed) of the matrix

$$A = \begin{bmatrix} -2 & 0 & 1 \\ 19 & 2 & -16 \\ -1 & 0 & 0 \end{bmatrix}.$$

Solution: We start by looking for the eigenvalues through the characteristic polynomial.

$$\det(A - \lambda I) = \det\left(\begin{bmatrix} -2-\lambda & 0 & 1 \\ 19 & 2-\lambda & -16 \\ -1 & 0 & 0-\lambda \end{bmatrix} \right)$$

To compute this determinant, we will expand along column 2, because it only has one non-zero entry. This gives

$$\begin{aligned} \det(A - \lambda I) &= (-1)^{2+2}(2-\lambda) \det\left(\begin{bmatrix} -2-\lambda & 1 \\ -1 & -\lambda \end{bmatrix} \right) \\ &= (2-\lambda)((-2-\lambda)(-\lambda) + 1) \\ &= (2-\lambda)(\lambda^2 + 2\lambda + 1) = (2-\lambda)(\lambda+1)^2 \end{aligned}$$

so we have an eigenvalue at 2 and a double eigenvalue at -1.

First, let's look for the eigenvector for eigenvalue 2. In this case, we know that the eigenvector must satisfy

$$\begin{bmatrix} -4 & 0 & 1 \\ 19 & 0 & -16 \\ -1 & 0 & -2 \end{bmatrix} \vec{v} = \vec{0}.$$

Row reducing the coefficient matrix will give

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

so that a corresponding eigenvector is

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

since we know that $v_1 = 0$ and $v_3 = 0$.

For $\lambda = -1$, we see that an eigenvector must satisfy

$$\begin{bmatrix} -1 & 0 & 1 \\ 19 & 3 & -16 \\ -1 & 0 & 1 \end{bmatrix} \vec{v} = \vec{0}.$$

We now look to row reduce this coefficient matrix.

$$\begin{bmatrix} -1 & 0 & 1 \\ 19 & 3 & -16 \\ -1 & 0 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & -1 \\ 19 & 3 & -16 \\ -1 & 0 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & -1 \\ 0 & 3 & 3 \\ 0 & 0 & 0 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Therefore, we know that

$$v_1 - v_3 = 0 \qquad v_2 + v_3 = 0.$$

If we pick $v_3 = 1$, then we know that $v_2 = -1$ and $v_1 = 1$, so the only eigenvector we get for $\lambda = -1$ is

$$\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}.$$

Since we only found one eigenvector for $\lambda = -1$ and $\lambda + 1$ was squared in the characteristic polynomial, this is a defective eigenvalue. Thus, we can look for a generalized eigenvalue

here, which means that we need to solve for a vector $\vec{w}$ with

$$\begin{bmatrix} -1 & 0 & 1 \\ 19 & 3 & -16 \\ -1 & 0 & 1 \end{bmatrix} \vec{w} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

We can then row reduce the augmented matrix to see what we can pick for $\vec{w}$.

$$\begin{bmatrix} -1 & 0 & 1 & 1 \\ 19 & 3 & -16 & -1 \\ -1 & 0 & 1 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & -1 & -1 \\ 19 & 3 & -16 & -1 \\ -1 & 0 & 1 & 1 \end{bmatrix}$$
$$\to \begin{bmatrix} 1 & 0 & -1 & -1 \\ 0 & 3 & 3 & 18 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$
$$\to \begin{bmatrix} 1 & 0 & -1 & -1 \\ 0 & 1 & 1 & 6 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Thus, the generalized eigenvector $\vec{w}$ must satisfy

$$w_1 - w_3 = -1 \qquad w_2 + w_3 = 6.$$

We can pick any non-zero numbers to do this, so we can take $w_3 = 1$, $w_2 = 5$ and $w_1 = 0$. Thus, the generalized eigenvector here is

$$\begin{bmatrix} 0 \\ 5 \\ 1 \end{bmatrix}.$$

3.6.5 Exercises

Exercise* 3.6.1*: *Find the eigenvalues and corresponding eigenvectors of the matrix*

$$\begin{bmatrix} -8 & -18 \\ 4 & 10 \end{bmatrix}$$

Exercise* 3.6.2*: *Find the eigenvalues and corresponding eigenvectors of the matrix*

$$\begin{bmatrix} -2 & 0 \\ 8 & -4 \end{bmatrix}$$

Exercise* 3.6.3*: *Find the eigenvalues and corresponding eigenvectors of the matrix*

$$\begin{bmatrix} -7 & 1 \\ -12 & 0 \end{bmatrix}$$

Exercise* 3.6.4*: *Find the eigenvalues and corresponding eigenvectors of the matrix*

$$\begin{bmatrix} -3 & 5 \\ -8 & 9 \end{bmatrix}$$

Exercise* 3.6.5*: *Find the eigenvalues and corresponding eigenvectors of the matrix*

$$\begin{bmatrix} 0 & 2 \\ -1 & -2 \end{bmatrix}$$

Exercise* 3.6.6*: *Find the eigenvalues and corresponding eigenvectors of the matrix*

$$\begin{bmatrix} -4 & 1 \\ -8 & 0 \end{bmatrix}$$

Exercise* 3.6.7*: *Find the eigenvalues and corresponding eigenvectors of the matrix*

$$\begin{bmatrix} 0 & -8 \\ 2 & 8 \end{bmatrix}$$

Exercise* 3.6.8*: *Find the eigenvalues and corresponding eigenvectors of the matrix*

$$\begin{bmatrix} 1 & -2 \\ 8 & -7 \end{bmatrix}$$

Exercise* 3.6.9*: *Find the eigenvalues and corresponding eigenvectors of the matrix*

$$\begin{bmatrix} 4 & 0 & 0 \\ -4 & 2 & 1 \\ -6 & 0 & 1 \end{bmatrix}$$

Exercise* 3.6.10*: *Find the eigenvalues and corresponding eigenvectors of the matrix*

$$\begin{bmatrix} -4 & 9 & 9 \\ -3 & 6 & 9 \\ 3 & -7 & -10 \end{bmatrix}$$

Exercise* 3.6.11*: *Find the eigenvalues and corresponding eigenvectors of the matrix*

$$\begin{bmatrix} -2 & 0 & 0 \\ 0 & 4 & 6 \\ 6 & -3 & -2 \end{bmatrix}$$

Exercise* 3.6.12*: *Find the eigenvalues and corresponding eigenvectors of the matrix*

$$\begin{bmatrix} 5 & 3 & 6 \\ 2 & 2 & 2 \\ -3 & -2 & -3 \end{bmatrix}$$

Exercise* 3.6.13: *Find the eigenvalues and eigenvectors for the matrix below. Compute generalized eigenvectors if needed to get to a total of two vectors.*

$$\begin{bmatrix} -11 & -9 \\ 12 & 10 \end{bmatrix}$$

Exercise* 3.6.14: *Find the eigenvalues and eigenvectors for the matrix below. Compute generalized eigenvectors if needed to get to a total of two vectors.*

$$\begin{bmatrix} 4 & -4 \\ 1 & 0 \end{bmatrix}$$

***Exercise* 3.6.15:** *This exercise will work through the process of finding the eigenvalues and corresponding eigenvectors of the matrix*

$$A = \begin{bmatrix} -2 & 0 & -3 \\ 12 & 5 & 12 \\ 0 & -1 & 1 \end{bmatrix}.$$

a) *Find the characteristic polynomial of this matrix by computing* $\det(A - \lambda I)$*. In finding this, use cofactor expansion along either column 1 or row 1 and do* ***not*** *expand out all of the terms. Use grouping to factor this polynomial.*

b) *This polynomial can be rewritten as* $-(\lambda - r_1)^2(\lambda - r_2)$ *where* r_1 *and* r_2 *are the eigenvalues of* A*. What are the eigenvalues? What is each of their algebraic multiplicity?*

c) *Find an eigenvector for eigenvalue* r_2 *above. What is the geometric multiplicity of this eigenvalue?*

d) *Find an eigenvector for eigenvalue* r_1*. What is the geometric multiplicity of this eigenvalue?*

e) *There is only one possible eigenvector for* r_1*, which means it is defective. Find a solution to the equation* $(A - r_1 I)\vec{w} = \vec{v}$*, where* $\vec{v}$ *is the eigenvector you found in the previous part. This is the generalized eigenvector for* r_1*.*

***Exercise* 3.6.16:** *We say that a matrix* A *is* diagonalizable *if there exist matrices* D *and* P *so that* $PDP^{-1} = A$*. This really means that* A *can be represented by a diagonal matrix in a different basis (as opposed to the standard basis). One way this can be done is with eigenvalues.*

a) *Consider the matrix* A *given by*

$$A = \begin{bmatrix} -4 & 6 \\ -1 & 1 \end{bmatrix}.$$

Find the eigenvalues and corresponding eigenvectors of this matrix.

b) *Form two matrices, D, a diagonal matrix with the eigenvalues of A on the diagonal, and E, a matrix whose columns are the eigenvectors of A in the* same order *as the eigenvalues were put into D. Write out these matrices.*

c) *Compute E^{-1}.*

d) *Work out the products EDE^{-1} and $E^{-1}AE$. What do you notice here?*

This shows that, in the case of a 2×2 matrix, if we have two distinct real eigenvalues, that matrix is diagonalizable, using the eigenvectors.

***Exercise* 3.6.17:** *Follow the process outlined in Exercise 3.6.16 to attempt to diagonalize the matrix*

$$\begin{bmatrix} 13 & 14 & 12 \\ -6 & -4 & -6 \\ -3 & -6 & -2 \end{bmatrix}$$

Hint: 1 is an eigenvalue.

***Exercise* 3.6.18:** *The diagonalization process decribed in Exercise 3.6.16 works for any case where there are real and distinct eigenvalues, as well as complex eigenvalues (but the algebra with the complex numbers gets complicated). It may or may not work in the case of repeated eigenvalues, and it fails whenever there are defective eigenvalues. Consider the matrix*

$$\begin{bmatrix} 4 & -1 \\ 1 & 2 \end{bmatrix}$$

a) *Find the eigenvalue(s) of this matrix, and see that we have a repeated eigenvalue.*

b) *Find the eigenvector for that eigenvalue, as well as a generalized eigenvector.*

c) *Build a matrix E like before, but this time put the eigenvector in the first column and the generalized eigenvector in the second. Compute E^{-1}.*

d) *Find the product $E^{-1}AE$. Before, this gave us a diagonal matrix, but what do we get now?*

The matrix we get here is almost diagonal, but not quite. It turns out that this is the best we can do for matrices with defective eigenvalues. This matrix is often called J and is the Jordan Form *of the matrix A.*

Exercise* 3.6.19: *Follow the process in Exercise 3.6.18 to find the Jordan Form of the matrix*

$$\begin{bmatrix} -7 & 5 & 5 \\ -4 & 5 & 7 \\ -6 & 3 & 1 \end{bmatrix}.$$

3.7 Related Topics in Linear Algebra

Attribution: [JL], §A.4.

Learning Objectives

After this section, you will be able to:

- Determine the kernel of a matrix using row reduction,
- Understand the connection between rank and nullity in a given matrix,
- Compute the inverse of a matrix using row reduction, and
- Use properties of the trace and determinant to analyze the eigenvalues of a matrix.

3.7.1 Kernel

The set of solutions of a linear equation $L\vec{x} = \vec{0}$, the kernel of L, is a subspace: If $\vec{x}$ and $\vec{y}$ are solutions, then

$$L(\vec{x} + \vec{y}) = L\vec{x} + L\vec{y} = \vec{0} + \vec{0} = \vec{0}, \qquad \text{and} \qquad L(\alpha\vec{x}) = \alpha L\vec{x} = \alpha\vec{0} = \vec{0}.$$

So $\vec{x} + \vec{y}$ and $\alpha\vec{x}$ are solutions. The dimension of the kernel is called the *nullity* of the matrix.

The same sort of idea governs the solutions of linear differential equations. We try to describe the kernel of a linear differential operator, and as it is a subspace, we look for a basis of this kernel. Much of this book is dedicated to finding such bases.

The kernel of a matrix is the same as the kernel of its reduced row echelon form. For a matrix in reduced row echelon form, the kernel is rather easy to find. If a vector $\vec{x}$ is applied to a matrix L, then each entry in $\vec{x}$ corresponds to a column of L, the column that the entry multiplies. To find the kernel, pick a non-pivot column make a vector that has a -1 in the entry corresponding to this non-pivot column and zeros at all the other entries corresponding to the other non-pivot columns. Then for all the entries corresponding to pivot columns make it precisely the value in the corresponding row of the non-pivot column to make the vector be a solution to $L\vec{x} = \vec{0}$. This procedure is best understood by example.

Example 3.7.1: Consider

$$L = \begin{bmatrix} \boxed{1} & 2 & 0 & 0 & 3 \\ 0 & 0 & \boxed{1} & 0 & 4 \\ 0 & 0 & 0 & \boxed{1} & 5 \end{bmatrix}.$$

This matrix is in reduced row echelon form, the pivots are marked. There are two non-pivot columns, so the kernel has dimension 2, that is, it is the span of 2 vectors. Let us find the first vector. We look at the first non-pivot column, the 2^{nd} column, and we put a -1 in the 2^{nd} entry of our vector. We put a 0 in the 5^{th} entry as the 5^{th} column is also a non-pivot

column:

$$\begin{bmatrix} ? \\ -1 \\ ? \\ ? \\ 0 \end{bmatrix}.$$

Let us fill the rest. When this vector hits the first row, we get a -2 and 1 times whatever the first question mark is. So make the first question mark 2. For the second and third rows, it is sufficient to make it the question marks zero. We are really filling in the non-pivot column into the remaining entries. Let us check while marking which numbers went where:

$$\begin{bmatrix} 1 & \boxed{2} & 0 & 0 & 3 \\ 0 & \boxed{0} & 1 & 0 & 4 \\ 0 & \boxed{0} & 0 & 1 & 5 \end{bmatrix} \begin{bmatrix} \boxed{2} \\ -1 \\ \boxed{0} \\ \boxed{0} \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Yay! How about the second vector. We start with

$$\begin{bmatrix} ? \\ 0 \\ ? \\ ? \\ -1. \end{bmatrix}$$

We set the first question mark to 3, the second to 4, and the third to 5. Let us check, marking things as previously,

$$\begin{bmatrix} 1 & 2 & 0 & 0 & \boxed{3} \\ 0 & 0 & 1 & 0 & \boxed{4} \\ 0 & 0 & 0 & 1 & \boxed{5} \end{bmatrix} \begin{bmatrix} \boxed{3} \\ 0 \\ \boxed{4} \\ \boxed{5} \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

There are two non-pivot columns, so we only need two vectors. We have found the basis of the kernel. So,

$$\text{kernel of } L = \text{span} \left\{ \begin{bmatrix} 2 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \\ 4 \\ 5 \\ -1 \end{bmatrix} \right\}$$

What we did in finding a basis of the kernel is we expressed all solutions of $L\vec{x} = \vec{0}$ as a linear combination of some given vectors.

The procedure to find the basis of the kernel of a matrix L:

(i) Find the reduced row echelon form of L.

(ii) Write down the basis of the kernel as above, one vector for each non-pivot column.

The rank of a matrix is the dimension of the column space, and that is the span on the pivot columns, while the kernel is the span of vectors one for each non-pivot column. So the two numbers must add to the number of columns.

Theorem 3.7.1 (Rank–Nullity)

If a matrix A has n columns, rank r, and nullity k (dimension of the kernel), then

$$n = r + k.$$

The theorem is immensely useful in applications. It allows one to compute the rank r if one knows the nullity k and vice versa, without doing any extra work.

Let us consider an example application, a simple version of the so-called *Fredholm alternative*. A similar result is true for differential equations. Consider

$$A\vec{x} = \vec{b},$$

where A is a square $n \times n$ matrix. There are then two mutually exclusive possibilities:

(i) A nonzero solution $\vec{x}$ to $A\vec{x} = \vec{0}$ exists.

(ii) The equation $A\vec{x} = \vec{b}$ has a unique solution $\vec{x}$ for every $\vec{b}$.

How does the Rank–Nullity theorem come into the picture? Well, if A has a nonzero solution $\vec{x}$ to $A\vec{x} = \vec{0}$, then the nullity k is positive. But then the rank $r = n - k$ must be less than n. In particular it means that the column space of A is of dimension less than n, so it is a subspace that does not include everything in $\mathbb{R}^n$. So $\mathbb{R}^n$ has to contain some vector $\vec{b}$ not in the column space of A. In fact, most vectors in $\mathbb{R}^n$ are not in the column space of A.

The idea of a kernel also comes up when defining and discussing eigenvectors. In order to find this vector, we are looking for a vector $\vec{v}$ so that

$$(A - \lambda I)\vec{v} = \vec{0}.$$

This means that we are looking for a vector $\vec{v}$ that is in the kernel of the matrix $(A - \lambda I)$. Since the kernel is also a subspace, this means that the set of all eigenvectors of a matrix A with a certain eigenvalue is a subspace, so it has a dimension. This dimension is number of linearly independent eigenvectors with that eigenvalue, so it is the geometric multiplicity of this eigenvalue. This also motivates why this is sometimes called the *eigenspace* for a given eigenvalue. Finding a basis of this subspace (which is also finding the kernel of the matrix $A - \lambda I$) is the exact same as the process of finding the eigenvectors of the matrix A.

3.7.2 Computing the inverse

If the matrix A is square and there exists a unique solution $\vec{x}$ to $A\vec{x} = \vec{b}$ for any $\vec{b}$ (there are no free variables), then A is invertible.

In particular, if $A\vec{x} = \vec{b}$ then $\vec{x} = A^{-1}\vec{b}$. Now we just need to compute what A^{-1} is. We can surely do elimination every time we want to find $A^{-1}\vec{b}$, but that would be ridiculous. The mapping A^{-1} is linear and hence given by a matrix, and we have seen that to figure out the matrix we just need to find where does A^{-1} take the standard basis vectors $\vec{e}_1$, $\vec{e}_2$, $\ldots$, $\vec{e}_n$.

That is, to find the first column of A^{-1} we solve $A\vec{x} = \vec{e}_1$, because then $A^{-1}\vec{e}_1 = \vec{x}$. To find the second column of A^{-1} we solve $A\vec{x} = \vec{e}_2$. And so on. It is really just n eliminations that we need to do. But it gets even easier. If you think about it, the elimination is the same for everything on the left side of the augmented matrix. Doing n eliminations separately we would redo most of the computations. Best is to do all at once.

Therefore, to find the inverse of A, we write an $n \times 2n$ augmented matrix $[\, A \mid I \,]$, where I is the identity matrix, whose columns are precisely the standard basis vectors. We then perform row reduction until we arrive at the reduced row echelon form. If A is invertible, then pivots can be found in every column of A, and so the reduced row echelon form of $[\, A \mid I \,]$ looks like $[\, I \mid A^{-1} \,]$. We then just read off the inverse A^{-1}. If you do not find a pivot in every one of the first n columns of the augmented matrix, then A is not invertible.

This is best seen by example.

Example 3.7.2: Find the inverse of the matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{bmatrix}.$$

Solution: We write the augmented matrix and we start reducing:

$$\left[\begin{array}{ccc|ccc} \boxed{1} & 2 & 3 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 & 1 & 0 \\ 3 & 1 & 0 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{ccc|ccc} \boxed{1} & 2 & 3 & 1 & 0 & 0 \\ 0 & -4 & -5 & -2 & 1 & 0 \\ 0 & -5 & -9 & -3 & 0 & 1 \end{array}\right] \to \left[\begin{array}{ccc|ccc} \boxed{1} & 2 & 3 & 1 & 0 & 0 \\ 0 & \boxed{1} & 5/4 & 1/2 & 1/4 & 0 \\ 0 & -5 & -9 & -3 & 0 & 1 \end{array}\right] \to$$

$$\to \left[\begin{array}{ccc|ccc} \boxed{1} & 2 & 3 & 1 & 0 & 0 \\ 0 & \boxed{1} & 5/4 & 1/2 & 1/4 & 0 \\ 0 & 0 & -11/4 & -1/2 & -5/4 & 1 \end{array}\right] \to \left[\begin{array}{ccc|ccc} \boxed{1} & 2 & 3 & 1 & 0 & 0 \\ 0 & \boxed{1} & 5/4 & 1/2 & 1/4 & 0 \\ 0 & 0 & \boxed{1} & 2/11 & 5/11 & -4/11 \end{array}\right] \to$$

$$\to \left[\begin{array}{ccc|ccc} \boxed{1} & 2 & 0 & 5/11 & -5/11 & 12/11 \\ 0 & \boxed{1} & 0 & 3/11 & -9/11 & 5/11 \\ 0 & 0 & \boxed{1} & 2/11 & 5/11 & -4/11 \end{array}\right] \to \left[\begin{array}{ccc|ccc} \boxed{1} & 0 & 0 & -1/11 & 3/11 & 2/11 \\ 0 & \boxed{1} & 0 & 3/11 & -9/11 & 5/11 \\ 0 & 0 & \boxed{1} & 2/11 & 5/11 & -4/11 \end{array}\right].$$

So

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{bmatrix}^{-1} = \begin{bmatrix} -1/11 & 3/11 & 2/11 \\ 3/11 & -9/11 & 5/11 \\ 2/11 & 5/11 & -4/11 \end{bmatrix}.$$

Not too terrible, no? Perhaps harder than inverting a 2×2 matrix for which we had a formula, but not too bad. Really in practice this is done efficiently by a computer.

3.7.3 Trace and Determinant of Matrices

The next thing to add into our toolbox of matrices is the idea of the trace of a matrix, and how it and the determinant relate to the eigenvalues of said matrix.

Definition 3.7.1

Let A be an $n \times n$ square matrix. The *trace* of A is the sum of all diagonal entries of A.

For example, if we have the matrix

$$\begin{bmatrix} 1 & 4 & -2 \\ 3 & 2 & 5 \\ 0 & 1 & 3 \end{bmatrix}$$

the trace is $1 + 2 + 3 = 6$.

The trace is important in our context because it also tells us something about the eigenvalues of a matrix. To work this out, let's consider the generic 2×2 matrix and how we would find the eigenvalues. If we have a 2×2 matrix of the form

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

we can write out the expression $\det(A - \lambda I)$ in order to find the eigenvalues. In this case, we would get

$$\det(A - \lambda I) = \det\left(\begin{bmatrix} a - \lambda & b \\ c & d - \lambda \end{bmatrix}\right) = (a - \lambda)(d - \lambda) - bc = \lambda^2 - (a + d)\lambda + (ad - bc).$$

However, the coefficients in this polynomial look familiar. $(ad - bc)$ is just the determinant of the matrix A, and $a + d$ is the trace. Therefore, for any 2×2 matrix, we could write the characteristic polynomial as

$$\det(A - \lambda I) = \lambda^2 - T\lambda + D \tag{3.3}$$

where T is the trace of the matrix and D is the determinant. On the other hand, assume that r_1 and r_2 are the two eigenvalues of this matrix (whether they be real, complex, or repeated). In that case, we know that this polynomial has r_1 and r_2 as roots. Therefore, it is equal to

$$\det(A - \lambda I) = (\lambda - r_1)(\lambda - r_2) = \lambda^2 - (r_1 + r_2)\lambda + r_1 r_2. \tag{3.4}$$

Matching up the coefficient of λ and the constant term in (3.3) and (3.4) gives the relation that

$$T = r_1 + r_2 \qquad D = r_1 r_2,$$

that is, the trace of the matrix is the sum of the eigenvalues, and the determinant of the matrix is the product of the eigenvalues. We only showed this fact for 2×2 matrices, but it does hold for matrices of all sizes, giving us the following theorem.

Theorem 3.7.1

Let A be an $n \times n$ square matrix with eigenvalues $\lambda_1,\ \lambda_2,\ \ldots, \lambda_n$, written with multiplicity if needed. Then

(a) The trace of A is $\lambda_1 + \lambda_2 + \cdots + \lambda_n$.

(b) The determinant of A is $(\lambda_1)(\lambda_2)\cdots(\lambda_n)$.

From the above statement, we note that if any of the eigenvalues is zero, the product of all eigenvalues will be zero, and so the matrix will have zero determinant. This gives an extra follow-up fact, and addition to Theorem 3.5.4.

Theorem 3.7.2

A matrix A is invertible if and only if all of it's eigenvalues are non-zero.

Example 3.7.3: Use the facts above to analyze the eigenvalues of the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 5 & 4 \end{bmatrix}.$$

Solution: From the matrix A, we can compute that the trace of A is $1 + 4 = 5$, and the determinant is $(1)(4) - (2)(5) = -6$. Based on the theorem above, we know that the two eigenvalues of this matrix must add to 5 and multiply to -6. While you could probably guess the numbers here, the important take-aways from this example are what we can learn.

The main fact to point out is that this is enough information, in the 2×2 case, to tell us that the eigenvalues have to be real and distinct. Since their product is a negative number, we can eliminate the other two options. If we have two complex roots, they must be of the form $x + iy$ and $x - iy$, and so the product is

$$(x + iy)(x - iy) = x^2 + ixy - ixy - i^2y^2 = x^2 + y^2$$

which is always positive, no matter what x and y are. Similarly, if we have a repeated eigvalue, the product will be that number squared, which is also positive. Therefore, if the determinant of a 2×2 matrix is negative, the eigenvalues must be real and distinct, with one being positive and one negative (otherwise the product can not be negative). These facts will be important when we start to analyze the solutions to systems of differential equations in Chapter 4. ⌟

Example 3.7.4: What can be said about the eigenvalues of the matrix

$$A = \begin{bmatrix} 0 & -1 & 0 \\ 2 & 2 & 0 \\ -7 & -3 & -1 \end{bmatrix}?$$

Solution: We can find the same information as the previous example. The trace of A is 1, and the determinant, by cofactor expansion along column 3, is $(-1)(0 + 2) = -2$. Therefore, the sum of the *three* eigenvalues is 1, and the product of them is -2. We don't actually have enough information here to determine what the eigenvalues are. The issue is that with three eigenvalues, there are many different ways to get to a product being negative. There could be three negative eigenvalues, two positive and one negative, or one negative real with two complex eigenvalues. However, the one thing we do know for sure is that there must be one negative real eigenvalue. For this particular example, we can compute that the eigenvalues are -1, $1 + i$, and $1 - i$, so we did end up in the complex case. ⌟

***Exercise* 3.7.1:** *Imagine that we have a 3×3 matrix with a positive determinant (it doesn't matter what the trace is). Think about all the scenarios and verify that at least one eigenvalue must be real and positive for this to happen.*

3.7.4 Exercises

***Exercise* 3.7.2:** *For the following matrices, find a basis for the kernel (nullspace).*

a) $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 5 \\ 1 & 1 & -4 \end{bmatrix}$ *b)* $\begin{bmatrix} 2 & -1 & -3 \\ 4 & 0 & -4 \\ -1 & 1 & 2 \end{bmatrix}$ *c)* $\begin{bmatrix} -4 & 4 & 4 \\ -1 & 1 & 1 \\ -5 & 5 & 5 \end{bmatrix}$ *d)* $\begin{bmatrix} -2 & 1 & 1 & 1 \\ -4 & 2 & 2 & 2 \\ 1 & 0 & 4 & 3 \end{bmatrix}$

Exercise* 3.7.3: *For the following matrices, find a basis for the kernel (nullspace).*

a) $\begin{bmatrix} 2 & 6 & 1 & 9 \\ 1 & 3 & 2 & 9 \\ 3 & 9 & 0 & 9 \end{bmatrix}$ *b)* $\begin{bmatrix} 2 & -2 & -5 \\ -1 & 1 & 5 \\ -5 & 5 & -3 \end{bmatrix}$ *c)* $\begin{bmatrix} 1 & -5 & -4 \\ 2 & 3 & 5 \\ -3 & 5 & 2 \end{bmatrix}$ *d)* $\begin{bmatrix} 0 & 4 & 4 \\ 0 & 1 & 1 \\ 0 & 5 & 5 \end{bmatrix}$

***Exercise* 3.7.4:** *Suppose a 5×5 matrix A has rank 3. What is the nullity?*

***Exercise* 3.7.5:** *Consider a square matrix A, and suppose that $\vec{x}$ is a nonzero vector such that $A\vec{x} = \vec{0}$. What does the Fredholm alternative say about invertibility of A?*

***Exercise* 3.7.6:** *Consider*

$$M = \begin{bmatrix} 1 & 2 & 3 \\ 2 & ? & ? \\ -1 & ? & ? \end{bmatrix}.$$

If the nullity of this matrix is 2, fill in the question marks. Hint: What is the rank?

Exercise* 3.7.7: *Suppose the column space of a 9×5 matrix A of dimension 3. Find*

a) Rank of A.

b) Nullity of A.

c) Dimension of the row space of A.

d) Dimension of the nullspace of A.

e) Size of the maximum subset of linearly independent rows of A.

***Exercise* 3.7.8:** *Compute the inverse of the given matrices*

a) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ *b)* $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ *c)* $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix}$

Exercise* 3.7.9: *Compute the inverse of the given matrices*

a) $\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ *b)* $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ *c)* $\begin{bmatrix} 2 & 4 & 0 \\ 2 & 2 & 3 \\ 2 & 4 & 1 \end{bmatrix}$

***Exercise* 3.7.10:** *By computing the inverse, solve the following systems for $\vec{x}$.*

a) $\begin{bmatrix} 4 & 1 \\ -1 & 3 \end{bmatrix} \vec{x} = \begin{bmatrix} 13 \\ 26 \end{bmatrix}$

b) $\begin{bmatrix} 3 & 3 \\ 3 & 4 \end{bmatrix} \vec{x} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$

Exercise 3.7.11:* *By computing the inverse, solve the following systems for $\vec{x}$.*

a) $\begin{bmatrix} -1 & 1 \\ 3 & 3 \end{bmatrix} \vec{x} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$ *b)* $\begin{bmatrix} 2 & 7 \\ 1 & 6 \end{bmatrix} \vec{x} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$

Exercise 3.7.12:* *For each of the following matrices below:*

a) Compute the trace and determinant of the matrix, and

b) Find the eigenvalues of the matrix and verify that the trace is the sum of the eigenvalues and the determinant is the product.

(i) $\begin{bmatrix} -4 & 2 \\ -9 & 5 \end{bmatrix}$ (ii) $\begin{bmatrix} 2 & -3 \\ 6 & -4 \end{bmatrix}$ (iii) $\begin{bmatrix} -10 & -12 \\ 6 & 8 \end{bmatrix}$. (iv) $\begin{bmatrix} -7 & -9 \\ 1 & -1 \end{bmatrix}$

Exercise 3.7.13:* *For each of the following matrices below:*

a) Compute the trace and determinant of the matrix, and

b) Find the eigenvalues of the matrix and verify that the trace is the sum of the eigenvalues and the determinant is the product.

(i) $\begin{bmatrix} -1 & -16 & -4 \\ 1 & 6 & 1 \\ -2 & -4 & 1 \end{bmatrix}$ (ii) $\begin{bmatrix} 1 & 2 & 0 \\ -12 & -13 & -4 \\ 16 & 14 & 3 \end{bmatrix}$ (iii) $\begin{bmatrix} 10 & -7 & -14 \\ 0 & 5 & 6 \\ 7 & -8 & -14 \end{bmatrix}$

Chapter 4

Systems of ODEs

4.1 Introduction to systems of ODEs

Attribution: [JL], §3.1.

Learning Objectives

After this section, you will be able to:

- Classify the order and number of components in a system of differential equations,
- Verify if a set of functions solves a system of differential equations, and
- Write a system of differential equations to fit a physical situation.

4.1.1 Systems

Often we do not have just one dependent variable and one equation. For instance, we may be looking at multiple populations that are changing over time, or watching how the amount of support for multiple candidates develops leading up to an election. And as we will see, we may end up with systems of several equations and several dependent variables even if we start with a single equation.

If we have several dependent variables, suppose y_1, y_2, ..., y_n, then we can have a differential equation involving all of them and their derivatives with respect to one independent variable x. For example, $y_1'' = f(y_1', y_2', y_1, y_2, x)$. Usually, when we have two dependent variables we have two equations such as

$$\begin{aligned} y_1'' &= f_1(y_1', y_2', y_1, y_2, x), \\ y_2'' &= f_2(y_1', y_2', y_1, y_2, x), \end{aligned}$$

for some functions f_1 and f_2. We call the above a *system of differential equations*. More precisely, the above is a *second order system* of ODEs as second order derivatives appear.

The system

$$\begin{aligned} x_1' &= g_1(x_1, x_2, x_3, t), \\ x_2' &= g_2(x_1, x_2, x_3, t), \\ x_3' &= g_3(x_1, x_2, x_3, t), \end{aligned}$$

is a *first order system*, where x_1, x_2, x_3 are the dependent variables, and t is the independent variable.

The terminology for systems is essentially the same as for single equations. For the system above, a *solution* is a set of three functions $x_1(t)$, $x_2(t)$, $x_3(t)$, such that

$$\begin{aligned} x_1'(t) &= g_1\big(x_1(t), x_2(t), x_3(t), t\big), \\ x_2'(t) &= g_2\big(x_1(t), x_2(t), x_3(t), t\big), \\ x_3'(t) &= g_3\big(x_1(t), x_2(t), x_3(t), t\big). \end{aligned}$$

In order to verify that something is a solution, we plug the different components into the solution to see that all of the equations are satisfied; if any one of the equations is not satisfied, then this set of functions is not a solution. We usually also have an *initial condition*. Just like for single equations we specify x_1, x_2, and x_3 for some fixed t. For example, $x_1(0) = a_1$, $x_2(0) = a_2$, $x_3(0) = a_3$ for some constants a_1, a_2, and a_3. For the second order system we would also specify the first derivatives at that same initial time point. And if we find a solution with constants in it, where by solving for the constants we find a solution for any initial condition, we call this solution the *general solution*. Best to look at a simple example.

Example 4.1.1: Sometimes a system is easy to solve by solving for one variable and then for the second variable. Take the first order system

$$\begin{aligned} y_1' &= y_1, \\ y_2' &= y_1 - y_2, \end{aligned}$$

with y_1, y_2 as the dependent variables and x as the independent variable. Consider initial conditions $y_1(0) = 1$, $y_2(0) = 2$ and solve the initial value problem.

Solution: We note that $y_1 = C_1 e^x$ is the general solution of the first equation, which we can get because this equation does not involve y_2 at all and we can get a solution via our normal first order equation methods. We then plug this y_1 into the second equation and get the equation $y_2' = C_1 e^x - y_2$, which is a linear first order equation that is easily solved for y_2. By the method of integrating factor we get

$$e^x y_2 = \frac{C_1}{2} e^{2x} + C_2,$$

or $y_2 = \frac{C_1}{2} e^x + C_2 e^{-x}$. The general solution to the system is, therefore,

$$y_1 = C_1 e^x, \qquad y_2 = \frac{C_1}{2} e^x + C_2 e^{-x}.$$

We solve for C_1 and C_2 given the initial conditions. We substitute $x = 0$ and find that $C_1 = 1$ and $C_2 = 3/2$. Thus the solution is $y_1 = e^x$, and $y_2 = (1/2)e^x + (3/2)e^{-x}$.

Generally, we will not be so lucky to be able to solve for each variable separately as in the example above, and we will have to solve for all variables at once. While we won't generally be able to solve for one variable and then the next, we will try to salvage as much as possible from this technique. It will turn out that in a certain sense we will still (try to) solve a bunch of single equations and put their solutions together. Let's not worry right now about how to solve systems yet.

We will mostly consider *linear systems*. The example above is an example of a *linear first order system*. It is linear as none of the dependent variables or their derivatives appear in nonlinear functions or with powers higher than one (x, y, x' and y', constants, and functions of t can appear, but not xy or $(y')^2$ or x^3). A more complicated example of a second order linear system is

$$\begin{aligned} y_1'' &= e^t y_1' + t^2 y_1 + 5y_2 + \sin(t), \\ y_2'' &= t y_1' - y_2' + 2y_1 + \cos(t). \end{aligned}$$

4.1.2 Applications

Let us consider some simple applications of systems and how to set up the equations.

Example 4.1.2: First, we consider salt and brine tanks, but this time water flows from one to the other and back. We again consider that the tanks are well-mixed.

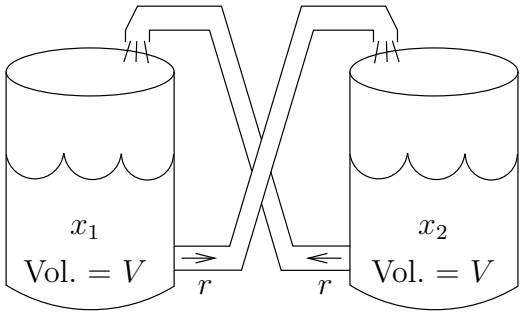


Figure 4.1: A closed system of two brine tanks.

Suppose we have two tanks, each containing volume V liters of salt brine. The amount of salt in the first tank is x_1 grams, and the amount of salt in the second tank is x_2 grams. The liquid is perfectly mixed and flows at the rate r liters per second out of each tank into the other. See Figure 4.1.

Solution: The rate of change of x_1, that is x_1', is the rate of salt coming in minus the rate going out. The rate coming in is the density of the salt in tank 2, that is $\frac{x_2}{V}$, times the rate r. The rate coming out is the density of the salt in tank 1, that is $\frac{x_1}{V}$, times the rate r. In other words it is

$$x_1' = \frac{x_2}{V} r - \frac{x_1}{V} r = \frac{r}{V} x_2 - \frac{r}{V} x_1 = \frac{r}{V}(x_2 - x_1).$$

Similarly we find the rate x_2', where the roles of x_1 and x_2 are reversed. All in all, the system of ODEs for this problem is

$$\begin{aligned} x_1' &= \frac{r}{V}(x_2 - x_1), \\ x_2' &= \frac{r}{V}(x_1 - x_2). \end{aligned}$$

In this system we cannot solve for x_1 or x_2 separately. We must solve for both x_1 and x_2 at once, which is intuitively clear since the amount of salt in one tank affects the amount in the other. We can't know x_1 before we know x_2, and vice versa.

We don't yet know how to find all the solutions, but intuitively we can at least find some solutions. Suppose we know that initially the tanks have the same amount of salt. That is, we have an initial condition such as $x_1(0) = x_2(0) = C$. Then clearly the amount of salt coming and out of each tank is the same, so the amounts are not changing. In other words, $x_1 = C$ and $x_2 = C$ (the constant functions) is a solution: $x_1' = x_2' = 0$, and $x_2 - x_1 = x_1 - x_2 = 0$, so the equations are satisfied.

Let us think about the setup a little bit more without solving it. Suppose the initial conditions are $x_1(0) = A$ and $x_2(0) = B$, for two different constants A and B. Since no salt is coming in or out of this closed system, the total amount of salt is constant. That is, $x_1 + x_2$ is constant, and so it equals $A + B$. Intuitively if A is bigger than B, then more salt will flow out of tank one than into it. Eventually, after a long time we would then expect the amount of salt in each tank to equalize. In other words, the solutions of both x_1 and x_2 should tend towards $\frac{A+B}{2}$. Once you know how to solve systems you will find out that this really is so. ⌟

Example 4.1.3: Another example that showcases how systems work is different ways that populations of animals can interact. There are two main interactions that we will consider. The first of these is of two "competing species." The idea here is that there are two species that are trying to coexist in a given area. On their own (without the other species), each one would grow exponentially, but any interaction between the two species is negative for both of them, because they share the types of food and other resources that they need to survive and grow. This gives rise to a system of differential equations of the form

$$\begin{aligned} \frac{dx_1}{dt} &= ax_1 - bx_1x_2 \\ \frac{dx_2}{dt} &= cx_2 - dx_1x_2. \end{aligned}$$

In the system here, the coefficient a represents the growth rate of species 1 on it's own, b represents the amount to which the competition for resources affects the growth rate of species 1, c represents the growth rate of species 2, and d represents the magnitude of how the competition affects the growth of species 2. This type of system can also be written to contain logistic growth terms for the two species, resulting in

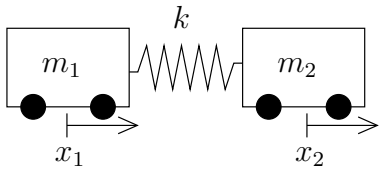
$$\begin{aligned} \frac{dx_1}{dt} &= ax_1(K_1 - x_1) - bx_1x_2 \\ \frac{dx_2}{dt} &= cx_2(K_2 - x_2) - dx_1x_2. \end{aligned}$$

The other main population model to consider is a "predator-prey" interaction. The key components of this model are that the prey population will grow on it's own and the interaction between the two populations is negative, because the presence of predator population will cause the prey population to decrease. On the other hand, the predator population will die off on it's own (without a food source) but the interaction with the prey population causes the predator population to increase. This gives rise to the system of differential equations

$$\begin{aligned} \frac{dx}{dt} &= ax - bxy \\ \frac{dy}{dt} &= -cy + dxy \end{aligned}$$

where x is the prey population and y is the predator population. We will take another look at both of these examples in § 5.3 once we have more terminology and techniques to discuss them.

Example 4.1.4: Let us look at a second order example. We return to the mass and spring setup, but this time we consider two masses.



Consider one spring with constant k and two masses m_1 and m_2. Think of the masses as carts that ride along a straight track with no friction. Let x_1 be the displacement of the first cart and x_2 be the displacement of the second cart. That is, we put the two carts somewhere with no tension on the spring, and we mark the position of the first and second cart and call those the zero positions. Then x_1 measures how far the first cart is from its zero position, and x_2 measures how far the second cart is from its zero position. The force exerted by the spring on the first cart is $k(x_2 - x_1)$, since $x_2 - x_1$ is how far the string is stretched (or compressed) from the rest position. The force exerted on the second cart is the opposite, thus the same thing with a negative sign. Newton's second law states that force equals mass times acceleration. So the system of equations is

$$\begin{aligned} m_1 x_1'' &= k(x_2 - x_1), \\ m_2 x_2'' &= -k(x_2 - x_1). \end{aligned}$$

Again, we cannot solve for the x_1 or x_2 variable separately. That we must solve for both x_1 and x_2 at once is intuitively clear, since where the first cart goes depends on exactly where the second cart goes and vice versa.

4.1.3 Changing to first order

Before we talk about how to handle systems, let us note that in some sense we need only consider first order systems. Let us take an n^{th} order differential equation

$$y^{(n)} = F(y^{(n-1)}, \ldots, y', y, x)$$

that we would like to convert into a first order system. To do this, we first consider what a first order system would look like. A first order system consists of a set of equations involving

the derivative of each of our variables. Let's start with the first variable $u_1 = y$. What is the derivative of y? Well, it's y', and we don't have a way to represent this in terms of our variables (u_1) without any derivatives. So, we add a new variable u_2 that we define to be y', which makes the first equation in our system $u_1' = u_2$.

Well, now we have u_2 and we need to determine what its derivative is. Since $u_2 = y'$, $u_2' = y''$. If the order of the equation n is 2, we then have an equation to define what y'' is in terms of y', y, and x, which are u_2, u_1, and x in our new system. If that's the case, we're done, and if not, we need to define a new variable $u_3 = y''$ so that $u_2' = u_3$. We can continue this process over and over again.

When do we stop? As illustrated in the previous example with $n = 2$, we stop when our derivative u_n' is the nth derivative of y. This works because our equation tells us exactly what $y^{(n)}$ is in terms of lower order terms, which we have already defined variables for. Thus, we define new variables $u_1, u_2, \ldots, u_n$ and write the system

$$\begin{aligned} u_1' &= u_2, \\ u_2' &= u_3, \\ &\vdots \\ u_{n-1}' &= u_n, \\ u_n' &= F(u_n, u_{n-1}, \ldots, u_2, u_1, x). \end{aligned}$$

We solve this system for u_1, u_2, $\ldots$, u_n. Once we have solved for the u's, we can discard u_2 through u_n and let $y = u_1$. This y solves the original equation.

Example 4.1.5: Take $x''' = 2x'' + 8x' + x + t$. Convert this equation into a first order system.

Solution: Letting $u_1 = x$, $u_2 = x'$, $u_3 = x''$, we find the system:

$$u_1' = u_2, \qquad u_2' = u_3, \qquad u_3' = 2u_3 + 8u_2 + u_1 + t.$$

Since this is a linear system, we can also write this in matrix-vector form, which will be useful for systems that we will analyze later. To do this, we define a vector $\vec{u}$ as

$$\vec{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}.$$

Then, we know that

$$\vec{u}\,' = \begin{bmatrix} u_1' \\ u_2' \\ u_3' \end{bmatrix} = \begin{bmatrix} u_2 \\ u_3 \\ 2u_3 + 8u_2 + u_1 + t \end{bmatrix}.$$

We want to rewrite this equation using the vector $\vec{u}$ and a matrix. We can rewrite this last vector as

$$\begin{bmatrix} u_1' \\ u_2' \\ u_3' \end{bmatrix} = \begin{bmatrix} u_2 \\ u_3 \\ u_1 + 8u_2 + 2u_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ t \end{bmatrix}$$

and the right-hand side of this equation can be written as

$$\vec{u}\,' = \begin{bmatrix} u_2 \\ u_3 \\ u_1 + 8u_2 + 2u_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ t \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 8 & 2 \end{bmatrix} \vec{u} + \begin{bmatrix} 0 \\ 0 \\ t \end{bmatrix}.$$

(Verify that the matrix multiplication works out here!) Therefore, we can write this first order system as

$$\vec{u}\,' = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 8 & 2 \end{bmatrix} \vec{u} + \begin{bmatrix} 0 \\ 0 \\ t \end{bmatrix}.$$

Note that if the equation above was non-linear, it would not be possible to write the system version in an appropriate matrix form. It is also important to know how to take initial conditions into account with these problems.

Example 4.1.6: Convert the initial value problem

$$x''' = 4e^t x'' - 3(x')^2 + t^2 \sin(x) + (t^2 + 1) \qquad x(0) = 2,\ x'(0) = -1,\ x''(0) = 4$$

into a system of first order equations. Simplify the expression as much as possible.

Solution: We follow the same procedure as the previous example. We define variables u_1, u_2, u_3 as

$$u_1 = x \quad u_2 = x' \quad u_3 = x''$$

so that we have the differential equations

$$u_1' = u_2 \quad u_2' = u_3 \quad u_3' = x''' = 4e^t u_3 - 3u_2^2 + t^2 \sin(u_1) + (t^2 + 1)$$

which we can write in vector form as

$$\vec{u}' = \begin{bmatrix} u_1' \\ u_2' \\ u_3' \end{bmatrix} = \begin{bmatrix} u_2 \\ u_3 \\ 4e^t u_3 - 3u_2^2 + t^2 \sin(u_1) + (t^2 + 1) \end{bmatrix}.$$

We would now want to try to convert this into matrix form. However, the matrix that we come up with should not depend on u at all. In this case, it would mean that we want to write this equation as

$$\vec{u}\,' = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ t^2 + 1 \end{bmatrix}$$

since the extra term needs to be everything that does not depend on u. However, while we can determine the first two rows of the matrix, we can not determine the last row. There is no way to pick terms *independent of* u to fill in the three stars in the bottom row in order to make the bottom term in the matrix-vector product to equal $4e^t u_3 - 3u_2^2 + t^2 \sin(u_1)$. The issue here is that the equation is non-linear; the u_2^2 term and the $\sin(u_1)$ term can not be

written in this way. Therefore, the best we can do is the vector form, and it can't be written in matrix form.

The last thing we need to deal with is the initial conditions. Since the conditions say that

$$x(0) = 2,\ x'(0) = -1,\ x''(0) = 4$$

and we have that $u_1 = x,\ u_2 = x',\ u_3 = x''$, this means that the initial condition should be

$$u_1(0) = 2,\ u_2(0) = -1,\ u_3(0) = 4,$$

or

$$\vec{u}(0) = \begin{bmatrix} 2 \\ -1 \\ 4 \end{bmatrix}.$$

Thus, the full way to write this initial value problem in system form is

$$\vec{u}' = \begin{bmatrix} u_2 \\ u_3 \\ 4e^t u_3 - 3u_2^2 + t^2 \sin(u_1) + (t^2+1) \end{bmatrix} \qquad \vec{u}(0) = \begin{bmatrix} 2 \\ -1 \\ 4 \end{bmatrix}.$$

A similar process can be followed for a system of higher order differential equations. For example, a system of k differential equations in k unknowns, all of order n, can be transformed into a first order system of $n \times k$ equations and $n \times k$ unknowns.

Example 4.1.7: Consider the system from the carts example,

$$m_1 x_1'' = k(x_2 - x_1), \qquad m_2 x_2'' = -k(x_2 - x_1).$$

Let $u_1 = x_1$, $u_2 = x_1'$, $u_3 = x_2$, $u_4 = x_2'$. The second order system becomes the first order system

$$u_1' = u_2, \qquad m_1 u_2' = k(u_3 - u_1), \qquad u_3' = u_4, \qquad m_2 u_4' = -k(u_3 - u_1).$$

Example 4.1.8: The idea works in reverse as well. Consider the system

$$x' = 2y - x, \qquad y' = x,$$

where the independent variable is t. We wish to solve for the initial conditions $x(0) = 1$, $y(0) = 0$.

Solution: If we differentiate the second equation, we get $y'' = x'$. We know what x' is in terms of x and y, and we know that $x = y'$. So,

$$y'' = x' = 2y - x = 2y - y'.$$

We now have the equation $y'' + y' - 2y = 0$. We know how to solve this equation and we find that $y = C_1 e^{-2t} + C_2 e^t$. Once we have y, we use the equation $y' = x$ to get x.

$$x = y' = -2C_1 e^{-2t} + C_2 e^t.$$

We solve for the initial conditions $1 = x(0) = -2C_1 + C_2$ and $0 = y(0) = C_1 + C_2$. Hence, $C_1 = -C_2$ and $1 = 3C_2$. So $C_1 = -1/3$ and $C_2 = 1/3$. Our solution is

$$x = \frac{2e^{-2t} + e^t}{3}, \qquad y = \frac{-e^{-2t} + e^t}{3}.$$

□

***Exercise* 4.1.1:** *Plug in and check that this really is the solution.*

It is useful to go back and forth between systems and higher order equations for other reasons. For example, software for solving ODE numerically (approximation) is generally for first order systems. So to use it, you have to take whatever ODE you want to solve and convert it to a first order system. In fact, it is not very hard to adapt computer code for the Euler or Runge–Kutta method for first order equations to handle first order systems. We essentially just treat the dependent variable not as a number but as a vector. In many mathematical computer languages there is almost no distinction in syntax.

4.1.4 Autonomous systems and vector fields

A system where the equations do not depend on the independent variable is called an *autonomous system*. For example the system $y' = 2y - x$, $y' = x$ is autonomous as t is the independent variable but does not appear in the equations.

For autonomous systems we can draw the so-called *direction field* or *vector field*, a plot similar to a slope field, but instead of giving a slope at each point, we give a direction (and a magnitude). The previous example, $x' = 2y - x$, $y' = x$, says that at the point (x, y) the direction in which we should travel to satisfy the equations should be the direction of the vector $(2y - x, x)$ with the speed equal to the magnitude of this vector. So we draw the vector $(2y - x, x)$ at the point (x, y) and we do this for many points on the xy-plane. For example, at the point $(1, 2)$ we draw the vector $\big(2(2) - 1, 1\big) = (3, 1)$, a vector pointing to the right and a little bit up, while at the point $(2, 1)$ we draw the vector $\big(2(1) - 2, 2\big) = (0, 2)$ a vector that points straight up. When drawing the vectors, we will scale down their size to fit many of them on the same direction field. We are mostly interested in their direction and relative size. See Figure 4.2 on the following page.

We can draw a path of the solution in the plane. Suppose the solution is given by $x = f(t)$, $y = g(t)$. We pick an interval of t (say $0 \le t \le 2$ for our example) and plot all the points $\big(f(t), g(t)\big)$ for t in the selected range. The resulting picture is called the *phase portrait* (or phase plane portrait). The particular curve obtained is called the *trajectory* or *solution curve*. See an example plot in Figure 4.3 on the next page. In the figure the solution starts at $(1, 0)$ and travels along the vector field for a distance of 2 units of t. We solved this system precisely, so we compute $x(2)$ and $y(2)$ to find $x(2) \approx 2.475$ and $y(2) \approx 2.457$. This point corresponds to the top right end of the plotted solution curve in the figure.

Notice the similarity to the diagrams we drew for autonomous systems in one dimension. But note how much more complicated things become when we allow just one extra dimension.

We can draw phase portraits and trajectories in the xy-plane even if the system is not autonomous. In this case however we cannot draw the direction field, since the field changes as t changes. For each t we would get a different direction field.

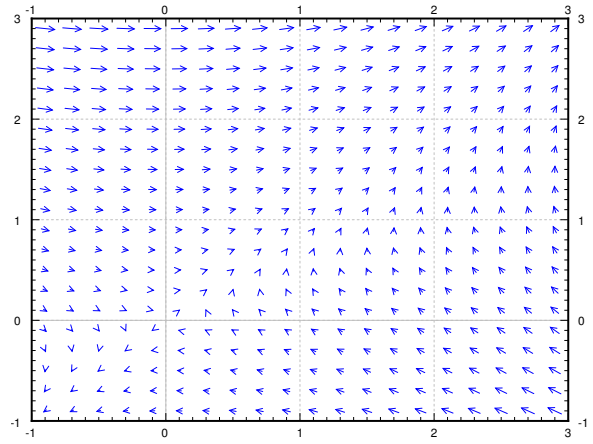


Figure 4.2: The direction field for $x' = 2y - x$, $y' = x$.

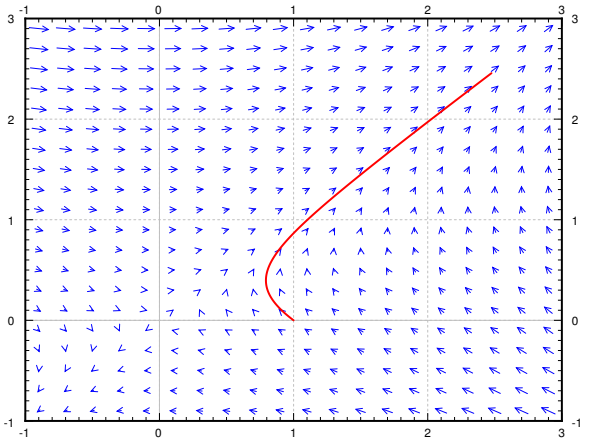


Figure 4.3: The direction field for $x' = 2y - x$, $y' = x$ with the trajectory of the solution starting at $(1, 0)$ for $0 \le t \le 2$.

4.1.5 Picard's theorem

Perhaps before going further, let us mention that Picard's theorem on existence and uniqueness still holds for systems of ODE. Let us restate this theorem in the setting of systems. A general first order system is of the form

$$\begin{aligned} x_1' &= F_1(x_1, x_2, \dots, x_n, t), \\ x_2' &= F_2(x_1, x_2, \dots, x_n, t), \\ &\vdots \\ x_n' &= F_n(x_1, x_2, \dots, x_n, t). \end{aligned} \tag{4.1}$$

Theorem 4.1.1 (Picard's theorem on existence and uniqueness for systems)

If for every $j = 1, 2, \dots, n$ and every $k = 1, 2, \dots, n$ each F_j is continuous and the derivative $\frac{\partial F_j}{\partial x_k}$ exists and is continuous near some $(x_1^0, x_2^0, \dots, x_n^0, t^0)$, then a solution to (4.1) subject to the initial condition $x_1(t^0) = x_1^0$, $x_2(t^0) = x_2^0$, $\dots$, $x_n(t^0) = x_n^0$ exists (at least for some small interval of t's) and is unique.

That is, a unique solution exists for any initial condition given that the system is reasonable (F_j and its partial derivatives in the x variables are continuous). As for single equations we may not have a solution for all time t, but at least for some short period of time.

As we can change any nth order ODE into a first order system, then we notice that this theorem provides also the existence and uniqueness of solutions for higher order equations that we have until now not stated explicitly.

4.1.6 Exercises

***Exercise* 4.1.2:** *Verify that* $x_1(t) = 2e^{-t} - 2e^{-2t}$, $x_2(t) = e^{-t} - 2e^{-2t}$ *solves the system* $x_1' = -2x_2$, $x_2' = x_1 - 3x_2$.

***Exercise* 4.1.3:** *Verify that* $x_1(t) = -2te^{-3t} - 2e^{-3t}$, $x_2(t) = 2te^{-3t} + 3e^{-3t}$ *solves the system* $x_1' = -5x_1 - 2x_2$, $x_2' = 2x_1 - x_2$.

***Exercise* 4.1.4:** *Find the general solution of* $x_1' = x_2 - x_1 + t$, $x_2' = x_2$.

***Exercise* 4.1.5:** *Find the general solution of* $x_1' = 3x_1 - x_2 + e^t$, $x_2' = x_1$.

Exercise* 4.1.6: *Find the general solution to* $y_1' = 3y_1$, $y_2' = y_1 + y_2$, $y_3' = y_1 + y_3$.

Exercise* 4.1.7: *Solve* $y' = 2x$, $x' = x + y$, $x(0) = 1$, $y(0) = 3$.

***Exercise* 4.1.8:** *Write* $ay'' + by' + cy = f(x)$ *as a first order system of ODEs.*

***Exercise* 4.1.9:** *Write* $x'' + y^2y' - x^3 = \sin(t)$, $y'' + (x' + y')^2 - x = 0$ *as a first order system of ODEs.*

Exercise* 4.1.10: *Write* $x''' = x + t$ *as a first order system.*

Exercise* 4.1.11: *Write* $y_1'' + y_1 + y_2 = t$, $y_2'' + y_1 - y_2 = t^2$ *as a first order system.*

***Exercise* 4.1.12:** *Write* $y^{(4)} - t^2y''' + e^ty' - (2t + 1)y = \cos(t)$ *as a first order system.*

***Exercise* 4.1.13:** *Write the initial value problem*

$$y'' - 2xy' + 3y = \sin(x) \qquad y(0) = 1,\ y'(0) = -2$$

as an initial value problem for a first order system of ODEs. Make sure to indicate how the initial condition appears as a part of this problem.

***Exercise* 4.1.14:** *Write the initial value problem*

$$y'' - (y + 1)^2y' - e^{xy} = \cos(x) \qquad y(0) = -1,\ y'(0) = 5$$

as an initial value problem for a first order system of ODEs. Make sure to indicate how the initial condition appears as a part of this problem. Can this be written in matrix form? Why or why not?

***Exercise* 4.1.15:** *Write the initial value problem*

$$y^{(4)} + e^xy'' - 4\cos(x)y' + (x^2 + 1)y = \frac{1}{x - 3} \qquad y(0) = 2,\ y'(0) = -3,\ y''(0) = 0,\ y^{(3)}(0) = 1$$

as an initial value problem for a first order system of ODEs. Make sure to indicate how the initial condition appears as a part of this problem.

***Exercise* 4.1.16:** *Suppose two masses on carts on frictionless surface are at displacements* x_1 *and* x_2 *as in Example 4.1.4 on page 273. Suppose that a rocket applies force* F *in the positive direction on cart* x_1. *Set up the system of equations.*

Exercise* 4.1.17: *Suppose two masses on carts on frictionless surface are at displacements x_1 and x_2 as in Example 4.1.4 on page 273. Suppose initial displacement is $x_1(0) = x_2(0) = 0$, and initial velocity is $x_1'(0) = x_2'(0) = a$ for some number a. Use your intuition to solve the system, explain your reasoning.*

***Exercise* 4.1.18:** *Suppose the tanks are as in Example 4.1.2 on page 271, starting both at volume V, but now the rate of flow from tank 1 to tank 2 is r_1, and rate of flow from tank 2 to tank one is r_2. In particular, the volumes will now be changing. Set up the system of equations.*

Exercise* 4.1.19: *Suppose the tanks are as in Example 4.1.2 on page 271 except that clean water flows in at the rate s liters per second into tank 1, and brine flows out of tank 2 and into the sewer also at the rate of s liters per second.*

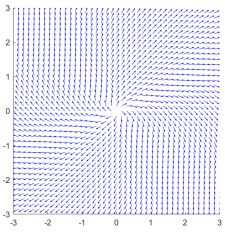
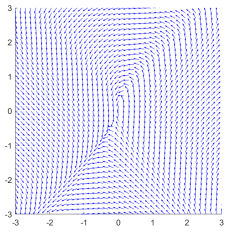
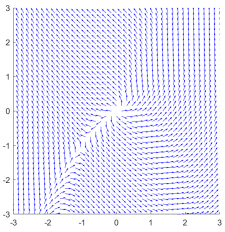
a) Draw the picture.

b) Set up the system of equations.

c) Intuitively, what happens as t goes to infinity, explain.

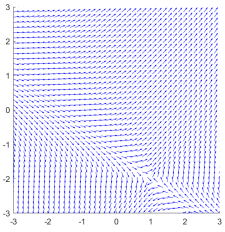
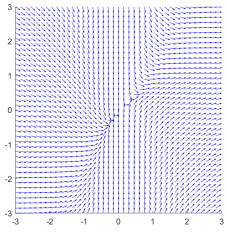
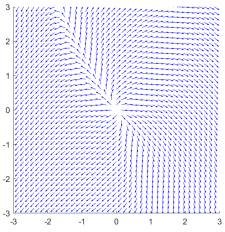
***Exercise* 4.1.20:** *Match the systems of differential equations below to their corresponding slope fields. Justify.*

$$(i)\ \begin{cases} \frac{dx}{dt} &= x + y \\ \frac{dy}{dt} &= 2y - x \end{cases} \qquad (ii)\ \begin{cases} \frac{dx}{dt} &= x - y \\ \frac{dy}{dt} &= x^2 + y \end{cases} \qquad (iii)\ \begin{cases} \frac{dx}{dt} &= x^2 - y^2 \\ \frac{dy}{dt} &= 3x - 1 \end{cases}$$

a)  b)  c) 

***Exercise* 4.1.21:** *Match the systems of differential equations below to their corresponding slope fields. Justify.*

$$(i)\ \begin{cases} \frac{dx}{dt} &= 2x + y \\ \frac{dy}{dt} &= y - x^2 \end{cases} \qquad (ii)\ \begin{cases} \frac{dx}{dt} &= x^2 \\ \frac{dy}{dt} &= x - y \end{cases} \qquad (iii)\ \begin{cases} \frac{dx}{dt} &= y + 2 \\ \frac{dy}{dt} &= x + y + 1 \end{cases}$$

a)  b)  c) 

4.2 Matrices and linear systems

Attribution: [JL], §3.2.

Learning Objectives

After this section, you will be able to:

- Define and perform addition and multiplication operations on matrices,
- Compute the determinant of a square matrix, and
- Find eigenvalues and eigenvectors of a square matrix.

This section is meant to summarize the parts of linear algebra that will be necessary in the process of developing and solving linear systems of differential equations. All of this information is covered in more detail in Chapter 3, so you can find more information there. If you went through that chapter already, this section will serve as a review.

4.2.1 Matrices and vectors

Before we start talking about linear systems of ODEs, we need to talk about matrices, so let us review these briefly. A *matrix* is an $m \times n$ array of numbers (m rows and n columns). For example, we denote a 3×5 matrix as follows

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \end{bmatrix}.$$

The numbers a_{ij} are called *elements* or *entries*.

By a *vector* we usually mean a *column vector*, that is an $m \times 1$ matrix. If we mean a *row vector*, we will explicitly say so (a row vector is a $1 \times n$ matrix). We usually denote matrices by upper case letters and vectors by lower case letters with an arrow such as $\vec{x}$ or $\vec{b}$. By $\vec{0}$ we mean the vector of all zeros.

We define some operations on matrices. We want 1×1 matrices to really act like numbers, so our operations have to be compatible with this viewpoint.

First, we can multiply a matrix by a *scalar* (a number). We simply multiply each entry in the matrix by the scalar. For example,

$$2 \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{bmatrix}.$$

Matrix addition is also easy. We add matrices element by element. For example,

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 1 & 1 & -1 \\ 0 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 2 \\ 4 & 7 & 10 \end{bmatrix}.$$

If the sizes do not match, then addition is not defined.

If we denote by 0 the matrix with all zero entries, by c, d scalars, and by A, B, C matrices, we have the following familiar rules:

$$\begin{aligned} A + 0 &= A = 0 + A, \\ A + B &= B + A, \\ (A + B) + C &= A + (B + C), \\ c(A + B) &= cA + cB, \\ (c + d)A &= cA + dA. \end{aligned}$$

Another useful operation for matrices is the so-called *transpose*. This operation just swaps rows and columns of a matrix. The transpose of A is denoted by A^T. Example:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

4.2.2 Matrix multiplication

Let us now define matrix multiplication. First we define the so-called *dot product* (or *inner product*) of two vectors. Usually this will be a row vector multiplied with a column vector of the same size. For the dot product we multiply each pair of entries from the first and the second vector and we sum these products. The result is a single number. For example,

$$\begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = a_1 b_1 + a_2 b_2 + a_3 b_3.$$

And similarly for larger (or smaller) vectors.

Armed with the dot product we define the *product of matrices*. First let us denote by $\text{row}_i(A)$ the i^{th} row of A and by $\text{column}_j(A)$ the j^{th} column of A. For an $m \times n$ matrix A and an $n \times p$ matrix B we can define the product AB. We let AB be an $m \times p$ matrix whose ij^{th} entry is the dot product

$$\text{row}_i(A) \cdot \text{column}_j(B).$$

Do note how the sizes match up: $m \times n$ multiplied by $n \times p$ is $m \times p$. Example:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} =$$

$$= \begin{bmatrix} 1 \cdot 1 + 2 \cdot 1 + 3 \cdot 1 & 1 \cdot 0 + 2 \cdot 1 + 3 \cdot 0 & 1 \cdot (-1) + 2 \cdot 1 + 3 \cdot 0 \\ 4 \cdot 1 + 5 \cdot 1 + 6 \cdot 1 & 4 \cdot 0 + 5 \cdot 1 + 6 \cdot 0 & 4 \cdot (-1) + 5 \cdot 1 + 6 \cdot 0 \end{bmatrix} = \begin{bmatrix} 6 & 2 & 1 \\ 15 & 5 & 1 \end{bmatrix}.$$

For multiplication we want an analogue of a 1. This analogue is the so-called *identity matrix*. The identity matrix is a square matrix with 1s on the diagonal and zeros everywhere else. It is usually denoted by I. For each size we have a different identity matrix and so

sometimes we may denote the size as a subscript. For example, the I_3 would be the 3×3 identity matrix

$$I = I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

We have the following rules for matrix multiplication. Suppose that A, B, C are matrices of the correct sizes so that the following make sense. Let α denote a scalar (number).

$$\begin{aligned} A(BC) &= (AB)C, \\ A(B+C) &= AB + AC, \\ (B+C)A &= BA + CA, \\ \alpha(AB) &= (\alpha A)B = A(\alpha B), \\ IA &= A = AI. \end{aligned}$$

A few warnings are in order.

(i) $AB \neq BA$ in general (it may be true by fluke sometimes). That is, matrices do not commute. For example, take $A = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ and $B = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 2 \end{smallmatrix}\right]$.

(ii) $AB = AC$ does not necessarily imply $B = C$, even if A is not 0.

(iii) $AB = 0$ does not necessarily mean that $A = 0$ or $B = 0$. Try, for example, $A = B = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$.

For the last two items to hold we would need to "divide" by a matrix. This is where the *matrix inverse* comes in. Suppose that A and B are $n \times n$ matrices such that

$$AB = I = BA.$$

Then we call B the inverse of A and we denote B by A^{-1}. If the inverse of A exists, then we call A *invertible*. If A is not invertible, we sometimes say A is *singular*.

If A is invertible, then $AB = AC$ does imply that $B = C$ (in particular the inverse of A is unique). We just multiply both sides by A^{-1} (on the left) to get $A^{-1}AB = A^{-1}AC$ or $IB = IC$ or $B = C$. We can also see from the definition that $(A^{-1})^{-1} = A$.

4.2.3 The determinant

For square matrices we define a useful quantity called the *determinant*. We define the determinant of a 1×1 matrix as the value of its only entry. For a 2×2 matrix we define

$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) \stackrel{\text{def}}{=} ad - bc.$$

Before trying to define the determinant for larger matrices, let us note the meaning of the determinant. Consider an $n \times n$ matrix as a mapping of the n-dimensional euclidean space $\mathbb{R}^n$ to itself, where $\vec{x}$ gets sent to $A\vec{x}$. In particular, a 2×2 matrix A is a mapping of the plane

to itself. The determinant of A is the factor by which the area of objects changes. If we take the unit square (square of side 1) in the plane, then A takes the square to a parallelogram of area $|\det(A)|$. The sign of $\det(A)$ denotes changing of orientation (negative if the axes get flipped). For example, let

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

Then $\det(A) = 1 + 1 = 2$. Let us see where the (unit) square with vertices $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$ gets sent. Clearly $(0,0)$ gets sent to $(0,0)$.

$$\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

The image of the square is another square with vertices $(0,0)$, $(1,-1)$, $(1,1)$, and $(2,0)$. The image square has a side of length $\sqrt{2}$ and is therefore of area 2.

If you think back to high school geometry, you may have seen a formula for computing the area of a parallelogram with vertices $(0,0)$, (a,c), (b,d) and $(a+b, c+d)$. And it is precisely

$$\left| \det \left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) \right|.$$

The vertical lines above mean absolute value. The matrix $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ carries the unit square to the given parallelogram.

Let us look at the determinant for larger matrices. We define A_{ij} as the matrix A with the i^{th} row and the j^{th} column deleted. To compute the determinant of a matrix, pick one row, say the i^{th} row and compute:

$$\boxed{\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij}).}$$

For the first row we get

$$\det(A) = a_{11}\det(A_{11}) - a_{12}\det(A_{12}) + a_{13}\det(A_{13}) - \cdots \begin{cases} +a_{1n}\det(A_{1n}) & \text{if } n \text{ is odd,} \\ -a_{1n}\det(A_{1n}) & \text{if } n \text{ even.} \end{cases}$$

We alternately add and subtract the determinants of the submatrices A_{ij} multiplied by a_{ij} for a fixed i and all j. For a 3×3 matrix, picking the first row, we get $\det(A) = a_{11}\det(A_{11}) - a_{12}\det(A_{12}) + a_{13}\det(A_{13})$. For example,

$$\begin{aligned} \det\left(\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \right) &= 1 \cdot \det\left(\begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix} \right) - 2 \cdot \det\left(\begin{bmatrix} 4 & 6 \\ 7 & 9 \end{bmatrix} \right) + 3 \cdot \det\left(\begin{bmatrix} 4 & 5 \\ 7 & 8 \end{bmatrix} \right) \\ &= 1(5 \cdot 9 - 6 \cdot 8) - 2(4 \cdot 9 - 6 \cdot 7) + 3(4 \cdot 8 - 5 \cdot 7) = 0. \end{aligned}$$

The numbers $(-1)^{i+j}\det(A_{ij})$ are called *cofactors* of the matrix and this way of computing the determinant is called the *cofactor expansion*. No matter which row you pick, you always

get the same number. It is also possible to compute the determinant by expanding along columns (picking a column instead of a row above). It is true that $\det(A) = \det(A^T)$.

A common notation for the determinant is a pair of vertical lines:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = \det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right).$$

I personally find this notation confusing as vertical lines usually mean a positive quantity, while determinants can be negative. Also think about how to write the absolute value of a determinant. I will not use this notation in this book.

Think of the determinants telling you the scaling of a mapping. If B doubles the sizes of geometric objects and A triples them, then AB (which applies B to an object and then A) should make size go up by a factor of 6. This is true in general:

$$\det(AB) = \det(A)\det(B).$$

This property is one of the most useful, and it is employed often to actually compute determinants. A particularly interesting consequence is to note what it means for existence of inverses. Take A and B to be inverses of each other, that is $AB = I$. Then

$$\det(A)\det(B) = \det(AB) = \det(I) = 1.$$

Neither $\det(A)$ nor $\det(B)$ can be zero. Let us state this as a theorem as it will be very important in the context of this course.

Theorem 4.2.1

An $n \times n$ matrix A is invertible if and only if $\det(A) \neq 0$.

In fact, $\det(A^{-1})\det(A) = 1$ says that $\det(A^{-1}) = \frac{1}{\det(A)}$. So we even know what the determinant of A^{-1} is before we know how to compute A^{-1}.

There is a simple formula for the inverse of a 2×2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Notice the determinant of the matrix $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ in the denominator of the fraction. The formula only works if the determinant is nonzero, otherwise we are dividing by zero.

4.2.4 Solving linear systems

One application of matrices we will need is to solve systems of linear equations. This is best shown by example. Suppose that we have the following system of linear equations

$$\begin{aligned} 2x_1 + 2x_2 + 2x_3 &= 2, \\ x_1 + x_2 + 3x_3 &= 5, \\ x_1 + 4x_2 + x_3 &= 10. \end{aligned}$$

Without changing the solution, we could swap equations in this system, we could multiply any of the equations by a nonzero number, and we could add a multiple of one equation to another equation. It turns out these operations always suffice to find a solution.

It is easier to write the system as a matrix equation. The system above can be written as

$$\begin{bmatrix} 2 & 2 & 2 \\ 1 & 1 & 3 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \\ 10 \end{bmatrix}.$$

To solve the system we put the coefficient matrix (the matrix on the left-hand side of the equation) together with the vector on the right and side and get the so-called *augmented matrix*

$$\left[\begin{array}{ccc|c} 2 & 2 & 2 & 2 \\ 1 & 1 & 3 & 5 \\ 1 & 4 & 1 & 10 \end{array}\right].$$

We apply the following three elementary operations.

(i) Swap two rows.

(ii) Multiply a row by a nonzero number.

(iii) Add a multiple of one row to another row.

We keep doing these operations until we get into a state where it is easy to read off the answer, or until we get into a contradiction indicating no solution, for example if we come up with an equation such as $0 = 1$.

Let us work through the example. First multiply the first row by $1/2$ to obtain

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 5 \\ 1 & 4 & 1 & 10 \end{array}\right].$$

Now subtract the first row from the second and third row.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 4 \\ 0 & 3 & 0 & 9 \end{array}\right]$$

Multiply the last row by $1/3$ and the second row by $1/2$.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 3 \end{array}\right]$$

Swap rows 2 and 3.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 \end{array}\right]$$

Subtract the last row from the first, then subtract the second row from the first.

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & -4 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 \end{array}\right]$$

If we think about what equations this augmented matrix represents, we see that $x_1 = -4$, $x_2 = 3$, and $x_3 = 2$. We try this solution in the original system and, voilà, it works!

***Exercise* 4.2.1:** *Check that the solution above really solves the given equations.*

We write this equation in matrix notation as

$$A\vec{x} = \vec{b},$$

where A is the matrix $\left[\begin{smallmatrix} 2 & 2 & 2 \\ 1 & 1 & 3 \\ 1 & 4 & 1 \end{smallmatrix}\right]$ and $\vec{b}$ is the vector $\left[\begin{smallmatrix} 2 \\ 5 \\ 10 \end{smallmatrix}\right]$. The solution can also be computed via the inverse,

$$\vec{x} = A^{-1}A\vec{x} = A^{-1}\vec{b}.$$

It is possible that the solution is not unique, or that no solution exists. It is easy to tell if a solution does not exist. If during the row reduction you come up with a row where all the entries except the last one are zero (the last entry in a row corresponds to the right-hand side of the equation), then the system is *inconsistent* and has no solution. For example, for a system of 3 equations and 3 unknowns, if you find a row such as $[\,0 \quad 0 \quad 0 \mid 1\,]$ in the augmented matrix, you know the system is inconsistent. That row corresponds to $0 = 1$.

You generally try to use row operations until the following conditions are satisfied. The first (from the left) nonzero entry in each row is called the *leading entry*.

(i) The leading entry in any row is strictly to the right of the leading entry of the row above.

(ii) Any zero rows are below all the nonzero rows.

(iii) All leading entries are 1.

(iv) All the entries above and below a leading entry are zero.

Such a matrix is said to be in *reduced row echelon form*. The variables corresponding to columns with no leading entries are said to be *free variables*. Free variables mean that we can pick those variables to be anything we want and then solve for the rest of the unknowns.

Example 4.2.1: The following augmented matrix is in reduced row echelon form.

$$\left[\begin{array}{ccc|c} 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Suppose the variables are x_1, x_2, and x_3. Then x_2 is the free variable, $x_1 = 3 - 2x_2$, and $x_3 = 1$.

On the other hand if during the row reduction process you come up with the matrix

$$\left[\begin{array}{ccc|c} 1 & 2 & 13 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 3 \end{array}\right],$$

there is no need to go further. The last row corresponds to the equation $0x_1 + 0x_2 + 0x_3 = 3$, which is preposterous. Hence, no solution exists.

4.2.5 Computing the inverse

If the matrix A is square and there exists a unique solution $\vec{x}$ to $A\vec{x} = \vec{b}$ for any $\vec{b}$ (there are no free variables), then A is invertible. Multiplying both sides by A^{-1}, you can see that $\vec{x} = A^{-1}\vec{b}$. So it is useful to compute the inverse if you want to solve the equation for many different right-hand sides $\vec{b}$.

We have a formula for the 2×2 inverse, but it is also not hard to compute inverses of larger matrices. While we will not have too much occasion to compute inverses for larger matrices than 2×2 by hand, let us touch on how to do it. Finding the inverse of A is actually just solving a bunch of linear equations. If we can solve $A\vec{x}_k = \vec{e}_k$ where $\vec{e}_k$ is the vector with all zeros except a 1 at the k^{th} position, then the inverse is the matrix with the columns $\vec{x}_k$ for $k = 1, 2, \ldots, n$ (exercise: why?). Therefore, to find the inverse we write a larger $n \times 2n$ augmented matrix $[\, A \mid I \,]$, where I is the identity matrix. We then perform row reduction. The reduced row echelon form of $[\, A \mid I \,]$ will be of the form $[\, I \mid A^{-1} \,]$ if and only if A is invertible. We then just read off the inverse A^{-1}.

4.2.6 Eigenvalues and eigenvectors of a matrix

Let A be a constant square matrix. Suppose there is a scalar λ and a nonzero vector $\vec{v}$ such that

$$A\vec{v} = \lambda\vec{v}.$$

We call λ an *eigenvalue* of A and we call $\vec{v}$ a corresponding *eigenvector*.

Example 4.2.2: The matrix $\left[\begin{smallmatrix} 2 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ has an eigenvalue $\lambda = 2$ with a corresponding eigenvector $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ as

$$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Let us see how to compute eigenvalues for any matrix. Rewrite the equation for an eigenvalue as

$$(A - \lambda I)\vec{v} = \vec{0}.$$

This equation has a nonzero solution $\vec{v}$ only if $A - \lambda I$ is not invertible. Were it invertible, we could write $(A - \lambda I)^{-1}(A - \lambda I)\vec{v} = (A - \lambda I)^{-1}\vec{0}$, which implies $\vec{v} = \vec{0}$. Therefore, A has the eigenvalue λ if and only if λ solves the equation

$$\det(A - \lambda I) = 0.$$

Consequently, we will be able to find an eigenvalue of A without finding a corresponding eigenvector. An eigenvector will have to be found later, once λ is known.

Example 4.2.3: Find all eigenvalues of $\left[\begin{smallmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{smallmatrix}\right]$.

Solution: We write

$$\det\left(\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right) = \det\left(\begin{bmatrix} 2-\lambda & 1 & 1 \\ 1 & 2-\lambda & 0 \\ 0 & 0 & 2-\lambda \end{bmatrix}\right) =$$
$$= (2-\lambda)\big((2-\lambda)^2 - 1\big) = -(\lambda-1)(\lambda-2)(\lambda-3).$$

So the eigenvalues are $\lambda = 1$, $\lambda = 2$, and $\lambda = 3$. ⌟

For an $n \times n$ matrix, the polynomial we get by computing $\det(A - \lambda I)$ is of degree n, and hence in general, we have n eigenvalues. Some may be repeated, some may be complex.

To find an eigenvector corresponding to an eigenvalue λ, we write

$$(A - \lambda I)\vec{v} = \vec{0},$$

and solve for a nontrivial (nonzero) vector $\vec{v}$. If λ is an eigenvalue, there will be at least one free variable, and so for each distinct eigenvalue λ, we can always find an eigenvector.

Example 4.2.4: Find an eigenvector of $\left[\begin{smallmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{smallmatrix}\right]$ corresponding to the eigenvalue $\lambda = 3$.

Solution: We write

$$(A - \lambda I)\vec{v} = \left(\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} - 3\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right)\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \vec{0}.$$

It is easy to solve this system of linear equations. We write down the augmented matrix

$$\left[\begin{array}{ccc|c} -1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{array}\right],$$

and perform row operations (exercise: which ones?) until we get:

$$\left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

The entries of $\vec{v}$ have to satisfy the equations $v_1 - v_2 = 0$, $v_3 = 0$, and v_2 is a free variable. We can pick v_2 to be arbitrary (but nonzero), let $v_1 = v_2$, and of course $v_3 = 0$. For example, if we pick $v_2 = 1$, then $\vec{v} = \left[\begin{smallmatrix} 1 \\ 1 \\ 0 \end{smallmatrix}\right]$. Let us verify that $\vec{v}$ really is an eigenvector corresponding to $\lambda = 3$:

$$\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 0 \end{bmatrix} = 3\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

Yay! It worked. ⌟

Exercise* 4.2.2** (easy):*** *Are eigenvectors unique? Can you find a different eigenvector for $\lambda = 3$ in the example above? How are the two eigenvectors related?*

***Exercise* 4.2.3:** *When the matrix is 2×2 you do not need to do row operations when computing an eigenvector, you can read it off from $A - \lambda I$ (if you have computed the eigenvalues correctly). Can you see why? Explain. Try it for the matrix $\left[\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right]$.*

4.2.7 Exercises

***Exercise* 4.2.4:** *Let A and B be the matrices below.*

$$A = \begin{bmatrix} 1 & 4 & -1 \\ 2 & 0 & 3 \\ 1 & -2 & 3 \end{bmatrix} \qquad B = \begin{bmatrix} 0 & 2 & 3 \\ 1 & -4 & -2 \\ 2 & -5 & 1 \end{bmatrix}$$

Compute $A + 3B$, AB, and BA.

***Exercise* 4.2.5:** *Solve $\left[\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right] \vec{x} = \left[\begin{smallmatrix} 5 \\ 6 \end{smallmatrix}\right]$ by using matrix inverse.*

***Exercise* 4.2.6:** *Compute determinant of $\left[\begin{smallmatrix} 9 & -2 & -6 \\ -8 & 3 & 6 \\ 10 & -2 & -6 \end{smallmatrix}\right]$.*

Exercise* 4.2.7: *Compute determinant of $\left[\begin{smallmatrix} 1 & 1 & 1 \\ 2 & 3 & -5 \\ 1 & -1 & 0 \end{smallmatrix}\right]$*

***Exercise* 4.2.8:** *Compute determinant of $\left[\begin{smallmatrix} 1 & 2 & 3 & 1 \\ 4 & 0 & 5 & 0 \\ 6 & 0 & 7 & 0 \\ 8 & 0 & 10 & 1 \end{smallmatrix}\right]$. Hint: Expand along the proper row or column to make the calculations simpler.*

***Exercise* 4.2.9:** *Compute inverse of $\left[\begin{smallmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{smallmatrix}\right]$.*

***Exercise* 4.2.10:** *For which h is $\left[\begin{smallmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & h \end{smallmatrix}\right]$ not invertible? Is there only one such h? Are there several? Infinitely many?*

Exercise* 4.2.11: *Find t such that $\left[\begin{smallmatrix} 1 & t \\ -1 & 2 \end{smallmatrix}\right]$ is not invertible.*

***Exercise* 4.2.12:** *For which h is $\left[\begin{smallmatrix} h & 1 & 1 \\ 0 & h & 0 \\ 1 & 1 & h \end{smallmatrix}\right]$ not invertible? Find all such h.*

Exercise* 4.2.13: *Solve the system of equations*

$$\begin{aligned} 4x_1 - 2x_2 + 4x_3 &= -8 \\ x_1 - 3x_3 &= 12 \\ -4x_1 + 4x_2 + 4x_3 &= -8 \end{aligned}$$

or determine that no solution exists.

Exercise* 4.2.14: *Solve the system of equations*

$$\begin{aligned} -x_1 - 4x_2 + 2x_3 &= 11 \\ 3x_1 - 3x_2 + x_3 &= 13 \\ -5x_1 - 5x_2 + 3x_3 &= 9 \end{aligned}$$

or determine that no solution exists.

Exercise* 4.2.15: *Solve the system of equations*

$$\begin{aligned} x_1 + 3x_2 - 3x_3 &= 1 \\ -3x_1 - 4x_2 + 4x_3 &= -3 \\ 4x_1 + 7x_2 - 7x_3 &= 7 \end{aligned}$$

or determine that no solution exists.

Exercise* 4.2.16: *Solve the system of equations*

$$\begin{aligned} x_1 + 3x_2 - x_3 &= 5 \\ 2x_1 + x_2 &= -3 \\ -3x_1 - 4x_2 + 2x_3 &= -6 \end{aligned}$$

or determine that no solution exists.

***Exercise* 4.2.17:** *Solve* $\begin{bmatrix} 9 & -2 & -6 \\ -8 & 3 & 6 \\ 10 & -2 & -6 \end{bmatrix} \vec{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$.

***Exercise* 4.2.18:** *Solve* $\begin{bmatrix} 5 & 3 & 7 \\ 8 & 4 & 4 \\ 6 & 3 & 3 \end{bmatrix} \vec{x} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$.

Exercise* 4.2.19: *Solve* $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \vec{x} = \begin{bmatrix} 10 \\ 20 \end{bmatrix}$.

***Exercise* 4.2.20:** *Solve* $\begin{bmatrix} 3 & 2 & 3 & 0 \\ 3 & 3 & 3 & 3 \\ 0 & 2 & 4 & 2 \\ 2 & 3 & 4 & 3 \end{bmatrix} \vec{x} = \begin{bmatrix} 2 \\ 0 \\ 4 \\ 1 \end{bmatrix}$.

***Exercise* 4.2.21:** *Find 3 nonzero* 2×2 *matrices* A, B, *and* C *such that* $AB = AC$ *but* $B \neq C$.

Exercise* 4.2.22: *Suppose* a, b, c *are nonzero numbers. Let* $M = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$, $N = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}$.

a) *Compute* M^{-1}. b) *Compute* N^{-1}.

Exercise* 4.2.23** (easy):*** *Let* A *be a* 3×3 *matrix with an eigenvalue of 3 and a corresponding eigenvector* $\vec{v} = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}$. *Find* $A\vec{v}$.

Exercise* 4.2.24: *Find the eigenvalues and eigenvectors for the matrix*

$$\begin{bmatrix} 0 & -2 \\ 1 & 3 \end{bmatrix}.$$

Exercise* 4.2.25: *Find the eigenvalues and eigenvectors for the matrix*

$$\begin{bmatrix} -8 & -5 \\ 8 & 4 \end{bmatrix}.$$

Exercise* 4.2.26: *Find the eigenvalues and eigenvectors for the matrix*

$$\begin{bmatrix} 7 & -3 & 7 \\ 9 & -5 & 7 \\ 0 & 0 & -3 \end{bmatrix}.$$

4.3 Linear systems of ODEs

Attribution: [JL], §3.3.

Learning Objectives

After this section, you will be able to:

- Use proper terminiology when discussing linear systems of differential equations and their solutions,
- Determine whether a set of functions is linearly independent, and
- Understand how the theory of non-homogeneous linear systems relates to the theory of non-linear equations.

In order to get into the details of how to talk about and deal with linear systems of differential equations, we first need to talk about matrix- or vector-valued functions. Such a function is just a matrix or vector whose entries depend on some variable. If t is the independent variable, we write a *vector-valued function* $\vec{x}(t)$ as

$$\vec{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix}.$$

Similarly a *matrix-valued function* $A(t)$ is

$$A(t) = \begin{bmatrix} a_{11}(t) & a_{12}(t) & \cdots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \cdots & a_{2n}(t) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(t) & a_{n2}(t) & \cdots & a_{nn}(t) \end{bmatrix}.$$

As long as the addition of vectors is defined, we can add vector-valued functions, and as long as the addition and multiplication of matrices are defined (like if they are the right size) we can multiply matrix-valued functions. In addition, the derivative $A'(t)$ or $\frac{dA}{dt}$ is just the matrix-valued function whose ij^{th} entry is $a'_{ij}(t)$. We used this idea previously when talking about how to write first order systems from higher order equations in § 4.1.

Rules of differentiation of matrix-valued functions are similar to rules for normal functions. Let $A(t)$ and $B(t)$ be matrix-valued functions. Let c a scalar and let C be a constant matrix. Then

$$\begin{aligned} \big(A(t) + B(t)\big)' &= A'(t) + B'(t), \\ \big(A(t)B(t)\big)' &= A'(t)B(t) + A(t)B'(t), \\ \big(cA(t)\big)' &= cA'(t), \\ \big(CA(t)\big)' &= CA'(t), \\ \big(A(t)\,C\big)' &= A'(t)\,C. \end{aligned}$$

Note the order of the multiplication in the last two expressions because matrix multiplication is not commutative.

A *first order linear system of ODEs* is a system that can be written as the vector equation

$$\vec{x}'(t) = P(t)\vec{x}(t) + \vec{f}(t),$$

where $P(t)$ is a matrix-valued function, and $\vec{x}(t)$ and $\vec{f}(t)$ are vector-valued functions. We will often suppress the dependence on t and only write $\vec{x}' = P\vec{x} + \vec{f}$. A solution of the system is a vector-valued function $\vec{x}$ satisfying the vector equation.

For example, the equations

$$\begin{aligned} x_1' &= 2tx_1 + e^t x_2 + t^2, \\ x_2' &= \frac{x_1}{t} - x_2 + e^t, \end{aligned}$$

can be written as

$$\vec{x}' = \begin{bmatrix} 2t & e^t \\ 1/t & -1 \end{bmatrix} \vec{x} + \begin{bmatrix} t^2 \\ e^t \end{bmatrix}.$$

We will mostly concentrate on equations that are not just linear, but are in fact *constant coefficient* equations. That is, the matrix P will be constant; it will not depend on t.

When $\vec{f} = \vec{0}$ (the zero vector), then we say the system is *homogeneous*. For homogeneous linear systems we have the principle of superposition, just like for single homogeneous equations.

Theorem 4.3.1 (Superposition)

Let $\vec{x}' = P\vec{x}$ be a linear homogeneous system of ODEs. Suppose that $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ are n solutions of the equation and $c_1, c_2, \ldots, c_n$ are any constants, then

$$\vec{x} = c_1\vec{x}_1 + c_2\vec{x}_2 + \cdots + c_n\vec{x}_n, \tag{4.2}$$

is also a solution. Furthermore, if this is a system of n equations (P is $n \times n$), and $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ are linearly independent, then every solution $\vec{x}$ can be written as (4.2).

Linear independence for vector-valued functions is the same idea as for normal functions. The vector-valued functions $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ are linearly independent if the only way to satisfy the equation

$$c_1\vec{x}_1 + c_2\vec{x}_2 + \cdots + c_n\vec{x}_n = \vec{0}$$

is by choosing the parameters $c_1 = c_2 = \cdots = c_n = 0$, where the equation must hold for all t.

Example 4.3.1: Determine if the sets $S_1 = \left\{ \vec{x}_1 = \begin{bmatrix} t^2 \\ t \end{bmatrix}, \ \vec{x}_2 = \begin{bmatrix} 0 \\ 1+t \end{bmatrix}, \ \vec{x}_3 = \begin{bmatrix} -t^2 \\ 1 \end{bmatrix} \right\}$ and $S_2 = \left\{ \vec{x}_1 = \begin{bmatrix} t^2 \\ t \end{bmatrix}, \ \vec{x}_2 = \begin{bmatrix} 0 \\ t \end{bmatrix}, \ \vec{x}_3 = \begin{bmatrix} -t^2 \\ 1 \end{bmatrix} \right\}$ are linearly independent.

Solution: The vector functions in S_1 are linearly dependent because $\vec{x}_1 + \vec{x}_3 = \vec{x}_2$, and this holds for all t. So $c_1 = 1$, $c_2 = -1$, and $c_3 = 1$ above will work.

On the other hand, the vector functions in S_2 are linearly independent, even though this is only a slight change from S_1. First write $c_1\vec{x}_1 + c_2\vec{x}_2 + c_3\vec{x}_3 = \vec{0}$ and note that it has to hold for all t. We get that

$$c_1\vec{x}_1 + c_2\vec{x}_2 + c_3\vec{x}_3 = \begin{bmatrix} c_1t^2 - c_3t^2 \\ c_1t + c_2t + c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

In other words $c_1t^2 - c_3t^2 = 0$ and $c_1t + c_2t + c_3 = 0$. If we set $t = 0$, then the second equation becomes $c_3 = 0$. But then the first equation becomes $c_1t^2 = 0$ for all t and so $c_1 = 0$. Thus the second equation is just $c_2t = 0$, which means $c_2 = 0$. So $c_1 = c_2 = c_3 = 0$ is the only solution and $\vec{x}_1$, $\vec{x}_2$, and $\vec{x}_3$ are linearly independent. ⌟

The linear combination $c_1\vec{x}_1 + c_2\vec{x}_2 + \cdots + c_n\vec{x}_n$ could always be also as

$$X(t)\,\vec{c},$$

where $X(t)$ is the matrix with columns $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$, and $\vec{c}$ is the column vector with entries $c_1, c_2, \ldots, c_n$. This is similar to the way that we could write linear combinations of vectors by putting them into a matrix, including how we talked about rank in § 3.4. Assuming that $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ are linearly independent and solutions to a given system of differential equations, the matrix-valued function $X(t)$ is called a *fundamental matrix*, or a *fundamental matrix solution*.

To solve nonhomogeneous first order linear systems, we use the same technique as we applied to solve single linear nonhomogeneous equations.

Theorem 4.3.2

Let $\vec{x}' = P\vec{x} + \vec{f}$ be a linear system of ODEs. Suppose $\vec{x}_p$ is one particular solution. Then every solution can be written as

$$\vec{x} = \vec{x}_c + \vec{x}_p,$$

where $\vec{x}_c$ is a solution to the associated homogeneous equation ($\vec{x}' = P\vec{x}$).

The procedure for systems is the same as for single equations. We find a particular solution to the nonhomogeneous equation, then we find the general solution to the associated homogeneous equation, and finally we add the two together.

Alright, suppose you have found the general solution of $\vec{x}' = P\vec{x} + \vec{f}$. Next suppose you are given an initial condition of the form

$$\vec{x}(t_0) = \vec{b}$$

for some fixed t_0 and a constant vector $\vec{b}$. Let $X(t)$ be a fundamental matrix solution of the associated homogeneous equation (i.e. columns of $X(t)$ are solutions). The general solution can be written as

$$\vec{x}(t) = X(t)\,\vec{c} + \vec{x}_p(t).$$

We are seeking a vector $\vec{c}$ such that

$$\vec{b} = \vec{x}(t_0) = X(t_0)\,\vec{c} + \vec{x}_p(t_0).$$

In other words, we are solving for $\vec{c}$ in the nonhomogeneous system of linear equations

$$X(t_0)\,\vec{c} = \vec{b} - \vec{x}_p(t_0).$$

Example 4.3.2: In § 4.1 we solved the system

$$\begin{aligned} x_1' &= x_1, \\ x_2' &= x_1 - x_2, \end{aligned}$$

with initial conditions $x_1(0) = 1$, $x_2(0) = 2$. Let us consider this problem in the language of this section.

The system is homogeneous, so $\vec{f}(t) = \vec{0}$. We write the system and the initial conditions as

$$\vec{x}' = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \vec{x}, \qquad \vec{x}(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

We found the general solution is $x_1 = c_1 e^t$ and $x_2 = \frac{c_1}{2} e^t + c_2 e^{-t}$. Letting $c_1 = 1$ and $c_2 = 0$, we obtain the solution $\begin{bmatrix} e^t \\ (1/2)e^t \end{bmatrix}$. Letting $c_1 = 0$ and $c_2 = 1$, we obtain $\begin{bmatrix} 0 \\ e^{-t} \end{bmatrix}$. These two solutions are linearly independent, as can be seen by setting $t = 0$, and noting that the resulting constant vectors are linearly independent. In matrix notation, a fundamental matrix solution is, therefore,

$$X(t) = \begin{bmatrix} e^t & 0 \\ \frac{1}{2}e^t & e^{-t} \end{bmatrix}.$$

To solve the initial value problem we solve for $\vec{c}$ in the equation

$$X(0)\,\vec{c} = \vec{b},$$

or in other words,

$$\begin{bmatrix} 1 & 0 \\ \frac{1}{2} & 1 \end{bmatrix} \vec{c} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

A single elementary row operation shows $\vec{c} = \begin{bmatrix} 1 \\ 3/2 \end{bmatrix}$. Our solution is

$$\vec{x}(t) = X(t)\,\vec{c} = \begin{bmatrix} e^t & 0 \\ \frac{1}{2}e^t & e^{-t} \end{bmatrix} \begin{bmatrix} 1 \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} e^t \\ \frac{1}{2}e^t + \frac{3}{2}e^{-t} \end{bmatrix}.$$

This new solution agrees with our previous solution from § 4.1.

4.3.1 Exercises

***Exercise* 4.3.1:** *Write the system* $x_1' = 2x_1 - 3tx_2 + \sin t$, $x_2' = e^t x_1 + 3x_2 + \cos t$ *in the form* $\vec{x}' = P(t)\vec{x} + \vec{f}(t)$.

Exercise* 4.3.2: *Write $x' = 3x - y + e^t$, $y' = tx$ in matrix notation.*

***Exercise* 4.3.3:** *Consider the third order differential equation*

$$y''' + (y' + 1)^2 = e^y + \sin(t + 1).$$

Convert this to a first order system and simplify as much as possible. Can you write this in the form $\vec{x}' = A\vec{x} + \vec{f}$? Why or why not?

***Exercise* 4.3.4:**

a) *Verify that the system $\vec{x}' = \left[\begin{smallmatrix} 1 & 3 \\ 3 & 1 \end{smallmatrix}\right] \vec{x}$ has the two solutions $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right] e^{4t}$ and $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right] e^{-2t}$.*

b) *Write down the general solution.*

c) *Write down the general solution in the form $x_1 = ?$, $x_2 = ?$ (i.e. write down a formula for each element of the solution).*

***Exercise* 4.3.5:** *Verify that $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right] e^t$ and $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right] e^t$ are linearly independent. Hint: Just plug in $t = 0$.*

Exercise* 4.3.6: *Are $\left[\begin{smallmatrix} e^{2t} \\ e^t \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} e^t \\ e^{2t} \end{smallmatrix}\right]$ linearly independent? Justify.*

***Exercise* 4.3.7:** *Verify that $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} e^t$ and $\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} e^t$ and $\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} e^{2t}$ are linearly independent. Hint: You must be a bit more tricky than in the previous exercises.*

Exercise* 4.3.8: *Are $\left[\begin{smallmatrix} \cosh(t) \\ 1 \end{smallmatrix}\right]$, $\left[\begin{smallmatrix} e^t \\ 1 \end{smallmatrix}\right]$, and $\left[\begin{smallmatrix} e^{-t} \\ 1 \end{smallmatrix}\right]$ linearly independent? Justify.*

***Exercise* 4.3.9:** *Verify that $\left[\begin{smallmatrix} t \\ t^2 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} t^3 \\ t^4 \end{smallmatrix}\right]$ are linearly independent.*

***Exercise* 4.3.10:** *Take the system $x_1' + x_2' = x_1$, $x_1' - x_2' = x_2$.*

a) *Write it in the form $A\vec{x}' = B\vec{x}$ for matrices A and B.*

b) *Compute A^{-1} and use that to write the system in the form $\vec{x}' = P\vec{x}$.*

Exercise* 4.3.11:

a) *Write $x_1' = 2tx_2$, $x_2' = 2tx_2$ in matrix notation.*

b) *Solve and write the solution in matrix notation.*

4.4 Eigenvalue method

Attribution: [JL], §3.4, 3.7.

Learning Objectives

After this section, you will be able to:

- Use the eigenvalue method to find straight-line solutions to constant-coefficient first order systems of ODE,
- Find general solutions to systems with real and distinct eigenvalues, and
- Solve initial value problems from all of these cases once the general solution has been found.

In this section we will learn how to solve linear homogeneous constant coefficient systems of ODEs by the eigenvalue method. Suppose we have such a system

$$\vec{x}' = P\vec{x},$$

where P is a constant square matrix. We wish to adapt the method for the single constant coefficient equation by trying the function $e^{\lambda t}$. However, $\vec{x}$ is a vector. So we try $\vec{x} = \vec{v}e^{\lambda t}$, where $\vec{v}$ is an arbitrary constant vector. We plug this $\vec{x}$ into the equation to get

$$\underbrace{\lambda \vec{v} e^{\lambda t}}_{\vec{x}'} = \underbrace{P\vec{v}e^{\lambda t}}_{P\vec{x}}.$$

We divide by $e^{\lambda t}$ and notice that we are looking for a scalar λ and a vector $\vec{v}$ that satisfy the equation

$$\lambda \vec{v} = P\vec{v}.$$

This means that we are looking for an eigenvalue λ with corresponding eigenvector $\vec{v}$ for the matrix P. When we can find these, we will get solutions to the original system of differential equations of the form

$$\vec{x}(t) = \vec{v}e^{\lambda t}.$$

We get the easiest route to solutions when the matrix P has all real eigenvalues and the eigenvalues are all distinct, and can extend to deal with the complications that arise from complex and repeated eigenvalues.

Another way to view these types of solutions are as "straight-line solutions." A system of differential equations of the form

$$\vec{x}' = P\vec{x},$$

is an autonomous system of differential equations, because there is no explicit dependence on t on the right-hand side. When we solved autonomous equations in § 1.7, we started by looking for equilibrium solutions and built up from there. In this particular case, we are looking for vectors $\vec{x}$ so that $P\vec{x} = 0$. As long as P is invertible, the only vector that satisfies this is $\vec{x} = 0$. So, that's not super interesting, and doesn't really tell us too much about the solution to the problem.

The next more involved type of solution we could look for is a straight-line solution. The idea is that this solution will either move directly (in a straight-line) towards or away from the origin. In the first order autonomous equation case, all of our solutions did this; they either moved towards or away from these equilibrium solutions. This may not be the case for systems, but we can try to find them. If a solution is going to move directly towards or away from the origin, then the direction of change for the solution must be parallel to the position vector. In Figure 4.4, the vectors that point in the same or opposite direction of $\vec{x}$ will give rise to a straight-line solution, but vectors that do not point in this direction will give solutions that do not follow a straight-line through the origin.

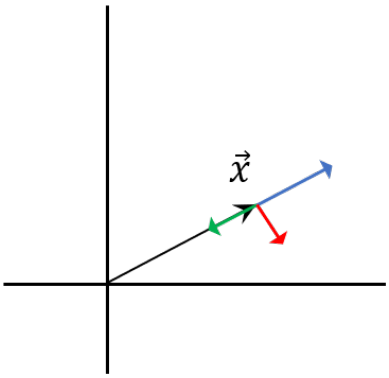


Figure 4.4: Position vector and possible direction vectors for straight line solutions.

This criterion means that we need to have

$$\vec{x}' = \lambda\vec{x}$$

for some constant λ. If this is the case, then we have

$$P\vec{x} = \lambda\vec{x}$$

and this is the equation for eigenvalues and eigenvectors of P. We are back to the same type of solution that we found previously.

4.4.1 The eigenvalue method with distinct real eigenvalues

OK. We have the system of equations

$$\vec{x}' = P\vec{x}.$$

We find the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of the matrix P, and corresponding eigenvectors $\vec{v}_1$, $\vec{v}_2, \ldots, \vec{v}_n$. Now we notice that the functions $\vec{v}_1 e^{\lambda_1 t}$, $\vec{v}_2 e^{\lambda_2 t}$, $\ldots$, $\vec{v}_n e^{\lambda_n t}$ are solutions of the homogeneous system of equations and hence $\vec{x} = c_1\vec{v}_1 e^{\lambda_1 t} + c_2\vec{v}_2 e^{\lambda_2 t} + \cdots + c_n\vec{v}_n e^{\lambda_n t}$ is a solution by superposition.

Theorem 4.4.1

Take $\vec{x}' = P\vec{x}$. If P is an $n \times n$ constant matrix that has n distinct real eigenvalues λ_1, $\lambda_2, \ldots, \lambda_n$, then there exist n linearly independent corresponding eigenvectors $\vec{v}_1$, $\vec{v}_2$, $\ldots, \vec{v}_n$, and the general solution to $\vec{x}' = P\vec{x}$ can be written as

$$\vec{x} = c_1\vec{v}_1 e^{\lambda_1 t} + c_2\vec{v}_2 e^{\lambda_2 t} + \cdots + c_n\vec{v}_n e^{\lambda_n t}.$$

The corresponding fundamental matrix solution is

$$X(t) = \begin{bmatrix} \vec{v}_1 e^{\lambda_1 t} & \vec{v}_2 e^{\lambda_2 t} & \cdots & \vec{v}_n e^{\lambda_n t} \end{bmatrix}.$$

That is, $X(t)$ is the matrix whose j^{th} column is $\vec{v}_j e^{\lambda_j t}$.

Example 4.4.1: Consider the system

$$\vec{x}' = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \vec{x}.$$

Find the general solution.

Solution: Earlier, we found the eigenvalues are $1, 2, 3$. We found the eigenvector $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ for the eigenvalue 3. Similarly we find the eigenvector $\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$ for the eigenvalue 1, and $\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$ for the eigenvalue 2 (exercise: check). Hence our general solution is

$$\vec{x} = c_1 \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} e^t + c_2 \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} e^{2t} + c_3 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} e^{3t} = \begin{bmatrix} c_1 e^t + c_3 e^{3t} \\ -c_1 e^t + c_2 e^{2t} + c_3 e^{3t} \\ -c_2 e^{2t} \end{bmatrix}.$$

In terms of a fundamental matrix solution,

$$\vec{x} = X(t)\,\vec{c} = \begin{bmatrix} e^t & 0 & e^{3t} \\ -e^t & e^{2t} & e^{3t} \\ 0 & -e^{2t} & 0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}.$$

***Exercise* 4.4.1:** *Check that this $\vec{x}$ really solves the system.*

Overall, the process for finding the solution for real and distinct eigenvalues is to first find the eigenvalues and eigenvectors of the matrix P. Once we have these, we get n linearly independent solutions of the form $\vec{x}_i(t) = \vec{v}_i e^{\lambda_i t}$, so that the general solution is of the form

$$\vec{x}(t) = c_1 \vec{v}_1 e^{\lambda_1 t} + c_2 \vec{v}_2 e^{\lambda_2 t} + \cdots + c_n \vec{v}_n e^{\lambda_n t}.$$

Then, if we need to solve for an initial condition, we figure out the coefficients c_1, c_2, ..., c_n to satisfy this condition.

Note: If we write a single homogeneous linear constant coefficient n^{th} order equation as a first order system (as we did in § 4.1), then the eigenvalue equation

$$\det(P - \lambda I) = 0$$

is essentially the same as the characteristic equation we got in § 2.1 and § 2.7. See the exercises for details about this.

Example 4.4.2: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} 0 & 4 \\ -3 & -7 \end{bmatrix} \vec{x} \qquad \vec{x}(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Solution: Since we are in the case of a constant-coefficient linear system, we start by looking for the eigenvalues and eigenvectors of the coefficient matrix P. To do this, we compute

$$\det(P - \lambda I) = (0 - \lambda)(-7 - \lambda) - (4)(-3) = \lambda^2 + 7\lambda + 12.$$

This polynomial factors as $(\lambda + 3)(\lambda + 4)$, and so the two eigenvalues are $\lambda_1 = -3$ and $\lambda_2 = -4$.

Next, we need to find the corresponding eigenvectors. For $\lambda = -3$, we get the matrix equation

$$(P + 3I)\vec{v} = \begin{bmatrix} 3 & 4 \\ -3 & -4 \end{bmatrix} \vec{v} = \vec{0}.$$

The two equations that you get here are redundant, which is $3v_1 + 4v_2 = 0$. One way to satisfy this is $v_1 = 4$, $v_2 = -3$, so that the eigenvector is $\left[\begin{smallmatrix} 4 \\ -3 \end{smallmatrix}\right]$.

For $\lambda = -4$, the matrix becomes

$$(P + 4I)\vec{v} = \begin{bmatrix} 4 & 4 \\ -3 & -3 \end{bmatrix} \vec{v} = 0$$

so the eigenvector here is $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$. Therefore, the general solution to this differential equation, by superposition, is

$$\vec{x}(t) = c_1 \begin{bmatrix} 4 \\ -3 \end{bmatrix} e^{-3t} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-4t}.$$

Finally, we have to solve the initial value problem using the initial conditions. If we plug in $t = 0$, we get the equation

$$\vec{x}(0) = c_1 \begin{bmatrix} 4 \\ -3 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

This results in needing to solve the system of equations

$$4c_1 + c_2 = 1 \qquad -3c_1 - c_2 = 1.$$

These can be solved in any way, including row reduction. We will start by adding the two equations together, which gives $c_1 = 2$, and then the first equation implies that $c_2 = -7$. Therefore, the solution to the initial value problem is

$$\vec{x}(t) = 2 \begin{bmatrix} 4 \\ -3 \end{bmatrix} e^{-3t} - 7 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-4t} = \begin{bmatrix} 8e^{-3t} - 7e^{-4t} \\ -6e^{-3t} + 7e^{-4t} \end{bmatrix}.$$

4.4.2 Phase Portraits

Now that we have these solutions, we want to get an idea for what they look like in the plane. We spent a lot of time in first order equations looking at direction fields, as well as phase lines for autonomous equations. We want to develope the same type of intuition for two-component systems in the plane, because much intuition can be obtained by studying this

simple case. Suppose we use coordinates (x, y) for the plane as usual, and suppose $P = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ is a 2×2 matrix. Consider the system

$$\begin{bmatrix} x \\ y \end{bmatrix}' = P \begin{bmatrix} x \\ y \end{bmatrix} \qquad \text{or} \qquad \begin{bmatrix} x \\ y \end{bmatrix}' = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{4.3}$$

The system is autonomous (compare this section to § 1.7) and so we can draw a vector field (see the end of § 4.1). We will be able to visually tell what the vector field looks like and how the solutions behave, once we find the eigenvalues and eigenvectors of the matrix P. The goal is to be able to sketch what the different trajectories of the solutions look like for a variety of initial conditions, as well as classify the general type of picture that results depending on the matrix P.

Case 1. Suppose that the eigenvalues of P are real and positive. We find two corresponding eigenvectors and plot them in the plane. For example, take the matrix $\left[\begin{smallmatrix} 1 & 1 \\ 0 & 2 \end{smallmatrix}\right]$. The eigenvalues are 1 and 2 and corresponding eigenvectors are $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$. See Figure 4.5.

Suppose the point (x, y) is on the line determined by an eigenvector $\vec{v}$ for an eigenvalue λ. That is, $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] = \alpha\vec{v}$ for some scalar α. Then

$$\begin{bmatrix} x \\ y \end{bmatrix}' = P \begin{bmatrix} x \\ y \end{bmatrix} = P(\alpha\vec{v}) = \alpha(P\vec{v}) = \alpha\lambda\vec{v}.$$

The derivative is a multiple of $\vec{v}$ and hence points along the line determined by $\vec{v}$. As $\lambda > 0$, the derivative points in the direction of $\vec{v}$ when α is positive and in the opposite direction when α is negative. Let us draw the lines determined by the eigenvectors, and let us draw arrows on the lines to indicate the directions. See Figure 4.6 on the following page.

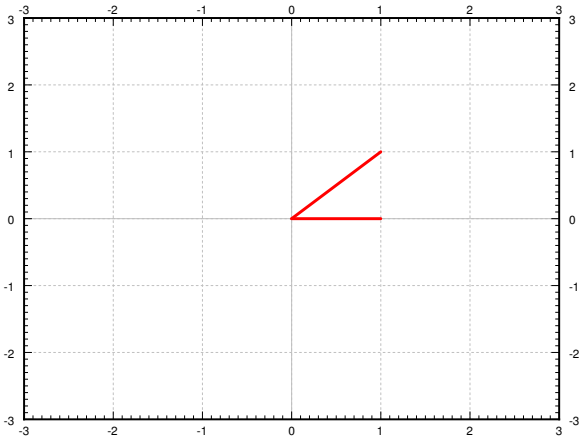


Figure 4.5: Eigenvectors of P.

We fill in the rest of the arrows for the vector field and we also draw a few solutions. See Figure 4.7 on the next page. The picture looks like a source with arrows coming out from the origin. Hence we call this type of picture a *source* or sometimes an *unstable node*. Notice the two eigenvectors are drawn on the entire vector field figure with arrows, and the straight-line solutions follow them.

Case 2. Suppose both eigenvalues are negative. For example, take the negation of the matrix in case 1, $\left[\begin{smallmatrix} -1 & -1 \\ 0 & -2 \end{smallmatrix}\right]$. The eigenvalues are -1 and -2 and corresponding eigenvectors are the same, $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$. The calculation and the picture are almost the same. The only difference is that the eigenvalues are negative and hence all arrows are reversed. We get the picture in Figure 4.8 on the following page. We call this kind of picture a *sink* or a *asymptotically stable node*.

Case 3. Suppose one eigenvalue is positive and one is negative. For example the matrix $\left[\begin{smallmatrix} 1 & 1 \\ 0 & -2 \end{smallmatrix}\right]$. The eigenvalues are 1 and -2 and corresponding eigenvectors are $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ -3 \end{smallmatrix}\right]$. We

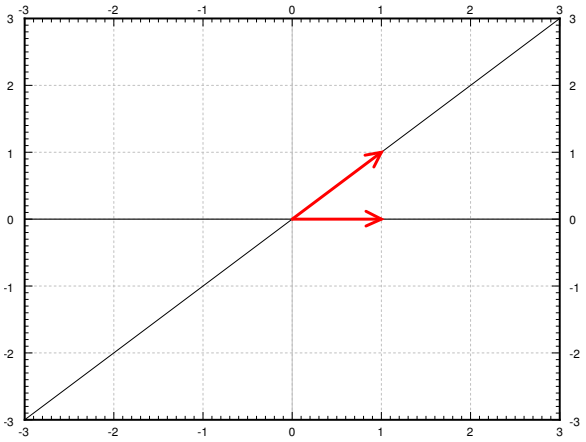


Figure 4.6: Eigenvectors of P with directions.

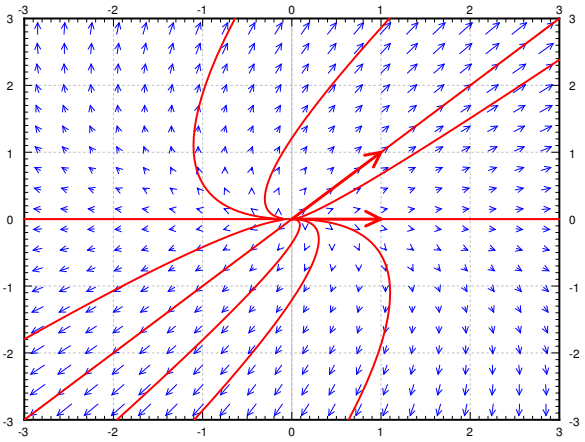


Figure 4.7: Example source vector field with eigenvectors and solutions.

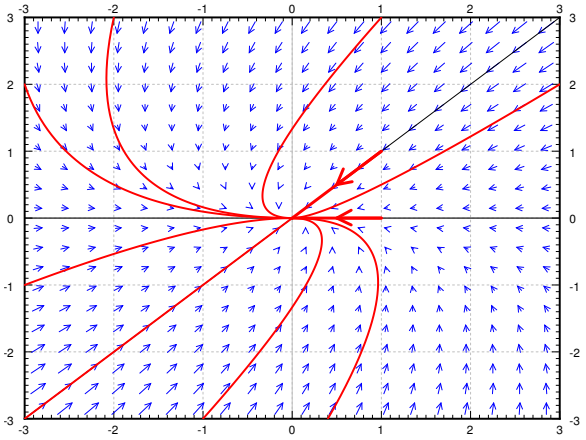


Figure 4.8: Example sink vector field with eigenvectors and solutions.

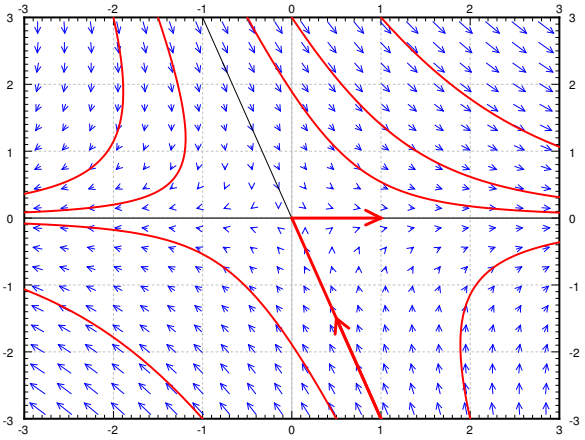


Figure 4.9: Example saddle vector field with eigenvectors and solutions.

reverse the arrows on one line (corresponding to the negative eigenvalue) and we obtain the picture in Figure 4.9. We call this picture a *saddle point*.

4.4.3 Exercises

***Exercise* 4.4.2:**

a) *Find the general solution of $x_1' = 2x_1$, $x_2' = 3x_2$ using the eigenvalue method (first write the system in the form $\vec{x}' = A\vec{x}$).*

b) *Solve the system by solving each equation separately and verify you get the same general solution.*

***Exercise* 4.4.3:** *Find the general solution of $x_1' = 3x_1 + x_2$, $x_2' = 2x_1 + 4x_2$ using the eigenvalue method.*

Exercise* 4.4.4: *Solve $x_1' = x_2$, $x_2' = x_1$ using the eigenvalue method.*

***Exercise* 4.4.5:** *Consider the second order equation given by*

$$y'' + 2y' - 8y = 0.$$

a) *Find the general solution of this problem using the methods of Chapter 2.*

b) *Convert this equation into a first order linear system using the transformation $\vec{x} = \begin{bmatrix} y \\ y' \end{bmatrix}$.*

c) *Find the eigenvalues and eigenvalues of the coefficient matrix and use that to find the general solution to the system.*

d) *Extract the first component of the general solution and compare that to the solution from part (a). How do they relate?*

e) *Look back through the work. How do the equation used to find the roots in (a) and the eigenvalues in (c) relate to each other?*

***Exercise* 4.4.6:** *Consider the second order equation given by*

$$y'' + 4y' + 5y = 0.$$

a) *Find the general solution of this problem using the methods of Chapter 2.*

b) *Convert this equation into a first order linear system using the transformation $\vec{x} = \begin{bmatrix} y \\ y' \end{bmatrix}$.*

c) *Find the eigenvalues and eigenvalues of the coefficient matrix and use that to find the general solution to the system.*

d) *Extract the first component of the general solution and compare that to the solution from part (a). How do they relate?*

e) *Look back through the work. How do the equation used to find the roots in (a) and the eigenvalues in (c) relate to each other?*

***Exercise* 4.4.7:** *Consider the second order equation given by*

$$y'' + by' - cy = 0.$$

for b and c two real numbers.

a) *Find the general solution of this problem using the methods of Chapter 2.*

b) *Convert this equation into a first order linear system using the transformation $\vec{x} = \begin{bmatrix} y \\ y' \end{bmatrix}$.*

c) *Find the eigenvalues and eigenvalues of the coefficient matrix and use that to find the general solution to the system.*

d) *Extract the first component of the general solution and compare that to the solution from part (a). How do they relate?*

e) *Look back through the work. How do the equation used to find the roots in (a) and the eigenvalues in (c) relate to each other?*

***Exercise* 4.4.8:** *Amino acid dating can be used by forensic scientists to determine the time of death in situations where other techniques might not work. These amino acids are sneaky, and they exist in a left-handed form (L) and a right-handed form (D), which are called enantiomers. While you're alive, your body keeps all your amino acids in the L form. Once you die, your body no longer regulates your amino acids, and every so often they flip a coin and decide whether to switch into the opposite form. This way, when someone finds your body in a dumpster, they can pull out your teeth and measure the racemization ratio, which is the ratio of D-enantiomers to L-enantiomers.*

Denote by $D(t)$ and $L(t)$, respectively, the proportions of D- and L-enantiomers found in your teeth, where t is measured in years after death. Since this is Math class, the proportions are governed by a system of differential equations, such as

$$\begin{bmatrix} L' \\ D' \end{bmatrix} = \begin{bmatrix} -.02 & .02 \\ .02 & -.02 \end{bmatrix} \begin{bmatrix} L \\ D \end{bmatrix}. \tag{4.4}$$

a) *Find the general solution to (4.4).*

b) *Solve (4.4) with initial conditions $D(0) = 0$ and $L(0) = 1$, and express the solution in component form. Describe what happens to the quantities $D(t)$ and $L(t)$ in the long run.*

c) *Given the above initial conditions, if the racemization ratio in your teeth is currently 1:3, how long ago did you die?*

***Exercise* 4.4.9:**

a) *Compute eigenvalues and eigenvectors of $A = \begin{bmatrix} 9 & -2 & -6 \\ -8 & 3 & 6 \\ 10 & -2 & -6 \end{bmatrix}$.*

b) *Find the general solution of $\vec{x}' = A\vec{x}$.*

Exercise* 4.4.10:

a) *Compute eigenvalues and eigenvectors of $A = \begin{bmatrix} 1 & 0 & 3 \\ -1 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix}$.*

b) *Solve the system $\vec{x}\,' = A\vec{x}$.*

***Exercise* 4.4.11:** *Let a, b, c, d, e, f be numbers. Find the eigenvalues of $\begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix}$.*

Exercise* 4.4.12: *Find the general solution of the system*

$$\vec{x}' = \begin{bmatrix} -7 & 1 \\ -12 & 0 \end{bmatrix} \vec{x}.$$

Exercise* 4.4.13: *Find the general solution of the system*

$$\vec{x}' = \begin{bmatrix} -13 & -12 \\ 9 & 8 \end{bmatrix} \vec{x}.$$

Exercise* 4.4.14: *Find the general solution of the system*

$$\vec{x}' = \begin{bmatrix} -2 & -6 & 0 \\ 4 & 8 & 0 \\ -4 & -7 & 3 \end{bmatrix} \vec{x}.$$

Exercise* 4.4.15: *Find the general solution of the system*

$$\vec{x}' = \begin{bmatrix} -6 & 2 & 4 \\ -2 & -1 & 4 \\ -2 & 1 & 0 \end{bmatrix} \vec{x}.$$

***Exercise* 4.4.16:** *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} -3 & 0 \\ 3 & -4 \end{bmatrix} \vec{x} \qquad \vec{x}(0) = \begin{bmatrix} -1 \\ 2 \end{bmatrix}.$$

***Exercise* 4.4.17:** *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} 1 & -3 \\ 2 & 6 \end{bmatrix} \vec{x} \qquad \vec{x}(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

***Exercise* 4.4.18:** *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} 7 & 4 & 0 \\ -8 & -5 & 0 \\ 17 & 7 & -2 \end{bmatrix} \vec{x} \qquad \vec{x}(0) = \begin{bmatrix} -3 \\ 2 \\ 2 \end{bmatrix}.$$

4.5 Eigenvalue method with complex eigenvalues

Attribution: [JL], §3.4, 3.7.

Learning Objectives

After this section, you will be able to:

- Use Euler's formula to find a real-valued general solution to a first order system with complex eigenvalues and
- Solve initial value problems from all of these cases once the general solution has been found.

As we have seen previously, a matrix may very well have complex eigenvalues even if all the entries are real. However, this may seem concerning going forward into solutions to differential equations that require these complex numbers in them. We will see in this section that we can still write solutions this way, but we no longer have straight-line solutions. Take, for example,

$$\vec{x}' = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \vec{x}.$$

Let us compute the eigenvalues of the matrix $P = \left[\begin{smallmatrix} 1 & 1 \\ -1 & 1 \end{smallmatrix}\right]$.

$$\det(P - \lambda I) = \det\left(\begin{bmatrix} 1-\lambda & 1 \\ -1 & 1-\lambda \end{bmatrix}\right) = (1-\lambda)^2 + 1 = \lambda^2 - 2\lambda + 2 = 0.$$

Thus $\lambda = 1 \pm i$. Corresponding eigenvectors are also complex. Start with $\lambda = 1 - i$.

$$\big(P - (1-i)I\big)\vec{v} = \vec{0},$$
$$\begin{bmatrix} i & 1 \\ -1 & i \end{bmatrix} \vec{v} = \vec{0}.$$

The equations $iv_1 + v_2 = 0$ and $-v_1 + iv_2 = 0$ are multiples of each other. This may be trickier to spot than the real version, but that is because they are *complex* multiples of each other. If we multiply the first equation by i, we get exactly the second one. So we only need to consider one of them. After picking $v_2 = 1$, for example, we have an eigenvector $\vec{v} = \left[\begin{smallmatrix} i \\ 1 \end{smallmatrix}\right]$. In similar fashion we find that $\left[\begin{smallmatrix} -i \\ 1 \end{smallmatrix}\right]$ is an eigenvector corresponding to the eigenvalue $1 + i$.

We could write the solution as

$$\vec{x} = c_1 \begin{bmatrix} i \\ 1 \end{bmatrix} e^{(1-i)t} + c_2 \begin{bmatrix} -i \\ 1 \end{bmatrix} e^{(1+i)t} = \begin{bmatrix} c_1 i e^{(1-i)t} - c_2 i e^{(1+i)t} \\ c_1 e^{(1-i)t} + c_2 e^{(1+i)t} \end{bmatrix}.$$

We would then need to look for complex values c_1 and c_2 to solve any initial conditions. It is perhaps not completely clear that we get a real solution. After solving for c_1 and c_2, we could use Euler's formula and do the whole song and dance we did before, but we will not. We will apply the formula in a smarter way first to find independent real solutions.

In this case, we only needed one of the two eigenvectors to get the general solution, which happens because the complex eigenvalues and eigenvectors always come in conjugate pairs.

First a small detour. The real part of a complex number z can be computed as $\frac{z+\bar{z}}{2}$, where the bar above z means $\overline{a+ib} = a-ib$. This operation is called the *complex conjugate*. If a is a real number, then $\bar{a} = a$. Similarly we bar whole vectors or matrices by taking the complex conjugate of every entry. Suppose a matrix P is real. Then $\overline{P} = P$, and so $\overline{P\vec{x}} = \overline{P}\,\overline{\vec{x}} = P\overline{\vec{x}}$. Also the complex conjugate of 0 is still 0, therefore,

$$\vec{0} = \overline{\vec{0}} = \overline{(P-\lambda I)\vec{v}} = (P-\bar{\lambda}I)\overline{\vec{v}}.$$

In other words, if $\lambda = a+ib$ is an eigenvalue, then so is $\bar{\lambda} = a-ib$. And if $\vec{v}$ is an eigenvector corresponding to the eigenvalue λ, then $\overline{\vec{v}}$ is an eigenvector corresponding to the eigenvalue $\bar{\lambda}$.

Suppose $a+ib$ is a complex eigenvalue of P, and $\vec{v}$ is a corresponding eigenvector. Then

$$\vec{x}_1 = \vec{v}e^{(a+ib)t}$$

is a solution (complex-valued) of $\vec{x}' = P\vec{x}$. Euler's formula shows that $\overline{e^{a+ib}} = e^{a-ib}$, and so

$$\vec{x}_2 = \overline{\vec{x}_1} = \overline{\vec{v}}e^{(a-ib)t}$$

is also a solution. As $\vec{x}_1$ and $\vec{x}_2$ are solutions, the function

$$\vec{x}_3 = \operatorname{Re}\vec{x}_1 = \operatorname{Re}\vec{v}e^{(a+ib)t} = \frac{\vec{x}_1 + \overline{\vec{x}_1}}{2} = \frac{\vec{x}_1+\vec{x}_2}{2} = \frac{1}{2}\vec{x}_1 + \frac{1}{2}\vec{x}_2$$

is also a solution. And $\vec{x}_3$ is real-valued! Similarly as $\operatorname{Im} z = \frac{z-\bar{z}}{2i}$ is the imaginary part, we find that

$$\vec{x}_4 = \operatorname{Im}\vec{x}_1 = \frac{\vec{x}_1 - \overline{\vec{x}_1}}{2i} = \frac{\vec{x}_1-\vec{x}_2}{2i}.$$

is also a real-valued solution. It turns out that $\vec{x}_3$ and $\vec{x}_4$ are linearly independent. We will use Euler's formula to separate out the real and imaginary part.

Returning to our problem,

$$\vec{x}_1 = \begin{bmatrix} i \\ 1 \end{bmatrix} e^{(1-i)t} = \begin{bmatrix} i \\ 1 \end{bmatrix} \left(e^t\cos t - ie^t \sin t\right) = \begin{bmatrix} ie^t\cos t + e^t\sin t \\ e^t\cos t - ie^t\sin t \end{bmatrix} = \begin{bmatrix} e^t\sin t \\ e^t\cos t \end{bmatrix} + i\begin{bmatrix} e^t\cos t \\ -e^t\sin t \end{bmatrix}.$$

Then

$$\operatorname{Re}\vec{x}_1 = \begin{bmatrix} e^t\sin t \\ e^t\cos t \end{bmatrix}, \qquad \text{and} \qquad \operatorname{Im}\vec{x}_1 = \begin{bmatrix} e^t\cos t \\ -e^t\sin t \end{bmatrix},$$

are the two real-valued linearly independent solutions we seek.

***Exercise* 4.5.1:** *Check that these really are solutions.*

This gives that we can write the general solution to this problem as

$$\vec{x} = c_1\begin{bmatrix} e^t\sin t \\ e^t\cos t \end{bmatrix} + c_2\begin{bmatrix} e^t\cos t \\ -e^t\sin t \end{bmatrix} = \begin{bmatrix} c_1e^t\sin t + c_2e^t\cos t \\ c_1e^t\cos t - c_2e^t\sin t \end{bmatrix}.$$

This solution is real-valued for real c_1 and c_2. We now solve for any initial conditions we may have. Notice that the i has been dropped from the part of the process where we split the

complex solution into real and imaginary parts. The point is that the real and imaginary parts of the solution are independently solutions to the equation, and so we can use them to form our basis of solutions with constants c_1 and c_2 in front of them. We want everything to be real, and this process allows us to do it.

Let us summarize as a theorem.

Theorem 4.5.1

Let P be a real-valued constant matrix. If P has a complex eigenvalue $a+ib$ and a corresponding eigenvector $\vec{v}$, then P also has a complex eigenvalue $a-ib$ with a corresponding eigenvector $\overline{\vec{v}}$. Furthermore, $\vec{x}' = P\vec{x}$ has two linearly independent real-valued solutions

$$\vec{x}_1 = \operatorname{Re} \vec{v}e^{(a+ib)t}, \qquad \text{and} \qquad \vec{x}_2 = \operatorname{Im} \vec{v}e^{(a+ib)t}.$$

The main point here is that the real and imaginary parts of these complex solutions are the real-valued independent solutions that we seek. Compare this to Theorem 2.2.2 in § 2.2, where we saw that the same idea worked for second order equation with complex roots.

For each pair of complex eigenvalues $a+ib$ and $a-ib$, we get two real-valued linearly independent solutions. We then go on to the next eigenvalue, which is either a real eigenvalue or another complex eigenvalue pair. If we have n distinct eigenvalues (real or complex), then we end up with n linearly independent solutions. If we had only two equations ($n=2$) as in the example above, then once we found two solutions we are finished, and our general solution is

$$\vec{x} = c_1\vec{x}_1 + c_2\vec{x}_2 = c_1\bigl(\operatorname{Re} \vec{v}e^{(a+ib)t}\bigr) + c_2\bigl(\operatorname{Im} \vec{v}e^{(a+ib)t}\bigr).$$

Example 4.5.1: Find the solution to the initial value problem

$$\vec{x}' = \begin{bmatrix} 1 & 4 \\ -2 & -3 \end{bmatrix} \vec{x} \qquad \vec{x}(0) = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

Solution: We start by looking for the eigenvalues and eigenvectors of the coefficient matrix. This results in the polynomial

$$\det(P-\lambda I) = (1-\lambda)(-3-\lambda) - (4)(-2) = \lambda^2 + 3\lambda - \lambda - 3 + 8 = \lambda^2 + 2\lambda + 5.$$

This polynomial does not factor, but the quadratic formula gives that the roots are

$$\lambda = \frac{-2 \pm \sqrt{4-(4)(1)(5)}}{2} = -1 \pm \frac{\sqrt{-16}}{2} = -1 \pm 2i.$$

Thus, we are in the complex roots case, and can work from there. We need to find the complex eigenvector for one of these eigenvalues and then split into real and imaginary parts to get the general solution.

For the eigenvalue $\lambda = -1+2i$, the matrix equation becomes

$$(P-\lambda I)\vec{v} = \begin{bmatrix} 1-(-1+2i) & 4 \\ -2 & -3-(-1+2i) \end{bmatrix} \vec{v} = \begin{bmatrix} 2-2i & 4 \\ -2 & -2-2i \end{bmatrix} \vec{v} = \vec{0}.$$

The two simultaneous equation that we need to solve for the vector v are

$$(2-2i)v_1 + 4v_2 = 0 \qquad -2v_1 + (-2-2i)v_2 = 0$$

and these equations don't appear to be redundant. However, this is because they are complex multiples of each other, not just real multiples. To see this, we can multiply the first equation by the complex conjugate of the first coefficient. The idea is that if we do so, this first coefficient will be real, and then we can compare it to the second equation. If we multiply the first equation by $2+2i$, since $(2+2i)(2-2i) = 8$, it becomes

$$8v_1 + 4(2+2i)v_2 = 0$$

and this is -4 times the second equation above. Therefore, they are redundant, and we can just pick one of them in order to find possible values of v_1 and v_2. If we divide this newest equation by 8, it becomes

$$v_1 + (1+i)v_2 = 0.$$

Based on this equation, we can pick $v_2 = -1$ and $v_1 = 1+i$. Therefore, the eigenvector for $\lambda = -1+2i$ is $\left[\begin{smallmatrix} 1+i \\ -1 \end{smallmatrix}\right]$. This means that a complex-valued solution to this differential equation is

$$\vec{x}(t) = \begin{bmatrix} 1+i \\ -1 \end{bmatrix} e^{(-1+2i)t}.$$

Now, we want to split this solution into real and imaginary parts in order to get a real-valued general solution. We apply Euler's formula to do so:

$$\begin{aligned}
\vec{x}(t) &= \begin{bmatrix} 1+i \\ -1 \end{bmatrix} e^{(-1+2i)t} \\
&= \begin{bmatrix} 1+i \\ -1 \end{bmatrix} e^{-t}(\cos(2t) + i\sin(2t)) \\
&= e^{-t} \begin{bmatrix} \cos(2t) + i\sin(2t) + i\cos(2t) - \sin(2t) \\ -\cos(2t) - i\sin(2t) \end{bmatrix} \\
&= \begin{bmatrix} e^{-t}\cos(2t) - e^{-t}\sin(2t) \\ -e^{-t}\cos(2t) \end{bmatrix} + i \begin{bmatrix} e^{-t}\sin(2t) + e^{-t}\cos(2t) \\ -e^{-t}\sin(2t) \end{bmatrix}.
\end{aligned}$$

Therefore, we can take the real and imaginary parts of this solution to get a general solution as

$$\vec{x}(t) = c_1 \begin{bmatrix} e^{-t}\cos(2t) - e^{-t}\sin(2t) \\ -e^{-t}\cos(2t) \end{bmatrix} + c_2 \begin{bmatrix} e^{-t}\sin(2t) + e^{-t}\cos(2t) \\ -e^{-t}\sin(2t) \end{bmatrix}.$$

***Exercise* 4.5.2:** *Work out the eigenvector and general solution from eigenvalue $\lambda = -1-2i$ and verify that it is an equivalent general solution to the one above.*

Finally, we need to solve the initial value problem. Plugging in $t = 0$ gives

$$\vec{x}(0) = c_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

The two equations that we get from here is $c_1 + c_2 = 1$ and $-c_1 = -2$, so that $c_1 = 2$ and $c_2 = -1$. Therefore, the solution to the initial value problem is

$$\vec{x}(t) = 2\begin{bmatrix} e^{-t}\cos(2t) - e^{-t}\sin(2t) \\ -e^{-t}\cos(2t) \end{bmatrix} - \begin{bmatrix} e^{-t}\sin(2t) + e^{-t}\cos(2t) \\ -e^{-t}\sin(2t) \end{bmatrix} = \begin{bmatrix} 2e^{-t}\cos(2t) - 3e^{-t}\sin(2t) \\ -2e^{-t}\cos(2t) + e^{-t}\sin(2t) \end{bmatrix}.$$

4.5.1 Phase Portraits

Similarly to the real eigenvalue situation, we have three different cases for the phase portrait when the eigenvalues of a 2x2 matrix are complex. As metioned before, our basis solutions that we are using to form the general solution are no longer just exponential terms. They involve sines and cosines, and so are not straight lines anymore. Therefore, these solutions will not have straight lines in them, but we can still uses these basis solutions to help determine and describe the overall behavior of the solutions to the system for a variety of initial conditions.

Case 1. Suppose the eigenvalues are purely imaginary. That is, suppose the eigenvalues are $\pm ib$. For example, let $P = \left[\begin{smallmatrix} 0 & 1 \\ -4 & 0 \end{smallmatrix}\right]$. The eigenvalues turn out to be $\pm 2i$ and eigenvectors are $\left[\begin{smallmatrix} 1 \\ 2i \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ -2i \end{smallmatrix}\right]$. Consider the eigenvalue $2i$ and its eigenvector $\left[\begin{smallmatrix} 1 \\ 2i \end{smallmatrix}\right]$. The real and imaginary parts of $\vec{v}e^{2it}$ are

$$\mathrm{Re}\left(\begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{2it}\right) = \begin{bmatrix} \cos(2t) \\ -2\sin(2t) \end{bmatrix}, \qquad \mathrm{Im}\left(\begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{2it}\right) = \begin{bmatrix} \sin(2t) \\ 2\cos(2t) \end{bmatrix}.$$

We can take any linear combination of them to get other solutions, which one we take depends on the initial conditions. Now note that the real part is a parametric equation for an ellipse. Same with the imaginary part and in fact any linear combination of the two. This is what happens in general when the eigenvalues are purely imaginary. So when the eigenvalues are purely imaginary, we get *ellipses* for the solutions. This type of picture is sometimes called a *center*. See Figure 4.10.

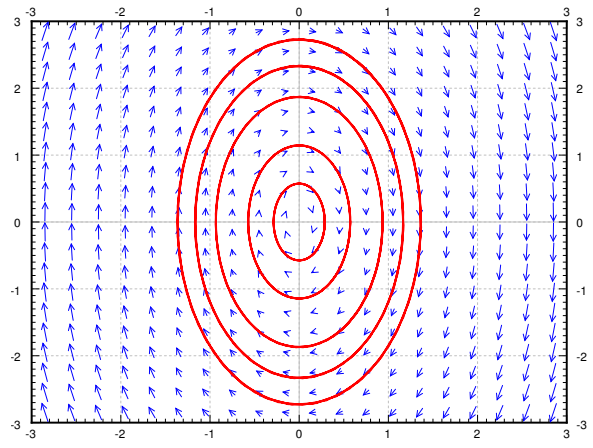


Figure 4.10: Example center vector field.

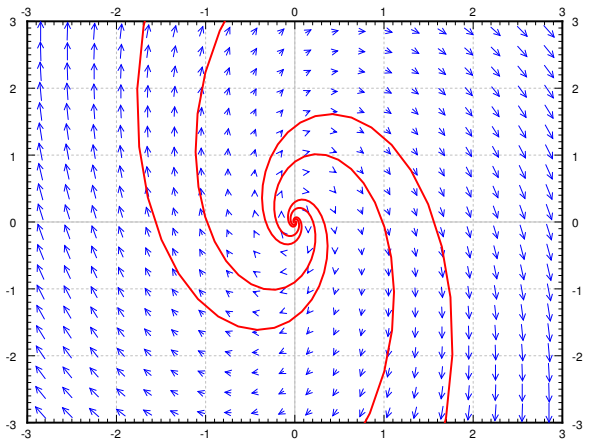


Figure 4.11: Example spiral source vector field.

Case 2. Now suppose the complex eigenvalues have a positive real part. That is, suppose the eigenvalues are $a \pm ib$ for some $a > 0$. For example, let $P = \left[\begin{smallmatrix} 1 & 1 \\ -4 & 1 \end{smallmatrix}\right]$. The eigenvalues turn out to be $1 \pm 2i$ and eigenvectors are $\left[\begin{smallmatrix} 1 \\ 2i \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ -2i \end{smallmatrix}\right]$. We take $1 + 2i$ and its eigenvector $\left[\begin{smallmatrix} 1 \\ 2i \end{smallmatrix}\right]$ and find the real and imaginary parts of $\vec{v}e^{(1+2i)t}$ are

$$\operatorname{Re}\left(\begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{(1+2i)t}\right) = e^t \begin{bmatrix} \cos(2t) \\ -2\sin(2t) \end{bmatrix}, \qquad \operatorname{Im}\left(\begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{(1+2i)t}\right) = e^t \begin{bmatrix} \sin(2t) \\ 2\cos(2t) \end{bmatrix}.$$

Note the e^t in front of the solutions. The solutions grow in magnitude while spinning around the origin. Hence we get a *spiral source*. See Figure 4.11 on the preceding page.

Case 3. Finally suppose the complex eigenvalues have a negative real part. That is, suppose the eigenvalues are $-a \pm ib$ for some $a > 0$. For example, let $P = \left[\begin{smallmatrix} -1 & -1 \\ 4 & -1 \end{smallmatrix}\right]$. The eigenvalues turn out to be $-1 \pm 2i$ and eigenvectors are $\left[\begin{smallmatrix} 1 \\ -2i \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ 2i \end{smallmatrix}\right]$. We take $-1 - 2i$ and its eigenvector $\left[\begin{smallmatrix} 1 \\ 2i \end{smallmatrix}\right]$ and find the real and imaginary parts of $\vec{v}e^{(-1-2i)t}$ are

$$\operatorname{Re}\left(\begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{(-1-2i)t}\right) = e^{-t} \begin{bmatrix} \cos(2t) \\ 2\sin(2t) \end{bmatrix}, \qquad \operatorname{Im}\left(\begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{(-1-2i)t}\right) = e^{-t} \begin{bmatrix} -\sin(2t) \\ 2\cos(2t) \end{bmatrix}.$$

Note the e^{-t} in front of the solutions. The solutions shrink in magnitude while spinning around the origin. Hence we get a *spiral sink*. See Figure 4.12.

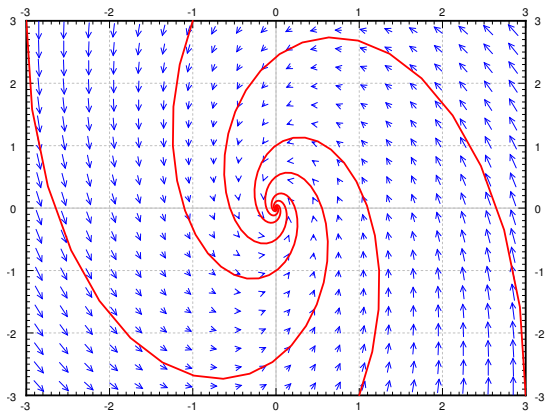


Figure 4.12: Example spiral sink vector field.

4.5.2 Exercises

***Exercise* 4.5.3:** *Find the general solution of $x_1' = x_1 - 2x_2$, $x_2' = 2x_1 + x_2$ using the eigenvalue method. Do not use complex exponentials in your solution.*

Exercise* 4.5.4: *Solve $x_1' = x_2$, $x_2' = -x_1$ using the eigenvalue method.*

***Exercise* 4.5.5:** *A 2×2 matrix A has complex eigenvector $\vec{v} = \begin{bmatrix} 1 \\ i \end{bmatrix}$ corresponding to eigenvalue $\lambda = -1 + 3i$.*

a) *Use Euler's Formula to find the (real-valued) general solution to the system* $\vec{x}' = A\vec{x}$.

b) *Sketch the phase portrait of this system.*

Exercise 4.5.6:*

a) *Compute eigenvalues and eigenvectors of* $A = \left[\begin{smallmatrix} 1 & 1 \\ -1 & 0 \end{smallmatrix}\right]$.

b) *Solve the system* $\vec{x}\,' = A\vec{x}$.

Exercise 4.5.7: *Consider the system*

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & -2 \\ 5 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

a) *Find the general solution.*

b) *Solve the IVP with initial conditions* $x(0) = 1, y(0) = 0$, *and determine the maximum* x*-coordinate on this trajectory.*

Exercise 4.5.8: *Find the general solution of the system*

$$\vec{x}' = \begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \vec{x}.$$

Exercise 4.5.9: *Find the general solution of the system*

$$\vec{x}' = \begin{bmatrix} 1 & 4 \\ -2 & -3 \end{bmatrix} \vec{x}.$$

Exercise 4.5.10: *Find the general solution of the system*

$$\vec{x}' = \begin{bmatrix} 2 & 0 & 3 \\ -6 & 2 & -9 \\ -3 & 0 & 2 \end{bmatrix} \vec{x}.$$

Exercise 4.5.11: *Find the general solution of the system*

$$\vec{x}' = \begin{bmatrix} -10 & -4 & 0 \\ 14 & 4 & 1 \\ 12 & 6 & -2 \end{bmatrix} \vec{x}.$$

Exercise 4.5.12: *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} 3 & -1 \\ 4 & 3 \end{bmatrix} \vec{x} \qquad \vec{x}(0) = \begin{bmatrix} 2 \\ -1 \end{bmatrix}.$$

Exercise 4.5.13: *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} -8 & -8 \\ 5 & 4 \end{bmatrix} \vec{x} \qquad \vec{x}(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Exercise 4.5.14: *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} -1 & 2 & -8 \\ 0 & 1 & -4 \\ 0 & 2 & -3 \end{bmatrix} \vec{x} \qquad \vec{x}(0) = \begin{bmatrix} 2 \\ 1 \\ -3 \end{bmatrix}.$$

4.6 Eigenvalue method with repeated eigenvalues

Attribution: [JL], §3.4, 3.7.

Learning Objectives

After this section, you will be able to:

- Find generalized eigenvectors to write a general solution to a first order system with repeated and defective eigenvalues, and
- Solve initial value problems from all of these cases once the general solution has been found.

There is one remaining case for the two-component first-order linear system: repeated eigenvalues. As we have seen previously, it may happen that a matrix A has some "repeated" eigenvalues. That is, the characteristic equation $\det(A - \lambda I) = 0$ may have repeated roots. This is actually unlikely to happen for a random matrix. If we take a small perturbation of A (we change the entries of A slightly), we get a matrix with distinct eigenvalues. As any system we want to solve in practice is an approximation to reality anyway, it is not absolutely indispensable to know how to solve these corner cases. On the other hand, these cases do come up in applications from time to time. Furthermore, if we have distinct but very close eigenvalues, the behavior is similar to that of repeated eigenvalues, and so understanding that case will give us insight into what is going on.

Geometric multiplicity

Take the diagonal matrix

$$A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}.$$

A has an eigenvalue 3 of multiplicity 2. We call the multiplicity of the eigenvalue in the characteristic equation the *algebraic multiplicity*. In this case, there also exist 2 linearly independent eigenvectors, $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$ corresponding to the eigenvalue 3. This means that the so-called *geometric multiplicity* of this eigenvalue is also 2. These terms have all been discussed previously in § 3.6.

In all the theorems where we required a matrix to have n distinct eigenvalues, we only really needed to have n linearly independent eigenvectors. For example, $\vec{x}' = A\vec{x}$ has the general solution

$$\vec{x} = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{3t} + c_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{3t}.$$

Let us restate the theorem about real eigenvalues. In the following theorem we will repeat eigenvalues according to (algebraic) multiplicity. So for the matrix A above, we would say that it has eigenvalues 3 and 3.

Theorem 4.6.1

Suppose the $n \times n$ matrix P has n real eigenvalues (not necessarily distinct), λ_1, λ_2, $\ldots$, λ_n, and there are n linearly independent corresponding eigenvectors $\vec{v}_1$, $\vec{v}_2$, $\ldots$, $\vec{v}_n$. Then the general solution to $\vec{x}' = P\vec{x}$ can be written as

$$\vec{x} = c_1\vec{v}_1 e^{\lambda_1 t} + c_2\vec{v}_2 e^{\lambda_2 t} + \cdots + c_n\vec{v}_n e^{\lambda_n t}.$$

The main difference in the statement here from the theorem in § 4.4 is that we are no longer assuming that we have n *distinct* eigenvalues. Instead, we need to assume that we end up with n linearly independent eigenvectors, which we get for free if the eigenvalues are all distinct, but we might also have that if we do not have all distinct eigenvalues.

The *geometric multiplicity* of an eigenvalue of algebraic multiplicity n is equal to the number of corresponding linearly independent eigenvectors. The geometric multiplicity is always less than or equal to the algebraic multiplicity. The theorem handles the case when these two multiplicities are equal for all eigenvalues. If for an eigenvalue the geometric multiplicity is equal to the algebraic multiplicity, then we say the eigenvalue is *complete*.

In other words, the hypothesis of the theorem could be stated as saying that if all the eigenvalues of P are complete, then there are n linearly independent eigenvectors and thus we have the given general solution.

If the geometric multiplicity of an eigenvalue is 2 or greater, then the set of linearly independent eigenvectors is not unique up to multiples as it was before. For example, for the diagonal matrix $A = \left[\begin{smallmatrix} 3 & 0 \\ 0 & 3 \end{smallmatrix}\right]$ we could also pick eigenvectors $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$, or in fact any pair of two linearly independent vectors. The number of linearly independent eigenvectors corresponding to λ is the number of free variables we obtain when solving $A\vec{v} = \lambda\vec{v}$. We pick specific values for those free variables to obtain eigenvectors. If you pick different values, you may get different eigenvectors.

Defective eigenvalues

If an $n \times n$ matrix has less than n linearly independent eigenvectors, it is said to be *deficient*. Then there is at least one eigenvalue with an algebraic multiplicity that is higher than its geometric multiplicity. We call this eigenvalue *defective* and the difference between the two multiplicities we call the *defect*.

Example 4.6.1: The matrix

$$\begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$$

has an eigenvalue 3 of algebraic multiplicity 2. Let us try to compute eigenvectors.

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \vec{0}.$$

We must have that $v_2 = 0$. Hence any eigenvector is of the form $\left[\begin{smallmatrix} v_1 \\ 0 \end{smallmatrix}\right]$. Any two such vectors are linearly dependent, and hence the geometric multiplicity of the eigenvalue is 1. Therefore, the defect is 1, and we can no longer apply the eigenvalue method directly to a system of ODEs with such a coefficient matrix.

Roughly, the key observation is that if λ is an eigenvalue of A of algebraic multiplicity m, then we can find certain m linearly independent vectors solving $(A-\lambda I)^k\vec{v}=\vec{0}$ for various powers k. We will call these *generalized eigenvectors*.

Let us continue with the example $A=\left[\begin{smallmatrix}3&1\\0&3\end{smallmatrix}\right]$ and the equation $\vec{x}'=A\vec{x}$. We found an eigenvalue $\lambda=3$ of (algebraic) multiplicity 2 and defect 1. We found one eigenvector $\vec{v}=\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$. We have one solution

$$\vec{x}_1=\vec{v}e^{3t}=\begin{bmatrix}1\\0\end{bmatrix}e^{3t}.$$

We are now stuck, we get no other solutions from standard eigenvectors. But we need two linearly independent solutions to find the general solution of the equation.

Let us try (in the spirit of repeated roots of the characteristic equation for a single equation) another solution of the form

$$\vec{x}_2=\vec{v}_1te^{3t},$$

since our modified guess for repeated roots from second order equations was te^{3t}. If we plug this guess into the equation, we get that

$$\vec{x}_2'=\vec{v}_1e^{3t}+3\vec{v}_1te^{3t}$$

and since the right-hand side of the equation is $A\vec{v}_1te^{3t}$, we need v_1 to satisfy

$$\vec{v}_1e^{3t}+3\vec{v}_1te^{3t}=A\vec{v}_1te^{3t}.$$

Since there is no e^{3t} term on the right-hand side of the equation, we are forced to pick $\vec{v}_1=\vec{0}$, and so we get the solution $\vec{x}_2=\vec{0}$, which is not good. This guess did not work.

The issue here is that we didn't have enough flexibility to actually get another solution to the differential equation, so we need something a little more complicated to make it work. To this end, we take a new guess of the form

$$\vec{x}_2=(\vec{v}_2+\vec{v}_1t)\,e^{3t}.$$

We differentiate to get

$$\vec{x}_2'=\vec{v}_1e^{3t}+3(\vec{v}_2+\vec{v}_1t)\,e^{3t}=(3\vec{v}_2+\vec{v}_1)\,e^{3t}+3\vec{v}_1te^{3t}.$$

As we are assuming that $\vec{x}_2$ is a solution, $\vec{x}_2'$ must equal $A\vec{x}_2$. So let's compute $A\vec{x}_2$:

$$A\vec{x}_2=A(\vec{v}_2+\vec{v}_1t)\,e^{3t}=A\vec{v}_2e^{3t}+A\vec{v}_1te^{3t}.$$

By looking at the coefficients of e^{3t} and te^{3t} we see $3\vec{v}_2+\vec{v}_1=A\vec{v}_2$ and $3\vec{v}_1=A\vec{v}_1$. This means that

$$(A-3I)\vec{v}_2=\vec{v}_1,\qquad\text{and}\qquad(A-3I)\vec{v}_1=\vec{0}.$$

Therefore, $\vec{x}_2$ is a solution if these two equations are satisfied. The second equation is satisfied if $\vec{v}_1$ is an eigenvector, and we found the eigenvector above, so let $\vec{v}_1=\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$. So, if we can find

a $\vec{v}_2$ that solves $(A-3I)\vec{v}_2 = \vec{v}_1$, then we are done. This is just a bunch of linear equations to solve and we are by now very good at that. Let us solve $(A-3I)\vec{v}_2 = \vec{v}_1$. Write

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

By inspection we see that letting $a = 0$ (a could be anything in fact) and $b = 1$ does the job. Hence we can take $\vec{v}_2 = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$. Our general solution to $\vec{x}' = A\vec{x}$ is

$$\vec{x} = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{3t} + c_2 \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} t \right) e^{3t} = \begin{bmatrix} c_1 e^{3t} + c_2 t e^{3t} \\ c_2 e^{3t} \end{bmatrix}.$$

Let us check that we really do have the solution. First $x_1' = c_1 3e^{3t} + c_2 e^{3t} + 3c_2 t e^{3t} = 3x_1 + x_2$. Good. Now $x_2' = 3c_2 e^{3t} = 3x_2$. Good.

In the example, if we plug $(A-3I)\vec{v}_2 = \vec{v}_1$ into $(A-3I)\vec{v}_1 = \vec{0}$ we find

$$(A-3I)(A-3I)\vec{v}_2 = \vec{0}, \qquad \text{or} \qquad (A-3I)^2\vec{v}_2 = \vec{0}.$$

Furthermore, if $(A-3I)\vec{w} \neq \vec{0}$, then $(A-3I)\vec{w}$ is an eigenvector, a multiple of $\vec{v}_1$. In this 2×2 case $(A-3I)^2$ is just the zero matrix (exercise). So any vector $\vec{w}$ solves $(A-3I)^2\vec{w} = \vec{0}$ and we just need a $\vec{w}$ such that $(A-3I)\vec{w} \neq \vec{0}$. Then we could use $\vec{w}$ for $\vec{v}_2$, and $(A-3I)\vec{w}$ for $\vec{v}_1$.

Note that the system $\vec{x}' = A\vec{x}$ has a simpler solution since A is a so-called *upper triangular matrix*, that is every entry below the diagonal is zero. In particular, the equation for x_2 does not depend on x_1. Mind you, not every defective matrix is triangular.

***Exercise* 4.6.1:** *Solve $\vec{x}' = \left[\begin{smallmatrix} 3 & 1 \\ 0 & 3 \end{smallmatrix}\right]\vec{x}$ by first solving for x_2 and then for x_1 independently. Check that you got the same solution as we did above.*

Let us describe the general algorithm. Suppose that λ is an eigenvalue of multiplicity 2, defect 1. First find an eigenvector $\vec{v}_1$ of λ. That is, $\vec{v}_1$ solves $(A-\lambda I)\vec{v}_1 = \vec{0}$. Then, find a vector $\vec{v}_2$ such that

$$(A-\lambda I)\vec{v}_2 = \vec{v}_1.$$

This gives us two linearly independent solutions

$$\begin{aligned} \vec{x}_1 &= \vec{v}_1 e^{\lambda t}, \\ \vec{x}_2 &= (\vec{v}_2 + \vec{v}_1 t)\, e^{\lambda t}, \end{aligned}$$

and so our general solution to the differential equation is

$$\vec{x}(t) = c_1 \vec{v}_1 e^{\lambda t} + c_2 (\vec{v}_2 + \vec{v}_1 t)\, e^{\lambda t}.$$

Example 4.6.2: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} -2 & 3 \\ -3 & 4 \end{bmatrix} \vec{x} \qquad \vec{x}(0) = \begin{bmatrix} 2 \\ -1 \end{bmatrix}.$$

Solution: First, we need to look for the eigenvalues of the coefficient matrix. These are found by

$$\det(A - \lambda I) = (-2 - \lambda)(4 - \lambda) - (3)(-3) = \lambda^2 - 2\lambda + 1 = 0.$$

Since this polynomial is $(\lambda - 1)^2$, this has a double root at $\lambda = 1$.

For $\lambda = 1$, we can hunt for the eigenvector as solutions to

$$(A - I)\vec{v} = \begin{bmatrix} -3 & 3 \\ -3 & 3 \end{bmatrix} \vec{v} = 0.$$

These two equations are redundant, and the first equation is $-3v_1 + 3v_2 = 0$, which can be solved by $v_1 = v_2 = 1$. Therefore, and an eigenvector for $\lambda = 1$ is $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$. Thus, we have a solution to this system of the form

$$\vec{x}_1(t) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^t.$$

Since we only found one eigenvector, we need to look for a generalized eigenvector as well. To do this, we want to solve the equation

$$(A - I)\vec{w} = \vec{v}$$

for the eigenvector $\vec{v}$ that we found previously. This means we need to solve

$$\begin{bmatrix} -3 & 3 \\ -3 & 3 \end{bmatrix} \vec{w} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and both rows of the vector equation result in the equation $-3w_1 + 3w_2 = 1$ for $\vec{w}$. We can pick *any* value of w_1 and w_2 to make this work. For the sake of this example, we will pick $w_1 = 0$ and $w_2 = 1/3$. Then, we have that our second linearly independent solution to the differential equation is

$$\vec{x}_2(t) = \left(\begin{bmatrix} 0 \\ 1/3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} t \right) e^t$$

and so the general solution to this system is

$$\vec{x}(t) = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^t + c_2 \left(\begin{bmatrix} 0 \\ 1/3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} t \right) e^t.$$

Finally, we can solve the initial value problem. Plugging in $t = 0$ gives

$$\vec{x}(0) = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 1/3 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix},$$

which gives that $c_1 = 2$ and then $2 + 1/3 c_2 = -1$, or $c_2 = -9$. Therefore, the solution to the initial value problem is

$$\vec{x}(t) = 2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^t - 9 \left(\begin{bmatrix} 0 \\ 1/3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} t \right) e^t.$$

***Exercise* 4.6.2:** *We could have also chosen $w_1 = -1/3$ and $w_2 = 0$ for the vector $\vec{w}$. Use this to get a different looking general solution. Then solve the same initial value problem to see that you end up with the same answer at the end of the process.*

Example 4.6.3: Consider the system

$$\vec{x}' = \begin{bmatrix} 2 & -5 & 0 \\ 0 & 2 & 0 \\ -1 & 4 & 1 \end{bmatrix} \vec{x}.$$

Find the general solution to this system using eigenvalues and eigenvectors.

Solution: Even though this is a three-component system, the process is exactly the same: find the eigenvalues, compute corresponding eigenvectors, then build them together into a general solution. Compute the eigenvalues,

$$0 = \det(A - \lambda I) = \det\left(\begin{bmatrix} 2-\lambda & -5 & 0 \\ 0 & 2-\lambda & 0 \\ -1 & 4 & 1-\lambda \end{bmatrix} \right) = (2-\lambda)^2(1-\lambda).$$

The eigenvalues are 1 and 2, where 2 has multiplicity 2. We leave it to the reader to find that $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ is an eigenvector for the eigenvalue $\lambda = 1$.

Let's focus on $\lambda = 2$. We compute eigenvectors:

$$\vec{0} = (A - 2I)\vec{v} = \begin{bmatrix} 0 & -5 & 0 \\ 0 & 0 & 0 \\ -1 & 4 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

The first equation says that $v_2 = 0$, so the last equation is $-v_1 - v_3 = 0$. Let v_3 be the free variable to find that $v_1 = -v_3$. Perhaps let $v_3 = -1$ to find an eigenvector $\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$. Problem is that setting v_3 to anything else just gets multiples of this vector and so we have a defect of 1. Let $\vec{v}_1$ be the eigenvector and let's look for a generalized eigenvector $\vec{v}_2$:

$$(A - 2I)\vec{v}_2 = \vec{v}_1,$$

or

$$\begin{bmatrix} 0 & -5 & 0 \\ 0 & 0 & 0 \\ -1 & 4 & -1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix},$$

where we used a, b, c as components of $\vec{v}_2$ for simplicity. The first equation says $-5b = 1$ so $b = -1/5$. The second equation says nothing. The last equation is $-a + 4b - c = -1$, or $a + 4/5 + c = 1$, or $a + c = 1/5$. We let c be the free variable and we choose $c = 0$. We find $\vec{v}_2 = \begin{bmatrix} 1/5 \\ -1/5 \\ 0 \end{bmatrix}$.

The general solution is therefore,

$$\vec{x} = c_1 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} e^t + c_2 \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} e^{2t} + c_3 \left(\begin{bmatrix} 1/5 \\ -1/5 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} t \right) e^{2t}.$$

This machinery can also be generalized to higher multiplicities and higher defects. We will not go over this method in detail, but let us just sketch the ideas. Suppose that A has an eigenvalue λ of multiplicity m. We find vectors such that

$$(A-\lambda I)^k \vec{v}_k = \vec{0}, \qquad \text{but} \qquad (A-\lambda I)^{k-1}\vec{v}_k \neq \vec{0}.$$

Such vectors are called *generalized eigenvectors* (then $\vec{v}_1 = (A-\lambda I)^{k-1}\vec{v}_k$ is an eigenvector). For the eigenvector $\vec{v}_1$ there is a chain of generalized eigenvectors $\vec{v}_2$ through $\vec{v}_k$ such that:

$$\begin{aligned}(A-\lambda I)\vec{v}_1 &= \vec{0},\\ (A-\lambda I)\vec{v}_2 &= \vec{v}_1,\\ &\vdots\\ (A-\lambda I)\vec{v}_k &= \vec{v}_{k-1}.\end{aligned}$$

Really once you find the $\vec{v}_k$ such that $(A-\lambda I)^k\vec{v}_k = \vec{0}$ but $(A-\lambda I)^{k-1}\vec{v}_k \neq \vec{0}$, you find the entire chain since you can compute the rest, $\vec{v}_{k-1} = (A-\lambda I)\vec{v}_k$, $\vec{v}_{k-2} = (A-\lambda I)\vec{v}_{k-1}$, etc. We form the linearly independent solutions

$$\begin{aligned}\vec{x}_1 &= \vec{v}_1 e^{\lambda t},\\ \vec{x}_2 &= (\vec{v}_2 + \vec{v}_1 t)\, e^{\lambda t},\\ &\vdots\\ \vec{x}_k &= \left(\vec{v}_k + \vec{v}_{k-1}t + \vec{v}_{k-2}\frac{t^2}{2} + \cdots + \vec{v}_2\frac{t^{k-2}}{(k-2)!} + \vec{v}_1\frac{t^{k-1}}{(k-1)!}\right) e^{\lambda t}.\end{aligned}$$

Recall that $k! = 1\cdot 2\cdot 3\cdots(k-1)\cdot k$ is the factorial. If you have an eigenvalue of geometric multiplicity ℓ, you will have to find ℓ such chains (some of them might be short: just the single eigenvector equation). We go until we form m linearly independent solutions where m is the algebraic multiplicity. We don't quite know which specific eigenvectors go with which chain, so start by finding $\vec{v}_k$ first for the longest possible chain and go from there.

For example, if λ is an eigenvalue of A of algebraic multiplicity 3 and defect 2, then solve

$$(A-\lambda I)\vec{v}_1 = \vec{0}, \qquad (A-\lambda I)\vec{v}_2 = \vec{v}_1, \qquad (A-\lambda I)\vec{v}_3 = \vec{v}_2.$$

That is, find $\vec{v}_3$ such that $(A-\lambda I)^3\vec{v}_3 = \vec{0}$, but $(A-\lambda I)^2\vec{v}_3 \neq \vec{0}$. Then you are done as $\vec{v}_2 = (A-\lambda I)\vec{v}_3$ and $\vec{v}_1 = (A-\lambda I)\vec{v}_2$. The 3 linearly independent solutions are

$$\vec{x}_1 = \vec{v}_1 e^{\lambda t}, \qquad \vec{x}_2 = (\vec{v}_2 + \vec{v}_1 t)\, e^{\lambda t}, \qquad \vec{x}_3 = \left(\vec{v}_3 + \vec{v}_2 t + \vec{v}_1\frac{t^2}{2}\right) e^{\lambda t}.$$

If on the other hand A has an eigenvalue λ of algebraic multiplicity 3 and defect 1, then solve

$$(A-\lambda I)\vec{v}_1 = \vec{0}, \qquad (A-\lambda I)\vec{v}_2 = \vec{0}, \qquad (A-\lambda I)\vec{v}_3 = \vec{v}_2.$$

Here $\vec{v}_1$ and $\vec{v}_2$ are actual honest eigenvectors, and $\vec{v}_3$ is a generalized eigenvector. So there are two chains. To solve, first find a $\vec{v}_3$ such that $(A-\lambda I)^2\vec{v}_3 = \vec{0}$, but $(A-\lambda I)\vec{v}_3 \neq \vec{0}$. Then $\vec{v}_2 = (A-\lambda I)\vec{v}_3$ is going to be an eigenvector. Then solve for an eigenvector $\vec{v}_1$ that is linearly independent from $\vec{v}_2$. You get 3 linearly independent solutions

$$\vec{x}_1 = \vec{v}_1 e^{\lambda t}, \qquad \vec{x}_2 = \vec{v}_2 e^{\lambda t}, \qquad \vec{x}_3 = (\vec{v}_3 + \vec{v}_2 t)\, e^{\lambda t}.$$

4.6.1 Phase Portraits

We also want to look at the phase portraits and direction field diagrams for repeated eigenvalues. There are two different options here, depending on if there are two linearly independent eigenvectors or not.

Case 1. If we have a repeated eigenvalue with two linearly independent eigenvectors, this means that our matrix A is of the form

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

for the repeated eigenvalue λ. This means that $A\vec{v} = \lambda\vec{v}$ for all vectors $\vec{v}$. So, every vector is part of a straight line solution, and so every solution goes either directly towards or directly away from the origin. This gives a *proper node* which can be a sink or a source depending on whether the eigenvalue is positive or negative.

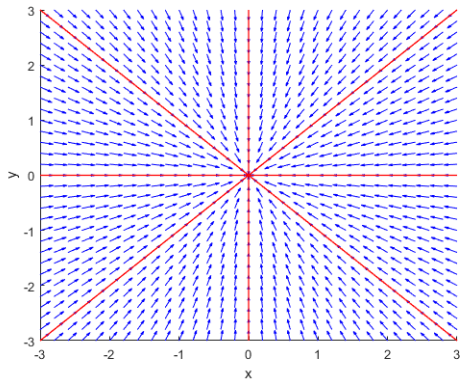


Figure 4.13: Example proper nodal sink vector field.

Case 2. If we have a repeated eigenvalue with only one linearly independent eigenvector, then we only have one straight-line solution. For instance, the matrix

$$A = \begin{bmatrix} 4 & -1 \\ 1 & 2 \end{bmatrix}$$

has only one eigenvector of $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ for eigenvalue 3. Like the nodal sources and sinks, the solutions will go to zero and infinity along the straight line solutions. In this case, because there is only one straight line, the phase portrait looks somewhere between a node and a spiral. This gives an *improper node* which can be a source or sink depending on the sign of the eigenvalue.

4.6.2 Exercises

***Exercise* 4.6.3:** *Compute eigenvalues and eigenvectors of* $\begin{bmatrix} -2 & -1 & -1 \\ 3 & 2 & 1 \\ -3 & -1 & 0 \end{bmatrix}$.

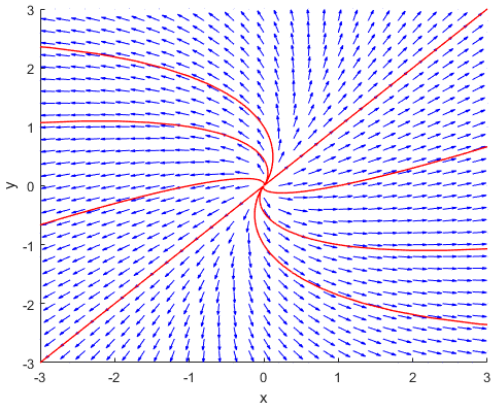


Figure 4.14: Example improper nodal source vector field.

***Exercise* 4.6.4:** *Let* $A = \left[\begin{smallmatrix} 5 & -3 \\ 3 & -1 \end{smallmatrix}\right]$. *Find the general solution of* $\vec{x}' = A\vec{x}$.

***Exercise* 4.6.5:** *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} -3 & 2 \\ 0 & -3 \end{bmatrix} \vec{x} \qquad \vec{x}(0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

***Exercise* 4.6.6:** *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} -5 & -2 \\ 8 & 3 \end{bmatrix} \vec{x} \qquad \vec{x}(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

***Exercise* 4.6.7:** *Assume* A *is a* 3×3 *matrix. The row-reduced echelon forms of* $A - \lambda I$ *are given for three different values of* λ*:*

$$A - 3I \rightsquigarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad A - 5I \rightsquigarrow \begin{pmatrix} 1 & 4 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \qquad A - 7I \rightsquigarrow \begin{pmatrix} 1 & -1 & 1/2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Find the general solution of the homogeneous system $\vec{x}' = A\vec{x}$.

***Exercise* 4.6.8:** *Consider the matrix* $A = \begin{pmatrix} 7 & 5 & -6 \\ 0 & -3 & 2 \\ 0 & -4 & 1 \end{pmatrix}$

a) *Determine the characteristic polynomial of* A *and give its eigenvalues.*

b) *How many (linearly independent) straight-line solutions does the system* $\vec{x}' = A\vec{x}$ *have? How do you know, without solving?*

***Exercise* 4.6.9:** *Let* $A = \begin{bmatrix} 1 & 5 & -18 \\ 2 & -1 & -5 \\ 1 & 1 & -6 \end{bmatrix}$.

a) *Show directly that* $\vec{v}_1 = \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix}$ *is an eigenvector of* A.

b) *All eigenvalues of* A *are the same. Find the general solution to* $\vec{x}' = A\vec{x}$.

Exercise 4.6.10: *Let* $A = \begin{bmatrix} 5 & -4 & 4 \\ 0 & 3 & 0 \\ -2 & 4 & -1 \end{bmatrix}$.

a) *What are the eigenvalues?*

b) *What is/are the defect(s) of the eigenvalue(s)?*

c) *Find the general solution of* $\vec{x}' = A\vec{x}$.

Exercise 4.6.11:* *Let* $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$.

a) *What are the eigenvalues?*

b) *What is/are the defect(s) of the eigenvalue(s)?*

c) *Find the general solution of* $\vec{x}\,' = A\vec{x}$.

Exercise 4.6.12: *Let* $A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$.

a) *What are the eigenvalues?*

b) *What is/are the defect(s) of the eigenvalue(s)?*

c) *Find the general solution of* $\vec{x}' = A\vec{x}$ *in two different ways and verify you get the same answer.*

Exercise 4.6.13:* *Let* $A = \begin{bmatrix} 1 & 3 & 3 \\ 1 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix}$.

a) *What are the eigenvalues?*

b) *What is/are the defect(s) of the eigenvalue(s)?*

c) *Find the general solution of* $\vec{x}\,' = A\vec{x}$.

Exercise 4.6.14: *Let* $A = \begin{bmatrix} 0 & 1 & 2 \\ -1 & -2 & -2 \\ -4 & 4 & 7 \end{bmatrix}$.

a) *What are the eigenvalues?*

b) *What is/are the defect(s) of the eigenvalue(s)?*

c) *Find the general solution of* $\vec{x}' = A\vec{x}$.

Exercise 4.6.15:* *Let* $A = \begin{bmatrix} 2 & 0 & 0 \\ -1 & -1 & 9 \\ 0 & -1 & 5 \end{bmatrix}$.

a) *What are the eigenvalues?*

b) *What is/are the defect(s) of the eigenvalue(s)?*

c) *Find the general solution of* $\vec{x}\,' = A\vec{x}$.

***Exercise* 4.6.16:** *Let* $A = \begin{bmatrix} 0 & 4 & -2 \\ -1 & -4 & 1 \\ 0 & 0 & -2 \end{bmatrix}$.

a) What are the eigenvalues?

b) What is/are the defect(s) of the eigenvalue(s)?

c) Find the general solution of $\vec{x}' = A\vec{x}$.

***Exercise* 4.6.17:** *Let* $A = \begin{bmatrix} 2 & 1 & -1 \\ -1 & 0 & 2 \\ -1 & -2 & 4 \end{bmatrix}$.

a) What are the eigenvalues?

b) What is/are the defect(s) of the eigenvalue(s)?

c) Find the general solution of $\vec{x}' = A\vec{x}$.

***Exercise* 4.6.18:** *Suppose that* A *is a* 2×2 *matrix with a repeated eigenvalue* λ. *Suppose that there are two linearly independent eigenvectors. Show that* $A = \lambda I$.

Exercise* 4.6.19: *Let* $A = \begin{bmatrix} a & a \\ b & c \end{bmatrix}$, *where* a, b, *and* c *are unknowns. Suppose that* 5 *is a doubled eigenvalue of defect 1, and suppose that* $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ *is a corresponding eigenvector. Find* A *and show that there is only one such matrix* A.

Exercise* 4.6.20: *For each system, (i) classify the system according to type as one of sink/source/saddle/center/ spiral source/spiral sink; (ii) solve the systems; (iii) sketch the phase portrait. Both real and complex eigenvalues appear.*

a) $\vec{x}' = \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix} \vec{x}$

b) $\vec{x}' = \begin{pmatrix} 3 & 2 \\ 0 & -2 \end{pmatrix} \vec{x}$

c) $\vec{x}' = \begin{pmatrix} -2 & -2 \\ 3 & -2 \end{pmatrix} \vec{x}$

d) $\vec{x}' = \begin{pmatrix} 3 & 5 \\ -5 & -3 \end{pmatrix} \vec{x}$

e) $\vec{x}' = \begin{pmatrix} 2 & 1/2 \\ -1 & 1 \end{pmatrix} \vec{x}$

f) $\vec{x}' = \begin{pmatrix} 3 & 3/2 \\ 3/2 & -1 \end{pmatrix} \vec{x}$

4.7 Two-dimensional systems and their vector fields

Attribution: [JL], §3.5.

Learning Objectives

After this section, you will be able to:

- Visualize and sketch the behavior of a two dimensional system based on the eigenvalues and eigenvectors.

In the last three sections, we looked at the different options for two-component constant-coefficient systems. We want to determine a nice way to put all of this together. We summarize the behavior of linear homogeneous two-dimensional systems given by a nonsingular matrix in Table 4.1. Systems where one of the eigenvalues is zero (the matrix is singular) come up in practice from time to time, see Example 4.1.2 on page 271, and the pictures are somewhat different (simpler in a way). See the exercises.

Eigenvalues	Behavior
real and both positive	source / unstable node
real and both negative	sink / asymptotically stable node
real and opposite signs	saddle
purely imaginary	center point / ellipses
complex with positive real part	spiral source
complex with negative real part	spiral sink
repeated with two eigenvectors	proper node (asympt. stable or unstable)
repeated with one eigenvector	improper node (asympt. stable or unstable)

Table 4.1: Summary of behavior of linear homogeneous two-dimensional systems.

The sketches of all of these different behaviors and phase portraits can be found in their respective sections. Make sure that you understand the terminology, general behavior, and sketches for each of these different cases.

4.7.1 Trace-Determinant Analysis

One other way to interpret and analyze this information is using the trace and determinant of the matrix. Recall from § 3.7 that the trace of a matrix is the sum of the diagonal entries of the matrix and the determinant of the matrix is computed from the entries and is a way to determine invertibility of the matrix. If we take a generic 2×2 matrix and find the characteristic polynomial, we get that for

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

the characteristic polynomial is

$$\det(A - \lambda I) = (a - \lambda)(d - \lambda) - (b)(c) = \lambda^2 - (a + d)\lambda + (ad - bc).$$

Since the trace of the matrix is $a + d$ and the determinant is $ad - bc$, we can rewrite this polynomial as

$$\lambda^2 - T\lambda + D = 0,$$

which also means that we can characterize the eigenvalues of the matrix in terms of the trace and determinant. We get that the eigenvalues are

$$\lambda = \frac{T \pm \sqrt{T^2 - 4D}}{2}. \tag{4.5}$$

There are a few important facts we can learn from this equation.

1. A lot depends on the value of $T^2 - 4D$. If $T^2 - 4D > 0$, then we will have two real distinct eigenvalues. If $T^2 - 4D = 0$, then there is a single repeated eigenvalue, and if $T^2 - 4D < 0$, we have complex eigenvalues.

2. If $D < 0$, then $T^2 - 4D > T^2$, which means that $\sqrt{T^2 - 4D} > |T|$. If we put this into (4.5), this will mean that the term that is after the $\pm$ will be larger than T in absolute value. Therefore, the two eigenvalues will be real and have opposite signs.

3. If $D \geq 0$, then the sign of the eigenvalues, or the sign of the real part in the complex case, is dictated by the sign of T. If $D \geq 0$, then $T^2 - 4D \leq T^2$, so that the part under the square root in (4.5) is always smaller in absolute value than T. Thus, both the plus and minus version will have values that are the same sign as T. If the expression is complex, then the real part is exactly $T/2$, which is the same sign as T.

All of this means we can make a new table characterizing the eigenvalues and how they are connected to the trace and determinant.

Eigenvalues	Trace and Determinant Classification
real and both positive	$T > 0$, $D > 0$, $T^2 - 4D > 0$
real and both negative	$T < 0$, $D > 0$, $T^2 - 4D > 0$
real and opposite signs	$D < 0$
purely imaginary	$T = 0$, $D > 0$
complex with positive real part	$T > 0$, $T^2 - 4D < 0$
complex with negative real part	$T < 0$, $T^2 - 4D < 0$
repeated	$T^2 - 4D = 0$

Table 4.2: Summary of behavior of linear homogeneous two-dimensional systems.

Since these are all based on the relation between T and D, we can also combine all of this into a figure to summarize the details. In Figure 4.15 on the following page, T is on

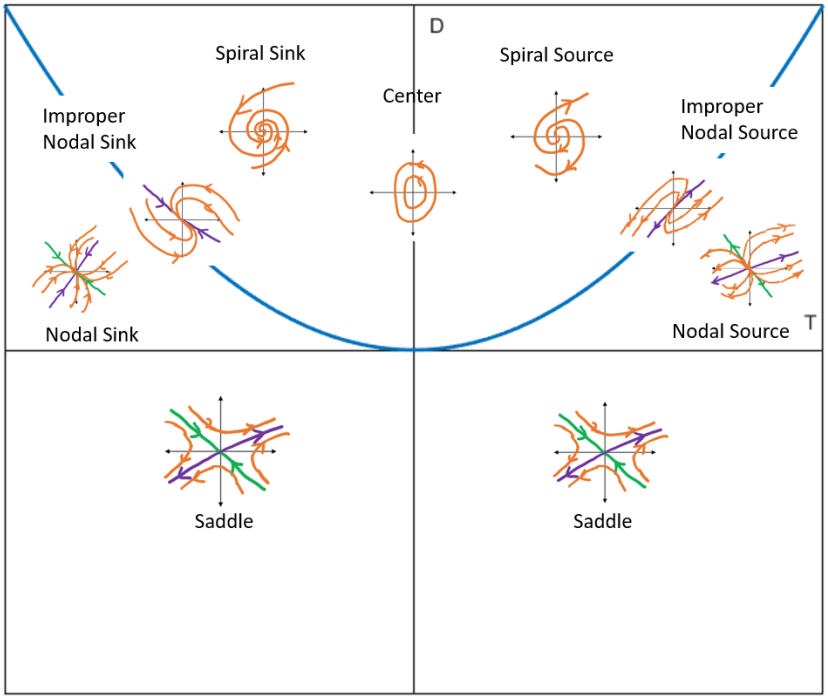


Figure 4.15: Trace-Determinant plane for analysis of two-component linear systems.

the horizontal axis and D is the vertical axis. The graph drawn is $D = {}^{T^2}/_4$, which is the important criteria that shows up in the table.

Figure 4.15 can be used to determine the behavior of a two-component system without actually needing to solve the differential equation. The point is that the signs and type of the eigenvalues determine the structure of the solution, and we can determine the important qualities of these using just the trace and determinant of a matrix.

Example 4.7.1: Use Trace-Determinant analysis to determine the overall behavior of the system

$$\vec{x}' = \begin{bmatrix} 1 & 4 \\ -2 & 3 \end{bmatrix} \vec{x}.$$

Solution: From the matrix, we can see that the trace is $1 + 3 = 4$ and the determinant is $(1)(3) - (4)(-2) = 11$. We see that $D > 0$ with $T^2 = 16$ and $4D = 44 > 16$. Therefore, we have $4D > T^2$, so we are above the curve on the graph, and so have a spiral. Since $T > 0$, this will be a spiral source.

Note: If you wanted to get a general solution or sketch a phase portrait for this differential equation, you would need to actually solve it out for that; you can not get enough information just from this image to sketch a proper phase portrait. ⌟

***Exercise* 4.7.1:** *Compute the eigenvalues for the system above, find the general solution, and verify that this is a spiral source. The numbers here will not work out great, so having the quick analysis that it is a spiral source is nice.*

4.7.2 Exercises

***Exercise* 4.7.2:** *Take the equation $mx'' + cx' + kx = 0$, with $m > 0$, $c \geq 0$, $k > 0$ for the mass-spring system.*

a) *Convert this to a system of first order equations.*

b) *Classify for what m, c, k do you get which behavior.*

c) *Can you explain from physical intuition why you do not get all the different kinds of behavior here?*

***Exercise* 4.7.3:** *What happens in the case when $P = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$? In this case the eigenvalue is repeated and there is only one independent eigenvector. What picture does this look like?*

***Exercise* 4.7.4:** *What happens in the case when $P = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$? Does this look like any of the pictures we have drawn?*

Exercise* 4.7.5: *Describe the behavior of the following systems without solving:*

a) $x' = x + y, \quad y' = x - y.$

b) $x_1' = x_1 + x_2, \quad x_2' = 2x_2.$

c) $x_1' = -2x_2, \quad x_2' = 2x_1.$

d) $x' = x + 3y, \quad y' = -2x - 4y.$

e) $x' = x - 4y, \quad y' = -4x + y.$

***Exercise* 4.7.6:** *Which behaviors are possible if P is diagonal, that is $P = \left[\begin{smallmatrix} a & 0 \\ 0 & b \end{smallmatrix}\right]$? You can assume that a and b are not zero.*

Exercise* 4.7.7: *Suppose that $\vec{x}\,' = A\vec{x}$ where A is a 2 by 2 matrix with eigenvalues $2 \pm i$. Describe the behavior.*

Exercise* 4.7.8: *For each of the following matrices A, describe the behavior and classify the phase portrait of the system given by $\vec{x}' = A\vec{x}$. Use the eigenvalues to determine this.*

a) $A = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix}$

b) $A = \begin{bmatrix} 3 & 5 \\ -1 & 1 \end{bmatrix}$

c) $A = \begin{bmatrix} 8 & -18 \\ 4 & -10 \end{bmatrix}$

d) $A = \begin{bmatrix} -2 & -4 \\ 0 & -3 \end{bmatrix}$

e) $A = \begin{bmatrix} 3 & -2 \\ 2 & -3 \end{bmatrix}$

f) $A = \begin{bmatrix} -3 & -4 \\ 1 & 1 \end{bmatrix}$

***Exercise* 4.7.9:** *For each of the matrices and systems in Exercise 4.7.8, perform the same analysis using the trace and determinant of the matrix.*

***Exercise* 4.7.10:** *Consider the system of differential equations given by*

$$\vec{x}' = \begin{bmatrix} -2 & -3 \\ 3 & -2 \end{bmatrix} \vec{x}.$$

a) *Use trace and determinant analysis to determine the behavior of this linear system.*

b) *Find the general solution to this system of differential equations and verify that it matches the analysis in (a).*

***Exercise* 4.7.11:** *Consider the system of differential equations given by*

$$\vec{x}' = \begin{bmatrix} -2 & -3 \\ 3 & 4 \end{bmatrix} \vec{x}.$$

a) *Use trace and determinant analysis to determine the behavior of this linear system.*

b) *Find the general solution to this system of differential equations and verify that it matches the analysis in (a).*

***Exercise* 4.7.12:** *Take the system from Example 4.1.2 on page 271, $x_1' = \frac{r}{V}(x_2 - x_1)$, $x_2' = \frac{r}{V}(x_1 - x_2)$. As we said, one of the eigenvalues is zero. What is the other eigenvalue, how does the picture look like and what happens when t goes to infinity.*

Exercise* 4.7.13: *Take $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]' = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right] \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$. Draw the vector field and describe the behavior. Is it one of the behaviors that we have seen before?*

***Exercise* 4.7.14:** *In this exercise, we will analyze "perturbations" or near-by matrices to the ones that are given. This will be important later in § 5.1. For each of the following matrices*

a) *Find the trace and determinant, and use them to classify the behavior of the linear system $\vec{x}' = A\vec{x}$ for the given matrix A.*

b) *Draw a sketch of the trace-determinant plane, including the curve $D = {}^{T^2}/_4$, and plot the point corresponding to the matrix on those axes.*

c) *Look at the points in a small (as small as you want) circle around the point you just drew. What does the behavior look like for systems whose matrices fall within that circle? What do these behaviors have in common with each other, and how do they differ?*

$$(i)\ \begin{bmatrix} 2 & 8 \\ -3 & -8 \end{bmatrix} \quad (ii)\ \begin{bmatrix} 15 & -12 \\ 16 & -13 \end{bmatrix} \quad (iii)\ \begin{bmatrix} 2 & -1 \\ 5 & -2 \end{bmatrix} \quad (iv)\ \begin{bmatrix} 4 & -1 \\ 2 & 2 \end{bmatrix}$$

$$(v)\ \begin{bmatrix} -2 & 2 \\ -2 & -6 \end{bmatrix} \quad (vi)\ \begin{bmatrix} -1 & -4 \\ 2 & 5 \end{bmatrix} \quad (vii)\ \begin{bmatrix} -5 & 6 \\ -3 & 1 \end{bmatrix} \quad (viii)\ \begin{bmatrix} 5 & -2 \\ 8 & -3 \end{bmatrix}$$

4.8 Nonhomogeneous systems

Attribution: [JL], §3.9.

Learning Objectives

After this section, you will be able to:

- Use the eigenvector decomposition or diagonalization to solve non-homogeneous systems,
- Use undetermined coefficients to solve non-homogeneous systems, and
- Use variation of parameters and fundamental matrices of solutions to solve non-homogeneous systems.

Now, we want to take a look at solving non-homogeneous linear systems. As discussed previously, the process here is the same as it was for second order non-homogeneous equations. We can solve the homogeneous equation and then need one particular solution to the non-homogeneous problem. Adding these together gives the general solution to the non-homogeneous problem, where we can pick constants to meet an initial condition if it is given. This section here will focus on a variety of methods to find this particular solution.

4.8.1 First order constant coefficient

Diagonalization

Diagonalization is a linear algebra-based process for adjusting a matrix into one that is diagonal. In order to see why this might be helpful in the process of solving non-homogeneous systems, or generating a particular solution to the non-homogeneous system, let's start by looking at a problem with a diagonal matrix to see how we could solve it.

Example 4.8.1: Find the general solution of the non-homogeneous system

$$\vec{x}'(t) = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{2t} \\ e^{-t} \end{bmatrix}.$$

Solution: If we write this system out in components, we get

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e^{2t} \\ e^{-t} \end{bmatrix},$$

or

$$x_1' = x_1 + e^{2t} \qquad x_2' = 3x_2 + e^{-t}.$$

These are two completely separated, or *decoupled* equations. We can solve each of these via first-order integrating factor methods. For the first, we get

$$\begin{aligned} x_1' - x_1 &= e^{2t} \\ (e^{-t}x_1)' &= e^t \\ e^{-t}x_1 &= e^t + C_1 \\ x_1(t) &= e^{2t} + C_1 e^t \end{aligned}$$

and for the second, we see that

$$\begin{aligned} x_2' - 3x_2 &= e^{-t} \\ (e^{-3t}x_2)' &= e^{2t} \\ e^{-3t}x_2 &= \frac{1}{2}e^{2t} + C_2 \\ x_2(t) &= \frac{1}{2}e^{-t} + C_2e^{3t} \end{aligned} \quad .$$

Therefore, the solution to this system is

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} e^{2t} + C_1e^t \\ \frac{1}{2}e^{-t} + C_2e^{3t} \end{bmatrix}$$

or, rewriting in a different form,

$$\vec{x}(t) = \begin{bmatrix} e^{2t} \\ \frac{1}{2}e^{-t} \end{bmatrix} + C_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^t + C_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{3t}.$$

Therefore, if we have a non-homogeneous system with a diagonal matrix, then we can separate the decoupled equations, solve them individually, and put them back together into a full solution. In this particular case, the eigenvectors of A were $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, and so the standard basis vectors were the directions in which A acts like a scalar. When the eigenvectors are not the standard basis vectors, we need to take them into account in order to use this method.

Take the equation

$$\vec{x}'(t) = A\vec{x}(t) + \vec{f}(t). \tag{4.6}$$

Assume A has n linearly independent eigenvectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n$ with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Build the matrices

$$E = [\vec{v}_1 \mid \vec{v}_2 \mid \cdots \mid \vec{v}_n] \qquad D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix},$$

that is, E is the matrix with the eigenvectors as columns, and D is a diagonal matrix with the eigenvalues on the diagonal in the same order as the eigenvectors are put into E. Since we have n eigenvectors, both of these are $n \times n$ square matrices. It is a fact from linear algebra that

$$A = EDE^{-1} \quad \text{or} \quad D = E^{-1}AE.$$

***Exercise* 4.8.1:** *For the matrix*

$$A = \begin{bmatrix} 6 & 2 \\ -4 & 0 \end{bmatrix}$$

compute the matrices E and D and verify that $EDE^{-1} = A$.

With this tool in hand, we look to approach our non-homogeneous system. We would like for the system to use the matrix D instead of the matrix A, because that is decoupled and we can solve it directly. To do this, we define a new unknown function $\vec{y}$ by the relation $\vec{x} = E\vec{y}$. If we plug this into (4.6), we get

$$E\vec{y}'(t) = AE\vec{y}(t) + \vec{f}(t).$$

Using the relation for A and the fact that E is a constant matrix, we get that

$$E\vec{y}'(t) = EDE^{-1}E\vec{y}(t) + \vec{f}(t) = ED\vec{y}(t) + \vec{f}(t).$$

If we multiply both sides of this equation by E^{-1}, we get

$$\vec{y}'(t) = D\vec{y}(t) + E^{-1}\vec{f}(t)$$

and this is now a decoupled system of equations. Once we compute $E^{-1}\vec{f}(t)$, we can then solve this directly because it is based on a decoupled system of differential equations to solve for the solution $\vec{y}$. Once we have $\vec{y}$, we can compute $\vec{x}$ as $\vec{x} = E\vec{y}$ to get our solution.

Example 4.8.2: Let $A = \left[\begin{smallmatrix} 1 & 3 \\ 3 & 1 \end{smallmatrix}\right]$. Solve $\vec{x}' = A\vec{x} + \vec{f}$ where $\vec{f}(t) = \left[\begin{smallmatrix} 2e^t \\ 2t \end{smallmatrix}\right]$ for $\vec{x}(0) = \left[\begin{smallmatrix} 3/16 \\ -5/16 \end{smallmatrix}\right]$.

Solution: The first step in this process is always to find the eigenvalues and eigenvectors of the coefficient matrix. We do this in the standard way

$$\det(A - \lambda I) = (1-\lambda)(1-\lambda) - (3)(3) = \lambda^2 - 2\lambda + 1 - 9 = \lambda^2 - 2\lambda - 8.$$

Since this factors as $(\lambda + 2)(\lambda - 4)$, the eigenvalues are -2 and 4. Using these (exercise!) we can show that the corresponding eigenvectors are $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$ for $\lambda = -2$ and $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ for $\lambda = 4$. Therefore, the general solution to the homogeneous problem is

$$\vec{x}(t) = c_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-2t} + c_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{4t}.$$

Now that we have this solution, we can work to solve the non-homogeneous problem. To do this, we form the matrices

$$E = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \qquad D = \begin{bmatrix} -2 & 0 \\ 0 & 4 \end{bmatrix}$$

and, using the fact that for a 2×2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

we can compute E^{-1} as

$$E^{-1} = \frac{1}{(1)(1) - (1)(-1)} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix}.$$

As an aside, we can check that $A = EDE^{-1}$ to make sure that we did this right.

$$\begin{aligned} EDE^{-1} &= \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} -2 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 2 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix} = A. \end{aligned}$$

Thus, we can proceed. From the general process of diagonalization, we know that the system we need to solve is

$$\vec{y}' = D\vec{y} + E^{-1}\vec{f} = \begin{bmatrix} -2 & 0 \\ 0 & 4 \end{bmatrix} \vec{y} + \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 2e^t \\ 2t \end{bmatrix}$$

for $\vec{y} = E^{-1}\vec{x}$, or $\vec{y}$ defined by $\vec{x} = E\vec{y}$. Computing the non-homogenoeus term gives

$$\begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 2e^t \\ 2t \end{bmatrix} = \begin{bmatrix} e^t - t \\ e^t + t \end{bmatrix}$$

so that we can now decouple the system

$$\begin{bmatrix} y_1' \\ y_2' \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} e^t - t \\ e^t + t \end{bmatrix}$$

into two separate first-order equations that we can solve

$$y_1' = -2y_1 + e^t - t \qquad y_2' = 4y_2 + e^t + t$$

by normal first-order integrating factor methods. For the y_1 equation, we want to use an integrating factor of e^{2t} to solve it as

$$\begin{aligned} y_1' + 2y_1 &= e^t - t \\ e^{2t}y_1' + 2e^{2t}y_1 &= e^{3t} - te^{2t} \\ (e^{2t}y_1)' &= e^{3t} - te^{2t} \\ e^{2t}y_1 &= \int e^{3t} - te^{2t}\, dt = \frac{1}{3}e^{3t} - \frac{1}{2}te^{2t} + \frac{1}{4}e^{2t} + C_1 \\ y_1 &= \frac{1}{3}e^t - \frac{1}{2}t + \frac{1}{4} + C_1e^{-2t}. \end{aligned}$$

For the second, we need the integrating factor e^{-4t} to solve

$$\begin{aligned} y_2' - 4y_2 &= e^t + t \\ e^{-4t}y_2' - 4e^{-4t}y_2 &= e^{-3t} + te^{-4t} \\ e^{-4t}y_2 &= \int e^{-3t} + te^{-4t}\, dt = -\frac{1}{3}e^{-3t} - \frac{1}{4}te^{-4t} - \frac{1}{16}e^{-4t} + C_2 \\ y_2 &= -\frac{1}{3}e^t - \frac{1}{4}t - \frac{1}{16} + C_2e^{4t}. \end{aligned}$$

Therefore, we have the vector solution

$$\vec{y}(t) = \begin{bmatrix} \frac{1}{3}e^t - \frac{1}{2}t + \frac{1}{4} + C_1e^{-2t} \\ -\frac{1}{3}e^t - \frac{1}{4}t - \frac{1}{16} + C_2e^{4t} \end{bmatrix}.$$

To get to the actual solution $\vec{x}$, we need to multiply this solution by the matrix E

$$\begin{aligned} \vec{x} = E\vec{y} &= \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{3}e^t - \frac{1}{2}t + \frac{1}{4} + C_1e^{-2t} \\ -\frac{1}{3}e^t - \frac{1}{4}t - \frac{1}{16} + C_2e^{4t} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{3}e^t - \frac{1}{2}t + \frac{1}{4} + C_1e^{-2t} + (-\frac{1}{3}e^t - \frac{1}{4}t - \frac{1}{16} + C_2e^{4t}) \\ -(\frac{1}{3}e^t - \frac{1}{2}t + \frac{1}{4} + C_1e^{-2t}) + (-\frac{1}{3}e^t - \frac{1}{4}t - \frac{1}{16} + C_2e^{4t}) \end{bmatrix} \\ &= \begin{bmatrix} -\frac{3}{4}t + \frac{3}{16} + C_1e^{-2t} + C_2e^{-4t} \\ -\frac{2}{3}e^t - \frac{1}{4}t - \frac{5}{16} - C_1e^{-2t} + C_2e^{4t} \end{bmatrix} \end{aligned}$$

which is a valid way to write the general solution. We can also write this solution in the form

$$\vec{x}(t) = \begin{bmatrix} 0 \\ -\frac{2}{3} \end{bmatrix} e^t + \begin{bmatrix} -\frac{3}{4} \\ -\frac{1}{4} \end{bmatrix} t + \begin{bmatrix} \frac{3}{16} \\ -\frac{5}{16} \end{bmatrix} + C_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-2t} + C_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{4t}$$

and we see that the general solution to the homogeneous problem shows up at the end of this solution.

Finally, we need to satisfy the initial conditions. If we plug in $t = 0$, we get

$$\vec{x}(0) = \begin{bmatrix} 0 \\ -\frac{2}{3} \end{bmatrix} + 0 + \begin{bmatrix} \frac{3}{16} \\ -\frac{5}{16} \end{bmatrix} + C_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix} + C_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3/16 \\ -5/16 \end{bmatrix}.$$

Rearranging this expression gives the two equations

$$C_1 + C_2 = 0 \qquad -C_1 + C_2 = \frac{2}{3}$$

which has solution $C_1 = -1/3$ and $C_2 = 1/3$. Therefore, the solution to the initial value problem is

$$\vec{x}(t) = \begin{bmatrix} 0 \\ -\frac{2}{3} \end{bmatrix} e^t + \begin{bmatrix} -\frac{3}{4} \\ -\frac{1}{4} \end{bmatrix} t + \begin{bmatrix} \frac{3}{16} \\ -\frac{5}{16} \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-2t} + \frac{1}{3} \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{4t}.$$

Another way to view this process is by thinking about it as eigenvector decomposition. (This approach is not necessary on a first reading. The next new information starts at the undetermined coefficients section.) The eigenvectors of A are the directions in which the matrix A basically acts like a scalar. If we can solve the differential equation in those directions, then it acts like a scalar equation, which we know how to solve. We can then reorient everything to get back to our original solution.

Again, we start with the equation

$$\vec{x}'(t) = A\vec{x}(t) + \vec{f}(t) \tag{4.7}$$

and assume A has n linearly independent eigenvectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n$ Write

$$\vec{x}(t) = \vec{v}_1\,\xi_1(t) + \vec{v}_2\,\xi_2(t) + \cdots + \vec{v}_n\,\xi_n(t). \tag{4.8}$$

That is, we wish to write our solution as a linear combination of eigenvectors of A. If we solve for the scalar functions ξ_1 through ξ_n, we have our solution $\vec{x}$. Let us decompose $\vec{f}$ in terms of the eigenvectors as well. We wish to write

$$\vec{f}(t) = \vec{v}_1\, g_1(t) + \vec{v}_2\, g_2(t) + \cdots + \vec{v}_n\, g_n(t). \tag{4.9}$$

That is, we wish to find g_1 through g_n that satisfy (4.9). Since all the eigenvectors are independent, the matrix $E = [\,\vec{v}_1 \quad \vec{v}_2 \quad \cdots \quad \vec{v}_n\,]$ is invertible. Write the equation (4.9) as $\vec{f} = E\vec{g}$, where the components of $\vec{g}$ are the functions g_1 through g_n. Then $\vec{g} = E^{-1}\vec{f}$. Hence it is always possible to find $\vec{g}$ when there are n linearly independent eigenvectors.

We plug (4.8) into (4.7), and note that $A\vec{v}_k = \lambda_k\vec{v}_k$:

$$\begin{aligned}
\overbrace{\vec{v}_1\xi_1' + \vec{v}_2\xi_2' + \cdots + \vec{v}_n\xi_n'}^{\vec{x}'} &= \overbrace{A\left(\vec{v}_1\xi_1 + \vec{v}_2\xi_2 + \cdots + \vec{v}_n\xi_n\right)}^{A\vec{x}} + \overbrace{\vec{v}_1 g_1 + \vec{v}_2 g_2 + \cdots + \vec{v}_n g_n}^{\vec{f}} \\
&= A\vec{v}_1\xi_1 + A\vec{v}_2\xi_2 + \cdots + A\vec{v}_n\xi_n + \vec{v}_1 g_1 + \vec{v}_2 g_2 + \cdots + \vec{v}_n g_n \\
&= \vec{v}_1\lambda_1\xi_1 + \vec{v}_2\lambda_2\xi_2 + \cdots + \vec{v}_n\lambda_n\xi_n + \vec{v}_1 g_1 + \vec{v}_2 g_2 + \cdots + \vec{v}_n g_n \\
&= \vec{v}_1(\lambda_1\xi_1 + g_1) + \vec{v}_2(\lambda_2\xi_2 + g_2) + \cdots + \vec{v}_n(\lambda_n\xi_n + g_n).
\end{aligned}$$

If we identify the coefficients of the vectors $\vec{v}_1$ through $\vec{v}_n$, we get the equations

$$\begin{aligned}
\xi_1' &= \lambda_1\xi_1 + g_1, \\
\xi_2' &= \lambda_2\xi_2 + g_2, \\
&\vdots \\
\xi_n' &= \lambda_n\xi_n + g_n.
\end{aligned}$$

Each one of these equations is independent of the others. They are all linear first order equations and can easily be solved by the standard integrating factor method for single equations. That is, for the k^{th} equation we write

$$\xi_k'(t) - \lambda_k\xi_k(t) = g_k(t).$$

We use the integrating factor $e^{-\lambda_k t}$ to find that

$$\frac{d}{dt}\left[\xi_k(t)\, e^{-\lambda_k t}\right] = e^{-\lambda_k t} g_k(t).$$

We integrate and solve for ξ_k to get

$$\xi_k(t) = e^{\lambda_k t}\int e^{-\lambda_k t} g_k(t)\, dt + C_k e^{\lambda_k t}.$$

If we are looking for just any particular solution, we can set C_k to be zero. If we leave these constants in, we get the general solution. Write $\vec{x}(t) = \vec{v}_1\xi_1(t) + \vec{v}_2\xi_2(t) + \cdots + \vec{v}_n\xi_n(t)$, and we are done.

As always, it is perhaps better to write these integrals as definite integrals. Suppose that we have an initial condition $\vec{x}(0) = \vec{b}$. Take $\vec{a} = E^{-1}\vec{b}$ to find $\vec{b} = \vec{v}_1 a_1 + \vec{v}_2 a_2 + \cdots + \vec{v}_n a_n$, just like before. Then if we write

$$\boxed{\xi_k(t) = e^{\lambda_k t} \int_0^t e^{-\lambda_k s} g_k(s)\, ds + a_k e^{\lambda_k t},}$$

we get the particular solution $\vec{x}(t) = \vec{v}_1\xi_1(t) + \vec{v}_2\xi_2(t) + \cdots + \vec{v}_n\xi_n(t)$ satisfying $\vec{x}(0) = \vec{b}$, because $\xi_k(0) = a_k$.

Let us remark that the technique we just outlined is the eigenvalue method applied to nonhomogeneous systems. If a system is homogeneous, that is, if $\vec{f} = \vec{0}$, then the equations we get are $\xi_k' = \lambda_k \xi_k$, and so $\xi_k = C_k e^{\lambda_k t}$ are the solutions and that's precisely what we got in § 4.4.

Example 4.8.3: (Same as the previous example) Let $A = \left[\begin{smallmatrix} 1 & 3 \\ 3 & 1 \end{smallmatrix}\right]$. Solve $\vec{x}' = A\vec{x} + \vec{f}$ where $\vec{f}(t) = \left[\begin{smallmatrix} 2e^t \\ 2t \end{smallmatrix}\right]$ for $\vec{x}(0) = \begin{bmatrix} 3/16 \\ -5/16 \end{bmatrix}$.

Solution: The eigenvalues of A are -2 and 4 and corresponding eigenvectors are $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ respectively. We write down the matrix E of the eigenvectors and compute its inverse (using the inverse formula for 2×2 matrices)

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

We are looking for a solution of the form $\vec{x} = \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]\xi_1 + \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]\xi_2$. We first need to write $\vec{f}$ in terms of the eigenvectors. That is we wish to write $\vec{f} = \left[\begin{smallmatrix} 2e^t \\ 2t \end{smallmatrix}\right] = \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right] g_1 + \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right] g_2$. Thus

$$\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = E^{-1}\begin{bmatrix} 2e^t \\ 2t \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 2e^t \\ 2t \end{bmatrix} = \begin{bmatrix} e^t - t \\ e^t + t \end{bmatrix}.$$

So $g_1 = e^t - t$ and $g_2 = e^t + t$.

We further need to write $\vec{x}(0)$ in terms of the eigenvectors. That is, we wish to write $\vec{x}(0) = \begin{bmatrix} 3/16 \\ -5/16 \end{bmatrix} = \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right] a_1 + \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right] a_2$. Hence

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = E^{-1}\begin{bmatrix} 3/16 \\ -5/16 \end{bmatrix} = \begin{bmatrix} 1/4 \\ -1/16 \end{bmatrix}.$$

So $a_1 = 1/4$ and $a_2 = -1/16$. We plug our $\vec{x}$ into the equation and get

$$\begin{aligned} \overbrace{\begin{bmatrix} 1 \\ -1 \end{bmatrix}\xi_1' + \begin{bmatrix} 1 \\ 1 \end{bmatrix}\xi_2'}^{\vec{x}'} &= \overbrace{A\begin{bmatrix} 1 \\ -1 \end{bmatrix}\xi_1 + A\begin{bmatrix} 1 \\ 1 \end{bmatrix}\xi_2}^{A\vec{x}} + \overbrace{\begin{bmatrix} 1 \\ -1 \end{bmatrix} g_1 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} g_2}^{\vec{f}} \\ &= \begin{bmatrix} 1 \\ -1 \end{bmatrix}(-2\xi_1) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} 4\xi_2 + \begin{bmatrix} 1 \\ -1 \end{bmatrix}(e^t - t) + \begin{bmatrix} 1 \\ 1 \end{bmatrix}(e^t + t). \end{aligned}$$

We get the two equations

$$\begin{aligned} \xi_1' &= -2\xi_1 + e^t - t, && \text{where } \xi_1(0) = a_1 = \frac{1}{4}, \\ \xi_2' &= 4\xi_2 + e^t + t, && \text{where } \xi_2(0) = a_2 = \frac{-1}{16}. \end{aligned}$$

We solve with integrating factor. Computation of the integral is left as an exercise to the student. You will need integration by parts.

$$\xi_1 = e^{-2t} \int e^{2t} \left(e^t - t\right) dt + C_1 e^{-2t} = \frac{e^t}{3} - \frac{t}{2} + \frac{1}{4} + C_1 e^{-2t}.$$

C_1 is the constant of integration. As $\xi_1(0) = 1/4$, then $1/4 = 1/3 + 1/4 + C_1$ and hence $C_1 = -1/3$. Similarly

$$\xi_2 = e^{4t} \int e^{-4t} \left(e^t + t\right) dt + C_2 e^{4t} = -\frac{e^t}{3} - \frac{t}{4} - \frac{1}{16} + C_2 e^{4t}.$$

As $\xi_2(0) = -1/16$ we have $-1/16 = -1/3 - 1/16 + C_2$ and hence $C_2 = 1/3$. The solution is

$$\vec{x}(t) = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \underbrace{\left(\frac{e^t - e^{-2t}}{3} + \frac{1-2t}{4} \right)}_{\xi_1} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \underbrace{\left(\frac{e^{4t} - e^t}{3} - \frac{4t+1}{16} \right)}_{\xi_2} = \begin{bmatrix} \frac{e^{4t}-e^{-2t}}{3} + \frac{3-12t}{16} \\ \frac{e^{-2t}+e^{4t}-2e^t}{3} + \frac{4t-5}{16} \end{bmatrix}.$$

That is, $x_1 = \frac{e^{4t}-e^{-2t}}{3} + \frac{3-12t}{16}$ and $x_2 = \frac{e^{-2t}+e^{4t}-2e^t}{3} + \frac{4t-5}{16}$. □

***Exercise* 4.8.2:** *Check that x_1 and x_2 solve the problem. Check both that they satisfy the differential equation and that they satisfy the initial conditions.*

Undetermined coefficients

The method of undetermined coefficients also works for systems. The only difference is that we use unknown vectors rather than just numbers. Same caveats apply to undetermined coefficients for systems as for single equations. This method does not always work for the same reasons that the corresponding method did not work for second order equations. We need to have a right-hand side of a proper form so that we can "guess" a solution of the correct form for the non-homogeneous solution. Furthermore, if the right-hand side is complicated, we have to solve for lots of variables. Each element of an unknown vector is an unknown number. In system of 3 equations with say say 4 unknown vectors (this would not be uncommon), we already have 12 unknown numbers to solve for. The method can turn into a lot of tedious work if done by hand. As the method is essentially the same as for single equations, let us just do an example.

Example 4.8.4: Let $A = \left[\begin{smallmatrix} -1 & 0 \\ -2 & 1 \end{smallmatrix}\right]$. Find a particular solution of $\vec{x}' = A\vec{x} + \vec{f}$ where $\vec{f}(t) = \left[\begin{smallmatrix} e^t \\ t \end{smallmatrix}\right]$.

Solution: Note that we can solve this system in an easier way (can you see how?), but for the purposes of the example, let us use the eigenvalue method plus undetermined coefficients.

The eigenvalues of A are -1 and 1 and corresponding eigenvectors are $\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ respectively. Hence our complementary solution is

$$\vec{x}_c = c_1 \begin{bmatrix}1\\1\end{bmatrix} e^{-t} + c_2 \begin{bmatrix}0\\1\end{bmatrix} e^{t},$$

for some arbitrary constants c_1 and c_2.

We would want to guess a particular solution of

$$\vec{x} = \vec{a}e^t + \vec{b}t + \vec{d}.$$

However, something of the form $\vec{a}e^t$ appears in the complementary solution. Because we do not yet know if the vector $\vec{a}$ is a multiple of $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$, we do not know if a conflict arises. It is possible that there is no conflict, but to be safe we should also try $\vec{k}te^t$. Here we find the crux of the difference between a single equation and systems. We try *both* terms $\vec{a}e^t$ and $\vec{k}te^t$ in the solution, not just the term $\vec{k}te^t$. Therefore, we try

$$\vec{x} = \vec{a}e^t + \vec{k}te^t + \vec{b}t + \vec{d}.$$

Thus we have 8 unknowns. We write $\vec{a} = \left[\begin{smallmatrix}a_1\\a_2\end{smallmatrix}\right]$, $\vec{b} = \left[\begin{smallmatrix}b_1\\b_2\end{smallmatrix}\right]$, $\vec{k} = \left[\begin{smallmatrix}k_1\\k_2\end{smallmatrix}\right]$, and $\vec{d} = \left[\begin{smallmatrix}d_1\\d_2\end{smallmatrix}\right]$. We plug $\vec{x}$ into the equation. First let us compute $\vec{x}'$.

$$\vec{x}' = \left(\vec{a} + \vec{k}\right) e^t + \vec{k}te^t + \vec{b} = \begin{bmatrix}a_1 + k_1\\a_2 + k_2\end{bmatrix} e^t + \begin{bmatrix}k_1\\k_2\end{bmatrix} te^t + \begin{bmatrix}b_1\\b_2\end{bmatrix}.$$

Now $\vec{x}'$ must equal $A\vec{x} + \vec{f}$, which is

$$\begin{aligned} A\vec{x} + \vec{f} &= A\vec{a}e^t + A\vec{k}te^t + A\vec{b}t + A\vec{d} + \vec{f} \\ &= \begin{bmatrix}-a_1\\-2a_1 + a_2\end{bmatrix} e^t + \begin{bmatrix}-k_1\\-2k_1 + k_2\end{bmatrix} te^t + \begin{bmatrix}-b_1\\-2b_1 + b_2\end{bmatrix} t + \begin{bmatrix}-d_1\\-2d_1 + d_2\end{bmatrix} + \begin{bmatrix}1\\0\end{bmatrix} e^t + \begin{bmatrix}0\\1\end{bmatrix} t \\ &= \begin{bmatrix}-a_1 + 1\\-2a_1 + a_2\end{bmatrix} e^t + \begin{bmatrix}-k_1\\-2k_1 + k_2\end{bmatrix} te^t + \begin{bmatrix}-b_1\\-2b_1 + b_2 + 1\end{bmatrix} t + \begin{bmatrix}-d_1\\-2d_1 + d_2\end{bmatrix}. \end{aligned}$$

We identify the coefficients of e^t, te^t, t and any constant vectors in $\vec{x}'$ and in $A\vec{x} + \vec{f}$ to find the equations:

$$\begin{aligned} a_1 + k_1 &= -a_1 + 1, & 0 &= -b_1, \\ a_2 + k_2 &= -2a_1 + a_2, & 0 &= -2b_1 + b_2 + 1, \\ k_1 &= -k_1, & b_1 &= -d_1, \\ k_2 &= -2k_1 + k_2, & b_2 &= -2d_1 + d_2. \end{aligned}$$

We could write the 8×9 augmented matrix and start row reduction, but it is easier to just solve the equations in an ad hoc manner. Immediately we see that $k_1 = 0$, $b_1 = 0$, $d_1 = 0$. Plugging these back in, we get that $b_2 = -1$ and $d_2 = -1$. The remaining equations that tell us something are

$$\begin{aligned} a_1 &= -a_1 + 1, \\ a_2 + k_2 &= -2a_1 + a_2. \end{aligned}$$

So $a_1 = 1/2$ and $k_2 = -1$. Finally, a_2 can be arbitrary and still satisfy the equations. We are looking for just a single solution so presumably the simplest one is when $a_2 = 0$. Therefore,

$$\vec{x} = \vec{a}e^t + \vec{k}te^t + \vec{b}t + \vec{d} = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix} e^t + \begin{bmatrix} 0 \\ -1 \end{bmatrix} te^t + \begin{bmatrix} 0 \\ -1 \end{bmatrix} t + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\, e^t \\ -te^t - t - 1 \end{bmatrix}.$$

That is, $x_1 = \frac{1}{2}\, e^t$, $x_2 = -te^t - t - 1$. We would add this to the complementary solution to get the general solution of the problem. Notice that both $\vec{a}e^t$ and $\vec{k}te^t$ were really needed. □

***Exercise* 4.8.3:** *Check that x_1 and x_2 solve the problem. Try setting $a_2 = 1$ and check we get a solution as well. What is the difference between the two solutions we obtained (one with $a_2 = 0$ and one with $a_2 = 1$)?*

As you can see, other than the handling of conflicts, undetermined coefficients works exactly the same as it did for single equations. However, the computations can get out of hand pretty quickly for systems. The equation we considered was pretty simple.

4.8.2 First order variable coefficient

Variation of parameters

Just as for a single equation, there is the method of variation of parameters. This method works for any linear system, even if it is not constant coefficient, provided we somehow solve the associated homogeneous problem.

Suppose we have the equation

$$\vec{x}' = A(t)\,\vec{x} + \vec{f}(t). \tag{4.10}$$

Further, suppose we solved the associated homogeneous equation $\vec{x}' = A(t)\,\vec{x}$ and found a fundamental matrix solution $X(t)$. If we find separate, linearly independent solutions, this matrix $X(t)$ can be generated by putting these solutions as the columns of a matrix. The general solution to the associated homogeneous equation is $X(t)\vec{c}$ for a constant vector $\vec{c}$. Just like for variation of parameters for single equation we try the solution to the nonhomogeneous equation of the form

$$\vec{x}_p = X(t)\,\vec{u}(t),$$

where $\vec{u}(t)$ is a vector-valued function instead of a constant. We substitute $\vec{x}_p$ into (4.10) to obtain

$$\underbrace{X'(t)\,\vec{u}(t) + X(t)\,\vec{u}'(t)}_{\vec{x}_p'(t)} = \underbrace{A(t)\,X(t)\,\vec{u}(t)}_{A(t)\vec{x}_p(t)} + \vec{f}(t).$$

But $X(t)$ is a fundamental matrix solution to the homogeneous problem. So $X'(t) = A(t)X(t)$, and

$$\cancel{X'(t)\,\vec{u}(t)} + X(t)\,\vec{u}'(t) = \cancel{X'(t)\,\vec{u}(t)} + \vec{f}(t).$$

Hence $X(t)\,\vec{u}'(t) = \vec{f}(t)$. If we compute $[X(t)]^{-1}$, then $\vec{u}'(t) = [X(t)]^{-1}\,\vec{f}(t)$. We integrate to obtain $\vec{u}$ and we have the particular solution $\vec{x}_p = X(t)\,\vec{u}(t)$. Let us write this as a formula

$$\boxed{\vec{x}_p = X(t) \int [X(t)]^{-1}\,\vec{f}(t)\,dt.}$$

Example 4.8.5: Find a particular solution to

$$\vec{x}' = \frac{1}{t^2+1}\begin{bmatrix} t & -1 \\ 1 & t \end{bmatrix}\vec{x} + \begin{bmatrix} t \\ 1 \end{bmatrix}(t^2+1), \tag{4.11}$$

given that the general solution to the homogeneous problem

$$\vec{x}' = \frac{1}{t^2+1}\begin{bmatrix} t & -1 \\ 1 & t \end{bmatrix}\vec{x}$$

is

$$\vec{x}_c(t) = c_1\begin{bmatrix} 1 \\ t \end{bmatrix} + c_2\begin{bmatrix} -t \\ 1 \end{bmatrix}.$$

Solution: Here $A = \frac{1}{t^2+1}\begin{bmatrix} t & -1 \\ 1 & t \end{bmatrix}$ is most definitely not constant, so it's a good thing that we have the general solution to this system. From this, we can build the matrix $X(t)$ as

$$X = \begin{bmatrix} 1 & -t \\ t & 1 \end{bmatrix}$$

, which is a fundamental matrix for this system and solves $X'(t) = A(t)X(t)$. Once we know the complementary solution we can find a solution to (4.11). First we find

$$[X(t)]^{-1} = \frac{1}{t^2+1}\begin{bmatrix} 1 & t \\ -t & 1 \end{bmatrix}.$$

Next we know a particular solution to (4.11) is

$$\begin{aligned}
\vec{x}_p &= X(t)\int [X(t)]^{-1}\vec{f}(t)\,dt \\
&= \begin{bmatrix} 1 & -t \\ t & 1 \end{bmatrix}\int \frac{1}{t^2+1}\begin{bmatrix} 1 & t \\ -t & 1 \end{bmatrix}\begin{bmatrix} t \\ 1 \end{bmatrix}(t^2+1)\,dt \\
&= \begin{bmatrix} 1 & -t \\ t & 1 \end{bmatrix}\int \begin{bmatrix} 2t \\ -t^2+1 \end{bmatrix} dt \\
&= \begin{bmatrix} 1 & -t \\ t & 1 \end{bmatrix}\begin{bmatrix} t^2 \\ -\frac{1}{3}t^3 + t \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{3}t^4 \\ \frac{2}{3}t^3 + t \end{bmatrix}.
\end{aligned}$$

Adding the complementary solution we find the general solution to (4.11):

$$\vec{x} = \begin{bmatrix} 1 & -t \\ t & 1 \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \frac{1}{3}t^4 \\ \frac{2}{3}t^3 + t \end{bmatrix} = \begin{bmatrix} c_1 - c_2 t + \frac{1}{3}t^4 \\ c_2 + (c_1+1)\,t + \frac{2}{3}t^3 \end{bmatrix}.$$

***Exercise* 4.8.4:** *Check that* $x_1 = \frac{1}{3}t^4$ *and* $x_2 = \frac{2}{3}t^3 + t$ *really solve* (4.11).

In the variation of parameters, we can obtain the general solution by adding in constants of integration. That is, we will add $X(t)\vec{c}$ for a vector of arbitrary constants. But that is precisely the complementary solution.

To conclude this section, we will solve one example using all three methods to be able to compare and contrast them. All of them have their benefits and drawbacks, and it's good to be able to do all three to be able to choose which to apply in a given circumstance.

Example 4.8.6: Find the general solution to the system of differential equations

$$\vec{x}' = \begin{bmatrix} -5 & -2 \\ 4 & 1 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{2t} + 1 \\ e^{2t} + 3 \end{bmatrix}.$$

Solution: No matter which of the three methods we want to use to solve this problem, we always need the eigenvalues and eigenvectors of the coefficient matrix in order to find the general solution to the homogeneous problem. These are found by

$$\det(A - \lambda I) = (-5 - \lambda)(1 - \lambda) - (-2)(4) = \lambda^2 + 4\lambda - 5 + 8 = \lambda^2 + 4\lambda + 3.$$

This polynomial factors as $(\lambda + 1)(\lambda + 3)$ so the eigenvalues are -1 and -3. For $\lambda = -1$, the system we need to solve is

$$(A + I)\vec{v} = \begin{bmatrix} -4 & -2 \\ 4 & 2 \end{bmatrix} \vec{v} = \vec{0}$$

which can be solved by the vector $\vec{v} = \left[\begin{smallmatrix} 1 \\ -2 \end{smallmatrix}\right]$. For $\lambda = -3$, the system is

$$(A + 3I)\vec{v} = \begin{bmatrix} -2 & -2 \\ 4 & 4 \end{bmatrix} \vec{v} = \vec{0}$$

which can be solved by the vector $\vec{v} = \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$. Therefore, the general solution to the homogeneous problem is

$$\vec{x}_c(t) = C_1 \begin{bmatrix} 1 \\ -2 \end{bmatrix} e^{-t} + C_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-3t}. \tag{4.12}$$

Now, we can divide into the different methods that we want to use to solve the non-homogeneous problem.

1. Diagonalization. For this method, we need the matrices E and D defined by

$$E = \begin{bmatrix} 1 & 1 \\ -2 & -1 \end{bmatrix} \qquad D = \begin{bmatrix} -1 & 0 \\ 0 & -3 \end{bmatrix}$$

and can then compute E^{-1} as

$$E^{-1} = \frac{1}{(1)(-1) - (1)(-2)} \begin{bmatrix} -1 & -1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 2 & 1 \end{bmatrix}.$$

We then compute

$$E^{-1}\vec{f} = \begin{bmatrix} -1 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} e^{2t} + 1 \\ e^{2t} + 3 \end{bmatrix} = \begin{bmatrix} -2e^{2t} - 4 \\ 3e^{2t} + 5 \end{bmatrix}$$

which gives rise to the decoupled system

$$\vec{y}' = \begin{bmatrix} -1 & 0 \\ 0 & -3 \end{bmatrix} \vec{y} + \begin{bmatrix} -2e^{2t} - 4 \\ 3e^{2t} + 5 \end{bmatrix}$$

where $\vec{y}$ is defined by $\vec{x} = E\vec{y}$. We can solve for y_1 and y_2 using normal first-order methods:

$$\begin{aligned} y_1' + y_1 &= -2e^{2t} - 4 \\ (e^t y_1)' &= -2e^{3t} - 4e^t \\ e^t y_1 &= -\frac{2}{3}e^{3t} - 4e^t + C_1 \\ y_1 &= -\frac{2}{3}e^{2t} - 4 + C_1 e^{-t} \end{aligned} \qquad \begin{aligned} y_2' + 3y_2 &= 3e^{2t} + 5 \\ (e^{3t} y_2)' &= 3e^{5t} + 5e^{3t} \\ e^{3t} y_2 &= \frac{3}{5}e^{5t} + \frac{5}{3}e^{3t} + C_2 \\ y_2 &= \frac{3}{5}e^{2t} + \frac{5}{3} + C_2 e^{-3t} \end{aligned}$$

Therefore, our solution for $\vec{y}$ is

$$\vec{y}(t) = \begin{bmatrix} -\frac{2}{3}e^{2t} - 4 + C_1 e^{-t} \\ \frac{3}{5}e^{2t} + \frac{5}{3} + C_2 e^{-3t} \end{bmatrix}$$

and by converting back to $\vec{x}$, we get

$$\begin{aligned} \vec{x}(t) = E\vec{y} &= \begin{bmatrix} 1 & 1 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} -\frac{2}{3}e^{2t} - 4 + C_1 e^{-t} \\ \frac{3}{5}e^{2t} + \frac{5}{3} + C_2 e^{-3t} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{15}e^{2t} - \frac{7}{3} + C_1 e^{-t} + C_2 e^{-3t} \\ \frac{11}{15}e^{2t} + \frac{19}{3} - 2C_1 e^{-t} - C_2 e^{-3t} \end{bmatrix}. \end{aligned}$$

Or, rewriting in a different way,

$$\vec{x}(t) = \begin{bmatrix} \frac{1}{15} \\ \frac{11}{15} \end{bmatrix} e^{2t} + \begin{bmatrix} -\frac{7}{3} \\ \frac{19}{3} \end{bmatrix} + C_1 \begin{bmatrix} 1 \\ -2 \end{bmatrix} e^{-t} + C_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-3t}. \tag{4.13}$$

Notice how the general solution to the homogeneous equation (4.12) shows up at the end of this expression.

2. Undetermined coefficients. Since the non-homogeneous part of our equation has terms of the form e^{2t} and constants, we should make a guess of the form

$$\vec{x}_p(t) = \vec{B}e^{2t} + \vec{D}.$$

We can plug this into our equation to get that

$$\vec{x}_p' = 2\vec{B}e^{2t} \tag{4.14}$$

and the right hand side of the equation is

$$\begin{bmatrix} -5 & -2 \\ 4 & 1 \end{bmatrix} \left(\vec{B}e^{2t} + \vec{D} \right) + \begin{bmatrix} e^{2t} + 1 \\ e^{2t} + 3 \end{bmatrix}.$$

Writing out $\vec{B}$ and $\vec{D}$ in components will give the right-hand side as

$$\begin{bmatrix} -5b_1 - 2b_2 \\ 4b_1 + b_2 \end{bmatrix} e^{2t} + \begin{bmatrix} -5d_1 - 2d_2 \\ 4d_1 + d_2 \end{bmatrix} + \begin{bmatrix} e^{2t} + 1 \\ e^{2t} + 3 \end{bmatrix}$$
$$= \begin{bmatrix} -5b_1 - 2b_2 + 1 \\ 4b_1 + b_2 + 1 \end{bmatrix} e^{2t} + \begin{bmatrix} -5d_1 - 2d_2 + 1 \\ 4d_1 + d_2 + 3 \end{bmatrix}.$$

We can now set this equal to the left-hand side in (4.14) to get the vector equation

$$\begin{bmatrix} 2b_1 \\ 2b_2 \end{bmatrix} e^{2t} = \begin{bmatrix} -5b_1 - 2b_2 + 1 \\ 4b_1 + b_2 + 1 \end{bmatrix} e^{2t} + \begin{bmatrix} -5d_1 - 2d_2 + 1 \\ 4d_1 + d_2 + 3 \end{bmatrix}$$

and we can match up the terms on the left and right sides to get a system that we need to solve:

$$\begin{aligned} 2b_1 &= -5b_1 - 2b_2 + 1 \\ 2b_2 &= 4b_1 + b_2 + 1 \\ 0 &= -5d_1 - 2d_2 + 1 \\ 0 &= 4d_1 + d_2 + 3. \end{aligned}$$

Let's start with the b equations. Rearranging these gives

$$7b_1 + 2b_2 = 1 \qquad -4b_1 + b_2 = 1$$

Subtracting two copies of the second equation from the first gives $15b_1 = -1$ or $b_1 = -1/15$, which gives $b_2 = 1 + \frac{4}{15} = \frac{19}{13}$. Next, we can solve the d equations, which we can rearrange to give

$$5d_1 + 2d_2 = 1 \qquad 4d_1 + d_2 = -3$$

Subtracting two copies of the second equation from the first gives $-3d_1 = 7$ so $d_1 = -7/3$, leading to $d_2 = -3 - 4(-7/3) = 19/3$. Therefore, a solution to the non-homogeneous problem is

$$\vec{x}_p(t) = \begin{bmatrix} -\frac{1}{15} \\ \frac{19}{15} \end{bmatrix} e^{2t} + \begin{bmatrix} -\frac{7}{3} \\ \frac{19}{3} \end{bmatrix}$$

and so we can add in the homogeneous solution from (4.12) to get the full general solution as

$$\begin{bmatrix} -\frac{1}{13} \\ \frac{17}{13} \end{bmatrix} e^{2t} + \begin{bmatrix} -\frac{7}{3} \\ \frac{19}{3} \end{bmatrix} + C_1 \begin{bmatrix} 1 \\ -2 \end{bmatrix} e^{-t} + C_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-3t}. \tag{4.15}$$

3. Variation of Parameters. For this method, we write down the fundamental matrix $X(t)$ by combining the two basis solutions into a matrix, as

$$X(t) = \begin{bmatrix} e^{-t} & e^{-3t} \\ -2e^{-t} & -e^{-3t} \end{bmatrix}$$

and compute the inverse matrix as

$$X^{-1}(t) = \frac{1}{(e^{-t})(-e^{-3t}) - (e^{-3t})(-2e^{-t})} \begin{bmatrix} -e^{-3t} & -e^{-3t} \\ 2e^{-t} & e^{-t} \end{bmatrix} = \begin{bmatrix} -e^{t} & -e^{t} \\ 2e^{3t} & e^{3t} \end{bmatrix}.$$

We can then work out the components of the method of variation of parameters.

$$\begin{aligned} X(t)^{-1}\vec{f} &= \begin{bmatrix} -e^t & -e^t \\ 2e^{3t} & e^{3t} \end{bmatrix} \begin{bmatrix} e^{2t}+1 \\ e^{2t}+3 \end{bmatrix} \\ &= \begin{bmatrix} -e^{3t}-e^t-e^{3t}-3e^t \\ 2e^{5t}+2e^{3t}+e^{5t}+3e^{3t} \end{bmatrix} \\ &= \begin{bmatrix} -2e^{3t}-4e^t \\ 3e^{5t}+5e^{3t} \end{bmatrix}. \end{aligned}$$

Integrating this expression gives

$$\int X(t)^{-1}\vec{f}\,dt = \begin{bmatrix} -\frac{2}{3}e^{3t}-4e^t+C_1 \\ \frac{3}{5}e^{5t}+\frac{5}{3}e^{3t}+C_2 \end{bmatrix},$$

and so the general solution to this system is

$$\begin{aligned} X(t)\int X(t)^{-1}\vec{f}\,dt &= \begin{bmatrix} e^{-t} & e^{-3t} \\ -2e^{-t} & -e^{-3t} \end{bmatrix} \begin{bmatrix} -\frac{2}{3}e^{3t}-4e^t+C_1 \\ \frac{3}{5}e^{5t}+\frac{5}{3}e^{3t}+C_2 \end{bmatrix} \\ &= \begin{bmatrix} -\frac{2}{3}e^{2t}-4+C_1e^{-t}+\frac{3}{5}e^{2t}+\frac{5}{3}+C_2e^{-3t} \\ \frac{4}{3}e^{2t}+8-2C_1e^{-t}-\frac{3}{5}e^{2t}-\frac{5}{3}-C_2e^{-3t} \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{15}e^{2t}-\frac{7}{3}+C_1e^{-t}+C_2e^{-3t} \\ \frac{11}{15}e^{2t}+\frac{19}{3}-2C_1e^{-t}-C_2e^{-3t} \end{bmatrix}. \end{aligned} \tag{4.16}$$

Notice again that the homogeneous solution (4.12) shows up at the end of these terms, so we do not need to add it in at the end.

Comparing the solutions (4.13), (4.15), and (4.16), we see that the three solutions generated by these three methods are all the same. ⌟

For this previous example, we only found the general solution. If the solution to an initial value problem was needed, we would need to wait until the very end, once we have figured out the solution to the non-homogeneous problem and added in the solution to the homogeneous problem to determine the value of the constants to meet the initial condition.

4.8.3 Exercises

***Exercise* 4.8.5:** *Find a particular solution to $x' = x + 2y + 2t$, $y' = 3x + 2y - 4$,*

a) using diagonalization, *b) using undetermined coefficients.*

Exercise* 4.8.6: *Find a particular solution to $x' = 5x + 4y + t$, $y' = x + 8y - t$,*

a) using diagonalization, *b) using undetermined coefficients.*

***Exercise* 4.8.7:** *Find the general solution to $x' = 4x + y - 1$, $y' = x + 4y - e^t$,*

a) using diagonalization, *b) using undetermined coefficients.*

Exercise* 4.8.8: *Find a particular solution to $x' = y + e^t$, $y' = x + e^t$,*

a) using diagonalization, *b) using undetermined coefficients.*

***Exercise* 4.8.9:** *Let $A = \begin{bmatrix} 2 & 2 \\ -2 & 7 \end{bmatrix}$. This matrix has eigenvectors $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$.*

a) Find a fundamental matrix, $\Psi(t)$, for the system $\vec{x}' = A\vec{x}$.

b) Use variation of parameters to solve the non-homogeneous system $\vec{x}' = A\vec{x} + \begin{bmatrix} e^{6t} \\ 0 \end{bmatrix}$.

c) If we used method of undetermined coefficients instead, what would be the appropriate guess for the form of the non-homogeneous solution?

Exercise* 4.8.10: *Solve $x_1' = x_2 + t$, $x_2' = x_1 + t$ with initial conditions $x_1(0) = 1$, $x_2(0) = 2$, using diagonalization.*

***Exercise* 4.8.11:** *For each of the following vector functions $\vec{f}(t)$, find the general solution to the system of differential equations given by*

$$\vec{x}' = \begin{bmatrix} -1 & -4 \\ 2 & 5 \end{bmatrix} \vec{x} + \vec{f}(t)$$

using any of the methods described in this section. Notice the similarities and differences between using these methods for different non-homogeneous parts.

a) $\vec{f}(t) = \begin{bmatrix} e^{2t} \\ 1 \end{bmatrix}$ *b)* $\vec{f}(t) = \begin{bmatrix} e^{-t} + 2 \\ e^{4t} - 1 \end{bmatrix}$ *c)* $\vec{f}(t) = \begin{bmatrix} e^{3t} \\ t \end{bmatrix}$

d) $\vec{f}(t) = \begin{bmatrix} \sin(3t) \\ 1 - \sin(3t) \end{bmatrix}$ *e)* $\vec{f}(t) = \begin{bmatrix} t + 2 \\ e^{-2t} \end{bmatrix}$ *f)* $\vec{f}(t) = \begin{bmatrix} te^t \\ 3 \end{bmatrix}$

***Exercise* 4.8.12:** *The variation of parameters method can also be applied to constant coefficient systems. Find the general solution of the system*

$$\vec{x}' = \begin{bmatrix} 3 & -1 \\ 2 & -2 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{2t} \\ e^t \end{bmatrix}$$

using

a) diagonalization *b) variation of parameters.*

Compare and contrast these methods. You can use undetermined coefficients to check your answer.

***Exercise* 4.8.13:** *Find the general solution to the differential equation*

$$\vec{x}' = \begin{bmatrix} -3 & -1 \\ 4 & -3 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{3t} + 1 \\ 2 \end{bmatrix}.$$

The best option is undetermined coefficients here because of the eigenvalues of the matrix. Diagonalization can be used, but care will be needed with solving the decoupled system because the coefficients will be complex.

***Exercise* 4.8.14:** *Find the general solution to the differential equation*

$$\vec{x}' = \begin{bmatrix} -5 & 16 \\ -1 & 3 \end{bmatrix} \vec{x} + \begin{bmatrix} \cos(2t) \\ \sin(2t) - 2\cos(2t) \end{bmatrix}.$$

The best option is undetermined coefficients here because of the eigenvalues of the matrix. We can't actually use diagonalization (try it and see why!).

***Exercise* 4.8.15:** *Find the general solution to the differential equation*

$$\vec{x}' = \begin{bmatrix} -2 & -12 \\ 2 & 8 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{2t} + e^{3t} \\ -2e^{2t} \end{bmatrix}.$$

***Exercise* 4.8.16:** *Consider the system*

$$\frac{dx}{dt} = x + 2y + 4e^{3t}; \ \frac{dy}{dt} = 3x - e^{3t}. \tag{4.17}$$

a) *Rewrite (4.17) in the form $\vec{x}' = A\vec{x} + \vec{g}(t)$, where $\vec{x}' = A\vec{x}$ is a homogeneous system, and $\vec{g}(t)$ is a vector-valued function.*

b) *Solve (4.17) using Method of Undetermined Coefficients.*

***Exercise* 4.8.17:**

a) *Use variation of parameters to solve the system* $\vec{x}' = \begin{bmatrix} 1 & -4 \\ 4 & -7 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{-3t} \\ 0 \end{bmatrix}$.

b) *What does that solution tell you about how to set up the guess for the method of undetermined coefficients when there is a repeated eigenvalue?*

***Exercise* 4.8.18:** *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} 5 & -6 \\ 3 & -1 \end{bmatrix} \vec{x} + \begin{bmatrix} t \\ 3 \end{bmatrix} \qquad \vec{x}(0) = \begin{bmatrix} 1 \\ -3 \end{bmatrix}.$$

***Exercise* 4.8.19:** *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} -4 & 2 \\ -9 & 5 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{3t} \\ e^{t} - 1 \end{bmatrix} \qquad \vec{x}(0) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

***Exercise* 4.8.20:** *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} 3 & 2 \\ 0 & 4 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{4t} \\ e^{3t} - t \end{bmatrix} \qquad \vec{x}(0) = \begin{bmatrix} 2 \\ -1 \end{bmatrix}.$$

***Exercise* 4.8.21:** *Take the equation* $\vec{x}' = \begin{bmatrix} \frac{1}{t} & -1 \\ 1 & \frac{1}{t} \end{bmatrix} \vec{x} + \begin{bmatrix} t^2 \\ -t \end{bmatrix}$.

a) *Check that* $\vec{x}_c = c_1 \begin{bmatrix} t\sin t \\ -t\cos t \end{bmatrix} + c_2 \begin{bmatrix} t\cos t \\ t\sin t \end{bmatrix}$ *is the complementary solution.*

b) *Use variation of parameters to find a particular solution.*

4.9 Second order systems and applications

Attribution: [JL], §3.6.

Learning Objectives

After this section, you will be able to:

- Use second order systems to model physical problems and
- Solve second order systems using diagonalization or eigenvalue methods.

4.9.1 Undamped mass-spring systems

While we did say that we will usually only look at first order systems, it is sometimes more convenient to study the system in the way it arises naturally. For example, suppose we have 3 masses connected by springs between two walls. We could pick any higher number, and the math would be essentially the same, but for simplicity we pick 3 right now. Let us also assume no friction, that is, the system is undamped. The masses are m_1, m_2, and m_3 and the spring constants are k_1, k_2, k_3, and k_4. Let x_1 be the displacement from rest position of the first mass, and x_2 and x_3 the displacement of the second and third mass. We make, as usual, positive values go right (as x_1 grows, the first mass is moving right). See Figure 4.16.

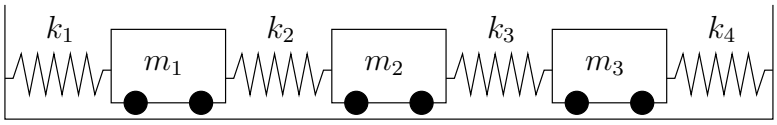


Figure 4.16: System of masses and springs.

This simple system turns up in unexpected places. For example, our world really consists of many small particles of matter interacting together. When we try the system above with many more masses, we obtain a good approximation to how an elastic material behaves.

Let us set up the equations for the three mass system. By Hooke's law, the force acting on the mass equals the spring compression times the spring constant. By Newton's second law, force is mass times acceleration. So if we sum the forces acting on each mass, put the right sign in front of each term, depending on the direction in which it is acting, and set this equal to mass times the acceleration, we end up with the desired system of equations.

$$\begin{aligned} m_1 x_1'' &= -k_1 x_1 + k_2(x_2 - x_1) &&= -(k_1 + k_2)x_1 + k_2 x_2, \\ m_2 x_2'' &= -k_2(x_2 - x_1) + k_3(x_3 - x_2) &&= k_2 x_1 - (k_2 + k_3)x_2 + k_3 x_3, \\ m_3 x_3'' &= -k_3(x_3 - x_2) - k_4 x_3 &&= k_3 x_2 - (k_3 + k_4)x_3. \end{aligned}$$

We define the matrices

$$M = \begin{bmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{bmatrix} \quad \text{and} \quad K = \begin{bmatrix} -(k_1 + k_2) & k_2 & 0 \\ k_2 & -(k_2 + k_3) & k_3 \\ 0 & k_3 & -(k_3 + k_4) \end{bmatrix}.$$

We write the equation simply as

$$M\vec{x}'' = K\vec{x}.$$

At this point we could introduce 3 new variables and write out a system of 6 first order equations. We claim this simple setup is easier to handle as a second order system. We call $\vec{x}$ the *displacement vector*, M the *mass matrix*, and K the *stiffness matrix*.

***Exercise* 4.9.1:** *Repeat this setup for 4 masses (find the matrices M and K). Do it for 5 masses. Can you find a prescription to do it for n masses?*

As with a single equation we want to "divide by M." This means computing the inverse of M. The masses are all nonzero and M is a diagonal matrix, so computing the inverse is easy:

$$M^{-1} = \begin{bmatrix} \frac{1}{m_1} & 0 & 0 \\ 0 & \frac{1}{m_2} & 0 \\ 0 & 0 & \frac{1}{m_3} \end{bmatrix}.$$

This fact follows readily by how we multiply diagonal matrices. As an exercise, you should verify that $MM^{-1} = M^{-1}M = I$.

Let $A = M^{-1}K$. We look at the system $\vec{x}'' = M^{-1}K\vec{x}$, or

$$\vec{x}'' = A\vec{x}.$$

Many real world systems can be modeled by this equation. For simplicity, we will only talk about the given masses-and-springs problem. We try a solution of the form

$$\vec{x} = \vec{v}e^{\alpha t}.$$

We compute that for this guess, $\vec{x}'' = \alpha^2\vec{v}e^{\alpha t}$. We plug our guess into the equation and get

$$\alpha^2\vec{v}e^{\alpha t} = A\vec{v}e^{\alpha t}.$$

We divide by $e^{\alpha t}$ to arrive at $\alpha^2\vec{v} = A\vec{v}$. Hence if α^2 is an eigenvalue of A and $\vec{v}$ is a corresponding eigenvector, we have found a solution.

In our example, and in other common applications, A has only real negative eigenvalues (and possibly a zero eigenvalue). So we study only this case. When an eigenvalue λ is negative, it means that $\alpha^2 = \lambda$ is negative. Hence there is some real number ω such that $-\omega^2 = \lambda$. Then $\alpha = \pm i\omega$. The solution we guessed was

$$\vec{x} = \vec{v}\,\big(\cos(\omega t) + i\sin(\omega t)\big).$$

By taking the real and imaginary parts (note that $\vec{v}$ is real), we find that $\vec{v}\cos(\omega t)$ and $\vec{v}\sin(\omega t)$ are linearly independent solutions.

If an eigenvalue is zero, it turns out that both $\vec{v}$ and $\vec{v}t$ are solutions, where $\vec{v}$ is an eigenvector corresponding to the eigenvalue 0.

***Exercise* 4.9.2:** *Show that if A has a zero eigenvalue and $\vec{v}$ is a corresponding eigenvector, then $\vec{x} = \vec{v}(a + bt)$ is a solution of $\vec{x}'' = A\vec{x}$ for arbitrary constants a and b.*

Theorem 4.9.1

Let A be a real $n \times n$ matrix with n distinct real negative (or zero) eigenvalues we denote by $-\omega_1^2 > -\omega_2^2 > \cdots > -\omega_n^2$, and corresponding eigenvectors by $\vec{v}_1$, $\vec{v}_2$, $\ldots$, $\vec{v}_n$. If A is invertible (that is, if $\omega_1 > 0$), then

$$\vec{x}(t) = \sum_{i=1}^{n} \vec{v}_i \big(a_i \cos(\omega_i t) + b_i \sin(\omega_i t)\big),$$

is the general solution of

$$\vec{x}'' = A\vec{x},$$

for some arbitrary constants a_i and b_i. If A has a zero eigenvalue, that is $\omega_1 = 0$, and all other eigenvalues are distinct and negative, then the general solution can be written as

$$\vec{x}(t) = \vec{v}_1(a_1 + b_1 t) + \sum_{i=2}^{n} \vec{v}_i \big(a_i \cos(\omega_i t) + b_i \sin(\omega_i t)\big).$$

We use this solution and the setup from the introduction of this section even when some of the masses and springs are missing. For example, when there are only 2 masses and only 2 springs, simply take only the equations for the two masses and set all the spring constants for the springs that are missing to zero.

4.9.2 Examples

Example 4.9.1: Consider the setup in Figure 4.17, with $m_1 = 2\,\text{kg}$, $m_2 = 1\,\text{kg}$, $k_1 = 4\,\text{N/m}$, and $k_2 = 2\,\text{N/m}$.

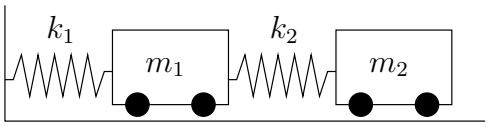


Figure 4.17: System of masses and springs.

Solution: The equations we write down are

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \vec{x}'' = \begin{bmatrix} -(4+2) & 2 \\ 2 & -2 \end{bmatrix} \vec{x},$$

or

$$\vec{x}'' = \begin{bmatrix} -3 & 1 \\ 2 & -2 \end{bmatrix} \vec{x}.$$

We find the eigenvalues of A to be $\lambda = -1, -4$ (exercise). We find corresponding eigenvectors to be $\left[\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$ respectively (exercise).

We check the theorem and note that $\omega_1 = 1$ and $\omega_2 = 2$. Hence the general solution is

$$\vec{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \big(a_1 \cos(t) + b_1 \sin(t)\big) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} \big(a_2 \cos(2t) + b_2 \sin(2t)\big).$$

The two terms in the solution represent the two so-called *natural* or *normal modes of oscillation*. And the two (angular) frequencies are the *natural frequencies*. The first natural frequency is 1, and second natural frequency is 2. The two modes are plotted in Figure 4.18.

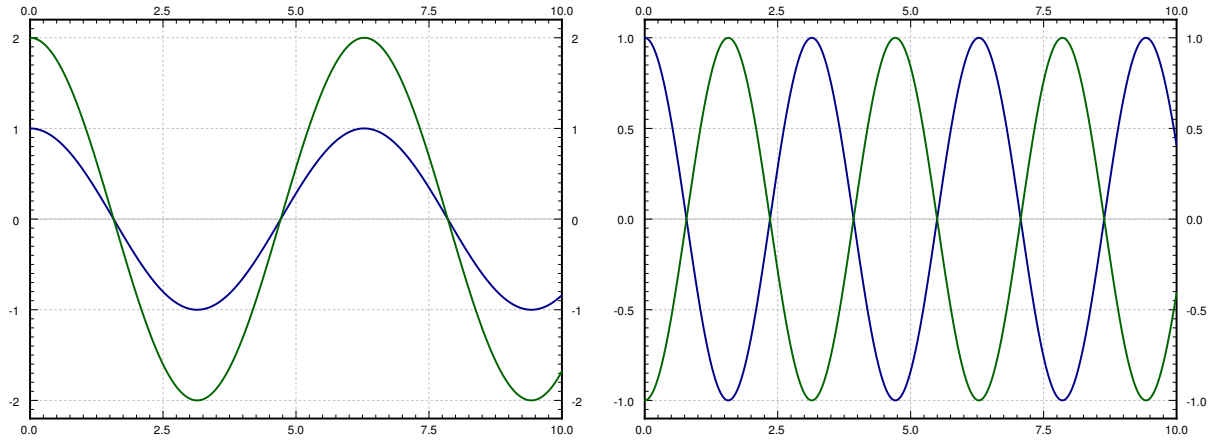


Figure 4.18: The two modes of the mass-spring system. In the left plot the masses are moving in unison and in the right plot are masses moving in the opposite direction.

Let us write the solution as

$$\vec{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} c_1 \cos(t - \alpha_1) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} c_2 \cos(2t - \alpha_2).$$

The first term,

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} c_1 \cos(t - \alpha_1) = \begin{bmatrix} c_1 \cos(t - \alpha_1) \\ 2c_1 \cos(t - \alpha_1) \end{bmatrix},$$

corresponds to the mode where the masses move synchronously in the same direction.

The second term,

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix} c_2 \cos(2t - \alpha_2) = \begin{bmatrix} c_2 \cos(2t - \alpha_2) \\ -c_2 \cos(2t - \alpha_2) \end{bmatrix},$$

corresponds to the mode where the masses move synchronously but in opposite directions.

The general solution is a combination of the two modes. That is, the initial conditions determine the amplitude and phase shift of each mode. As an example, suppose we have initial conditions

$$\vec{x}(0) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \vec{x}'(0) = \begin{bmatrix} 0 \\ 6 \end{bmatrix}.$$

We use the a_j, b_j constants to solve for initial conditions. First

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \vec{x}(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} a_1 + \begin{bmatrix} 1 \\ -1 \end{bmatrix} a_2 = \begin{bmatrix} a_1 + a_2 \\ 2a_1 - a_2 \end{bmatrix}.$$

We solve (exercise) to find $a_1 = 0$, $a_2 = 1$. To find the b_1 and b_2, we differentiate first:

$$\vec{x}' = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \bigl(-a_1 \sin(t) + b_1 \cos(t)\bigr) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} \bigl(-2a_2 \sin(2t) + 2b_2 \cos(2t)\bigr).$$

Now we solve:

$$\begin{bmatrix} 0 \\ 6 \end{bmatrix} = \vec{x}'(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} b_1 + \begin{bmatrix} 1 \\ -1 \end{bmatrix} 2b_2 = \begin{bmatrix} b_1 + 2b_2 \\ 2b_1 - 2b_2 \end{bmatrix}.$$

Again solve (exercise) to find $b_1 = 2$, $b_2 = -1$. So our solution is

$$\vec{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} 2\sin(t) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} \bigl(\cos(2t) - \sin(2t)\bigr) = \begin{bmatrix} 2\sin(t) + \cos(2t) - \sin(2t) \\ 4\sin(t) - \cos(2t) + \sin(2t) \end{bmatrix}.$$

The graphs of the two displacements, x_1 and x_2 of the two carts is in Figure 4.19.

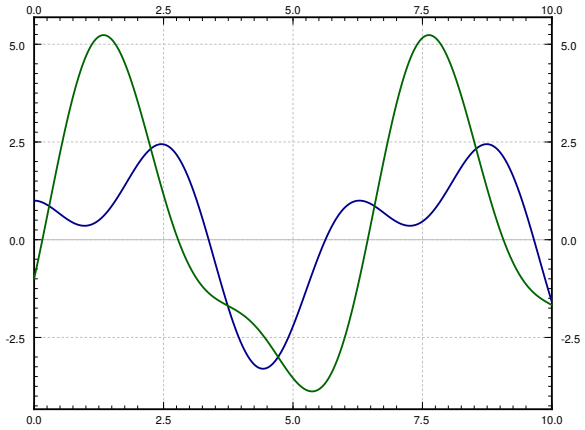


Figure 4.19: Superposition of the two modes given the initial conditions.

Example 4.9.2: We have two toy rail cars. Car 1 of mass 2 kg is traveling at 3 m/s towards the second rail car of mass 1 kg. There is a bumper on the second rail car that engages at the moment the cars hit (it connects to two cars) and does not let go. The bumper acts like a spring of spring constant $k = 2$ N/m. The second car is 10 meters from a wall. See Figure 4.20 on the next page.

We want to ask several questions. At what time after the cars link does impact with the wall happen? What is the speed of car 2 when it hits the wall?

Solution: OK, let us first set the system up. Let $t = 0$ be the time when the two cars link up. Let x_1 be the displacement of the first car from the position at $t = 0$, and let x_2 be the displacement of the second car from its original location. Then the time when $x_2(t) = 10$ is

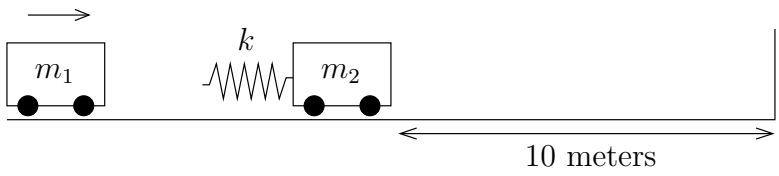


Figure 4.20: The crash of two rail cars.

exactly the time when impact with wall occurs. For this t, $x_2'(t)$ is the speed at impact. This system acts just like the system of the previous example but without k_1. Hence the equation is

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \vec{x}'' = \begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \vec{x},$$

or

$$\vec{x}'' = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix} \vec{x}.$$

We compute the eigenvalues of A. It is not hard to see that the eigenvalues are 0 and -3 (exercise). Furthermore, eigenvectors are $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ -2 \end{smallmatrix}\right]$ respectively (exercise). Then $\omega_1 = 0$, $\omega_2 = \sqrt{3}$, and by the second part of the theorem the general solution is

$$\begin{aligned} \vec{x} &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} (a_1 + b_1 t) + \begin{bmatrix} 1 \\ -2 \end{bmatrix} \Big(a_2 \cos(\sqrt{3}\, t) + b_2 \sin(\sqrt{3}\, t) \Big) \\ &= \begin{bmatrix} a_1 + b_1 t + a_2 \cos(\sqrt{3}\, t) + b_2 \sin(\sqrt{3}\, t) \\ a_1 + b_1 t - 2a_2 \cos(\sqrt{3}\, t) - 2b_2 \sin(\sqrt{3}\, t) \end{bmatrix}. \end{aligned}$$

We now apply the initial conditions. First the cars start at position 0 so $x_1(0) = 0$ and $x_2(0) = 0$. The first car is traveling at $3\,{}^{\mathrm{m}}\!/_{\mathrm{s}}$, so $x_1'(0) = 3$ and the second car starts at rest, so $x_2'(0) = 0$. The first conditions says

$$\vec{0} = \vec{x}(0) = \begin{bmatrix} a_1 + a_2 \\ a_1 - 2a_2 \end{bmatrix}.$$

It is not hard to see that $a_1 = a_2 = 0$. We set $a_1 = 0$ and $a_2 = 0$ in $\vec{x}(t)$ and differentiate to get

$$\vec{x}'(t) = \begin{bmatrix} b_1 + \sqrt{3}\, b_2 \cos(\sqrt{3}\, t) \\ b_1 - 2\sqrt{3}\, b_2 \cos(\sqrt{3}\, t) \end{bmatrix}.$$

So

$$\begin{bmatrix} 3 \\ 0 \end{bmatrix} = \vec{x}'(0) = \begin{bmatrix} b_1 + \sqrt{3}\, b_2 \\ b_1 - 2\sqrt{3}\, b_2 \end{bmatrix}.$$

Solving these two equations we find $b_1 = 2$ and $b_2 = \frac{1}{\sqrt{3}}$. Hence the position of our cars is (until the impact with the wall)

$$\vec{x} = \begin{bmatrix} 2t + \frac{1}{\sqrt{3}} \sin(\sqrt{3}\, t) \\ 2t - \frac{2}{\sqrt{3}} \sin(\sqrt{3}\, t) \end{bmatrix}.$$

Note how the presence of the zero eigenvalue resulted in a term containing t. This means that the cars will be traveling in the positive direction as time grows, which is what we expect.

What we are really interested in is the second expression, the one for x_2. We have $x_2(t) = 2t - \frac{2}{\sqrt{3}}\sin(\sqrt{3}\,t)$. See Figure 4.21 for the plot of x_2 versus time.

Just from the graph we can see that time of impact will be a little more than 5 seconds from time zero. For this we have to solve the equation $10 = x_2(t) = 2t - \frac{2}{\sqrt{3}}\sin(\sqrt{3}\,t)$. Using a computer (or even a graphing calculator) we find that $t_{\text{impact}} \approx 5.22$ seconds.

The speed of the second car is $x_2' = 2 - 2\cos(\sqrt{3}\,t)$. At the time of impact (5.22 seconds from $t = 0$) we get $x_2'(t_{\text{impact}}) \approx 3.85$. The maximum speed is the maximum of $2 - 2\cos(\sqrt{3}\,t)$, which is 4. We are traveling at almost the maximum speed when we hit the wall.

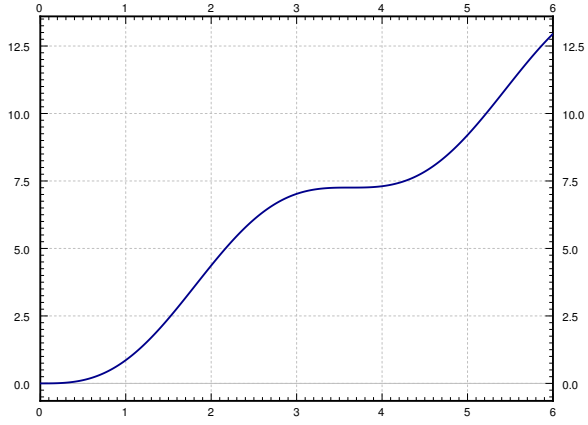


Figure 4.21: Position of the second car in time (ignoring the wall).

Suppose that Bob is a tiny person sitting on car 2. Bob has a Martini in his hand and would like not to spill it. Let us suppose Bob would not spill his Martini when the first car links up with car 2, but if car 2 hits the wall at any speed greater than zero, Bob will spill his drink. Suppose Bob can move car 2 a few meters towards or away from the wall (he cannot go all the way to the wall, nor can he get out of the way of the first car). Is there a "safe" distance for him to be at? A distance such that the impact with the wall is at zero speed?

The answer is yes. Looking at Figure 4.21, we note the "plateau" between $t = 3$ and $t = 4$. There is a point where the speed is zero. To find it we solve $x_2'(t) = 0$. This is when $\cos(\sqrt{3}\,t) = 1$ or in other words when $t = \frac{2\pi}{\sqrt{3}}, \frac{4\pi}{\sqrt{3}}, \ldots$ and so on. We plug in the first value to obtain $x_2\left(\frac{2\pi}{\sqrt{3}}\right) = \frac{4\pi}{\sqrt{3}} \approx 7.26$. So a "safe" distance is about 7 and a quarter meters from the wall.

Alternatively Bob could move away from the wall towards the incoming car 2, where another safe distance is $x_2\left(\frac{4\pi}{\sqrt{3}}\right) = \frac{8\pi}{\sqrt{3}} \approx 14.51$ and so on. We can use all the different t such that $x_2'(t) = 0$. Of course $t = 0$ is also a solution, corresponding to $x_2 = 0$, but that means standing right at the wall. ⌟

4.9.3 Forced oscillations

Finally we move to forced oscillations. Suppose that now our system is

$$\vec{x}'' = A\vec{x} + \vec{F}\cos(\omega t). \tag{4.18}$$

That is, we are adding periodic forcing to the system in the direction of the vector $\vec{F}$.

As before, this system just requires us to find one particular solution $\vec{x}_p$, add it to the general solution of the associated homogeneous system $\vec{x}_c$, and we will have the general

solution to (4.18). Let us suppose that ω is not one of the natural frequencies of $\vec{x}'' = A\vec{x}$, then we can guess

$$\vec{x}_p = \vec{c}\cos(\omega t),$$

where $\vec{c}$ is an unknown constant vector. Note that we do not need to use sine since there are only second derivatives. We solve for $\vec{c}$ to find $\vec{x}_p$. This is really just the method of *undetermined coefficients* for systems. Let us differentiate $\vec{x}_p$ twice to get

$$\vec{x}_p'' = -\omega^2\vec{c}\cos(\omega t).$$

Plug $\vec{x}_p$ and $\vec{x}_p''$ into equation (4.18):

$$\overbrace{-\omega^2\vec{c}\cos(\omega t)}^{\vec{x}_p''} = \overbrace{A\vec{c}\cos(\omega t)}^{A\vec{x}_p} + \vec{F}\cos(\omega t).$$

We cancel out the cosine and rearrange the equation to obtain

$$(A + \omega^2 I)\vec{c} = -\vec{F}.$$

So

$$\vec{c} = (A + \omega^2 I)^{-1}(-\vec{F}).$$

Of course this is possible only if $(A + \omega^2 I) = \big(A - (-\omega^2)I\big)$ is invertible. That matrix is invertible if and only if $-\omega^2$ is not an eigenvalue of A. That is true if and only if ω is not a natural frequency of the system.

We simplified things a little bit. If we wish to have the forcing term to be in the units of force, say Newtons, then we must write

$$M\vec{x}'' = K\vec{x} + \vec{G}\cos(\omega t).$$

If we then write things in terms of $A = M^{-1}K$, we have

$$\vec{x}'' = M^{-1}K\vec{x} + M^{-1}\vec{G}\cos(\omega t) \qquad \text{or} \qquad \vec{x}'' = A\vec{x} + \vec{F}\cos(\omega t),$$

where $\vec{F} = M^{-1}\vec{G}$.

Example 4.9.3: Let us take the example in Figure 4.17 on page 348 with the same parameters as before: $m_1 = 2$, $m_2 = 1$, $k_1 = 4$, and $k_2 = 2$. Now suppose that there is a force $2\cos(3t)$ acting on the second cart.

Solution: The equation is

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}\vec{x}'' = \begin{bmatrix} -4 & 2 \\ 2 & -2 \end{bmatrix}\vec{x} + \begin{bmatrix} 0 \\ 2 \end{bmatrix}\cos(3t) \qquad \text{or} \qquad \vec{x}'' = \begin{bmatrix} -3 & 1 \\ 2 & -2 \end{bmatrix}\vec{x} + \begin{bmatrix} 0 \\ 2 \end{bmatrix}\cos(3t).$$

We solved the associated homogeneous equation before and found the complementary solution to be

$$\vec{x}_c = \begin{bmatrix} 1 \\ 2 \end{bmatrix}\big(a_1\cos(t) + b_1\sin(t)\big) + \begin{bmatrix} 1 \\ -1 \end{bmatrix}\big(a_2\cos(2t) + b_2\sin(2t)\big).$$

The natural frequencies are 1 and 2. As 3 is not a natural frequency, we try $\vec{c}\cos(3t)$. We invert $(A + 3^2 I)$:

$$\left(\begin{bmatrix} -3 & 1 \\ 2 & -2 \end{bmatrix} + 3^2 I \right)^{-1} = \begin{bmatrix} 6 & 1 \\ 2 & 7 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{7}{40} & \frac{-1}{40} \\ \frac{-1}{20} & \frac{3}{20} \end{bmatrix}.$$

Hence,

$$\vec{c} = (A + \omega^2 I)^{-1}(-\vec{F}) = \begin{bmatrix} \frac{7}{40} & \frac{-1}{40} \\ \frac{-1}{20} & \frac{3}{20} \end{bmatrix} \begin{bmatrix} 0 \\ -2 \end{bmatrix} = \begin{bmatrix} \frac{1}{20} \\ \frac{-3}{10} \end{bmatrix}.$$

Combining with the general solution of the associated homogeneous problem, we get that the general solution to $\vec{x}'' = A\vec{x} + \vec{F}\cos(\omega t)$ is

$$\vec{x} = \vec{x}_c + \vec{x}_p = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \big(a_1 \cos(t) + b_1 \sin(t) \big) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} \big(a_2 \cos(2t) + b_2 \sin(2t) \big) + \begin{bmatrix} \frac{1}{20} \\ \frac{-3}{10} \end{bmatrix} \cos(3t).$$

We then solve for the constants a_1, a_2, b_1, and b_2 using any initial conditions we are given. ⌋

Note that given force $\vec{f}$, we write the equation as $M\vec{x}'' = K\vec{x} + \vec{f}$ to get the units right. Then we write $\vec{x}'' = M^{-1}K\vec{x} + M^{-1}\vec{f}$. The term $\vec{g} = M^{-1}\vec{f}$ in $\vec{x}'' = A\vec{x} + \vec{g}$ is in units of force per unit mass.

If ω is a natural frequency of the system, *resonance* may occur, because we will have to try a particular solution of the form

$$\vec{x}_p = \vec{c}\, t \sin(\omega t) + \vec{d}\cos(\omega t).$$

That is assuming that the eigenvalues of the coefficient matrix are distinct. Next, note that the amplitude of this solution grows without bound as t grows.

4.9.4 Non-Homogeneous Solutions

Undetermined coefficients

Let the equation be

$$\vec{x}'' = A\vec{x} + \vec{F}(t),$$

where A is a constant matrix. If $\vec{F}(t)$ is of the form $\vec{F}_0 \cos(\omega t)$, then as two derivatives of cosine is again cosine we can try a solution of the form

$$\vec{x}_p = \vec{c}\cos(\omega t),$$

and we do not need to introduce sines.

If the $\vec{F}$ is a sum of cosines, note that we still have the superposition principle. If $\vec{F}(t) = \vec{F}_0 \cos(\omega_0 t) + \vec{F}_1 \cos(\omega_1 t)$, then we would try $\vec{a}\cos(\omega_0 t)$ for the problem $\vec{x}'' = A\vec{x} + \vec{F}_0 \cos(\omega_0 t)$, and we would try $\vec{b}\cos(\omega_1 t)$ for the problem $\vec{x}'' = A\vec{x} + \vec{F}_1 \cos(\omega_1 t)$. Then we sum the solutions.

However, if there is duplication with the complementary solution, or the equation is of the form $\vec{x}'' = A\vec{x}' + B\vec{x} + \vec{F}(t)$, then we need to do the same thing as we do for first order systems.

You will never go wrong with putting in more terms than needed into your guess. You will find that the extra coefficients will turn out to be zero. But it is useful to save some time and effort.

Eigenvector decomposition

If we have the system

$$\vec{x}'' = A\vec{x} + \vec{f}(t),$$

we can do *eigenvector decomposition*, just like for first order systems.

Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues and $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n$ be eigenvectors. Again form the matrix $E = \begin{bmatrix} \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \end{bmatrix}$. Write

$$\vec{x}(t) = \vec{v}_1\, \xi_1(t) + \vec{v}_2\, \xi_2(t) + \cdots + \vec{v}_n\, \xi_n(t).$$

Decompose $\vec{f}$ in terms of the eigenvectors

$$\vec{f}(t) = \vec{v}_1\, g_1(t) + \vec{v}_2\, g_2(t) + \cdots + \vec{v}_n\, g_n(t),$$

where, again, $\vec{g} = E^{-1}\vec{f}$.

We plug in, and as before we obtain

$$\begin{aligned} \overbrace{\vec{v}_1\xi_1'' + \vec{v}_2\xi_2'' + \cdots + \vec{v}_n\xi_n''}^{\vec{x}''} &= \overbrace{A\left(\vec{v}_1\xi_1 + \vec{v}_2\xi_2 + \cdots + \vec{v}_n\xi_n\right)}^{A\vec{x}} + \overbrace{\vec{v}_1 g_1 + \vec{v}_2 g_2 + \cdots + \vec{v}_n g_n}^{\vec{f}} \\ &= A\vec{v}_1\xi_1 + A\vec{v}_2\xi_2 + \cdots + A\vec{v}_n\xi_n + \vec{v}_1 g_1 + \vec{v}_2 g_2 + \cdots + \vec{v}_n g_n \\ &= \vec{v}_1\lambda_1\xi_1 + \vec{v}_2\lambda_2\xi_2 + \cdots + \vec{v}_n\lambda_n\xi_n + \vec{v}_1\, g_1 + \vec{v}_2\, g_2 + \cdots + \vec{v}_n\, g_n \\ &= \vec{v}_1(\lambda_1\xi_1 + g_1) + \vec{v}_2(\lambda_2\xi_2 + g_2) + \cdots + \vec{v}_n(\lambda_n\xi_n + g_n). \end{aligned}$$

We identify the coefficients of the eigenvectors to get the equations

$$\begin{aligned} \xi_1'' &= \lambda_1\xi_1 + g_1, \\ \xi_2'' &= \lambda_2\xi_2 + g_2, \\ &\vdots \\ \xi_n'' &= \lambda_n\xi_n + g_n. \end{aligned}$$

Each one of these equations is independent of the others. We solve each equation using the methods of chapter 2. We write $\vec{x}(t) = \vec{v}_1\xi_1(t) + \vec{v}_2\xi_2(t) + \cdots + \vec{v}_n\xi_n(t)$, and we are done; we have a particular solution. We find the general solutions for ξ_1 through ξ_n, and again $\vec{x}(t) = \vec{v}_1\xi_1(t) + \vec{v}_2\,\xi_2(t) + \cdots + \vec{v}_n\xi_n(t)$ is the general solution (and not just a particular solution).

Example 4.9.4: Let us do the same example from before using this method.

Solution: The equation is

$$\vec{x}'' = \begin{bmatrix} -3 & 1 \\ 2 & -2 \end{bmatrix}\vec{x} + \begin{bmatrix} 0 \\ 2 \end{bmatrix}\cos(3t).$$

The eigenvalues are -1 and -4, with eigenvectors $\left[\begin{smallmatrix}1\\2\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}1\\-1\end{smallmatrix}\right]$. Therefore $E = \left[\begin{smallmatrix}1 & 1\\2 & -1\end{smallmatrix}\right]$ and $E^{-1} = \frac{1}{3}\left[\begin{smallmatrix}1 & 1\\2 & -1\end{smallmatrix}\right]$. Therefore,

$$\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = E^{-1}\vec{f}(t) = \frac{1}{3}\begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}\begin{bmatrix} 0 \\ 2\cos(3t) \end{bmatrix} = \begin{bmatrix} \frac{2}{3}\cos(3t) \\ \frac{-2}{3}\cos(3t) \end{bmatrix}.$$

So after the whole song and dance of plugging in, the equations we get are

$$\xi_1'' = -\xi_1 + \frac{2}{3}\cos(3t), \qquad \xi_2'' = -4\,\xi_2 - \frac{2}{3}\cos(3t).$$

For each equation we use the method of undetermined coefficients. We try $C_1\cos(3t)$ for the first equation and $C_2\cos(3t)$ for the second equation. We plug in to get

$$\begin{aligned} -9C_1\cos(3t) &= -C_1\cos(3t) + \frac{2}{3}\cos(3t), \\ -9C_2\cos(3t) &= -4C_2\cos(3t) - \frac{2}{3}\cos(3t). \end{aligned}$$

We solve each of these equations separately. We get $-9C_1 = -C_1 + 2/3$ and $-9C_2 = -4C_2 - 2/3$. And hence $C_1 = -1/12$ and $C_2 = 2/15$. So our particular solution is

$$\vec{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}\left(\frac{-1}{12}\cos(3t)\right) + \begin{bmatrix} 1 \\ -1 \end{bmatrix}\left(\frac{2}{15}\cos(3t)\right) = \begin{bmatrix} 1/20 \\ -3/10 \end{bmatrix}\cos(3t).$$

This solution matches what we got previously. □

4.9.5 Exercises

***Exercise* 4.9.3:** *Find a particular solution to*

$$\vec{x}'' = \begin{bmatrix} -3 & 1 \\ 2 & -2 \end{bmatrix}\vec{x} + \begin{bmatrix} 0 \\ 2 \end{bmatrix}\cos(2t).$$

Exercise* 4.9.4: *Find the general solution to* $\left[\begin{smallmatrix}1&0&0\\0&2&0\\0&0&3\end{smallmatrix}\right]\vec{x}'' = \left[\begin{smallmatrix}-3&0&0\\2&-4&0\\0&6&-3\end{smallmatrix}\right]\vec{x} + \left[\begin{smallmatrix}\cos(2t)\\0\\0\end{smallmatrix}\right]$.

***Exercise* 4.9.5** (challenging)**:** *Let us take the example in Figure 4.17 on page 348 with the same parameters as before: $m_1 = 2$, $k_1 = 4$, and $k_2 = 2$, except for m_2, which is unknown. Suppose that there is a force $\cos(5t)$ acting on the first mass. Find an m_2 such that there exists a particular solution where the first mass does not move.*

Note: This idea is called dynamic damping. *In practice there will be a small amount of damping and so any transient solution will disappear and after long enough time, the first mass will always come to a stop.*

***Exercise* 4.9.6:** *Let us take the Example 4.9.2 on page 350, but that at time of impact, car 2 is moving to the left at the speed of 3 m/s.*

a) *Find the behavior of the system after linkup.*

b) *Will the second car hit the wall, or will it be moving away from the wall as time goes on?*

c) *At what speed would the first car have to be traveling for the system to essentially stay in place after linkup?*

***Exercise* 4.9.7:** *Let us take the example in Figure 4.17 on page 348 with parameters* $m_1 = m_2 = 1$, $k_1 = k_2 = 1$. *Does there exist a set of initial conditions for which the first cart moves but the second cart does not? If so, find those conditions. If not, argue why not.*

Exercise* 4.9.8: *Suppose there are three carts of equal mass* m *and connected by two springs of constant* k *(and no connections to walls). Set up the system and find its general solution.*

Exercise* 4.9.9: *Suppose a cart of mass 2 kg is attached by a spring of constant* $k = 1$ *to a cart of mass 3 kg, which is attached to the wall by a spring also of constant* $k = 1$. *Suppose that the initial position of the first cart is 1 meter in the positive direction from the rest position, and the second mass starts at the rest position. The masses are not moving and are let go. Find the position of the second mass as a function of time.*

***Exercise* 4.9.10:** *Find the general solution to* $x_1'' = -6x_1 + 3x_2 + \cos(t)$, $x_2'' = 2x_1 - 7x_2 + 3\cos(t)$,

a) *using eigenvector decomposition,* b) *using undetermined coefficients.*

***Exercise* 4.9.11:** *Find the general solution to* $x_1'' = -6x_1 + 3x_2 + \cos(2t)$, $x_2'' = 2x_1 - 7x_2 + 3\cos(2t)$,

a) *using eigenvector decomposition,* b) *using undetermined coefficients.*

Exercise* 4.9.12: *Solve* $x_1'' = -3x_1 + x_2 + t$, $x_2'' = 9x_1 + 5x_2 + \cos(t)$ *with initial conditions* $x_1(0) = 0$, $x_2(0) = 0$, $x_1'(0) = 0$, $x_2'(0) = 0$, *using eigenvector decomposition.*

4.10 Matrix exponentials

Attribution: [JL], §3.8.

Learning Objectives

After this section, you will be able to:

- Compute the exponential of a matrix and
- Use the matrix exponential to solve linear systems of differential equations.

4.10.1 Definition

There is another way of finding a fundamental matrix solution of a system. Consider the constant coefficient equation

$$\vec{x}' = P\vec{x}.$$

If this would be just one equation (when P is a number or a 1×1 matrix), then the solution would be

$$\vec{x} = e^{Pt}.$$

That doesn't make sense if P is a larger matrix, but essentially the same computation that led to the above works for matrices when we define e^{Pt} properly. First let us write down the Taylor series for e^{at} for some number a:

$$e^{at} = 1 + at + \frac{(at)^2}{2} + \frac{(at)^3}{6} + \frac{(at)^4}{24} + \cdots = \sum_{k=0}^{\infty} \frac{(at)^k}{k!}.$$

Recall $k! = 1 \cdot 2 \cdot 3 \cdots k$ is the factorial, and $0! = 1$. We differentiate this series term by term

$$\frac{d}{dt}\left(e^{at}\right) = 0 + a + a^2 t + \frac{a^3 t^2}{2} + \frac{a^4 t^3}{6} + \cdots = a\left(1 + at + \frac{(at)^2}{2} + \frac{(at)^3}{6} + \cdots\right) = ae^{at}.$$

Maybe we can try the same trick with matrices.

Definition 4.10.1

For an $n \times n$ matrix A we define the *matrix exponential* as

$$e^A \stackrel{\text{def}}{=} I + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 + \cdots + \frac{1}{k!}A^k + \cdots$$

Let us not worry about convergence. The series really does always converge. We usually write Pt as tP by convention when P is a matrix. With this small change and by the exact same calculation as above we have that

$$\frac{d}{dt}\left(e^{tP}\right) = Pe^{tP}.$$

Now P and hence e^{tP} is an $n \times n$ matrix. What we are looking for is a vector. In the 1×1 case we would at this point multiply by an arbitrary constant to get the general solution. In the matrix case we multiply by a column vector $\vec{c}$.

Theorem 4.10.1

Let P be an $n \times n$ matrix. Then the general solution to $\vec{x}' = P\vec{x}$ is

$$\vec{x} = e^{tP}\vec{c},$$

where $\vec{c}$ is an arbitrary constant vector. In fact, $\vec{x}(0) = \vec{c}$.

Let us check:

$$\frac{d}{dt}\vec{x} = \frac{d}{dt}\left(e^{tP}\vec{c}\right) = Pe^{tP}\vec{c} = P\vec{x}.$$

Hence e^{tP} is a fundamental matrix solution of the homogeneous system. So if we can compute the matrix exponential, we have another method of solving constant coefficient homogeneous systems. It also makes it easy to solve for initial conditions. To solve $\vec{x}' = A\vec{x}$, $\vec{x}(0) = \vec{b}$, we take the solution

$$\vec{x} = e^{tA}\vec{b}.$$

This equation follows because $e^{0A} = I$, so $\vec{x}(0) = e^{0A}\vec{b} = \vec{b}$.

We mention a drawback of matrix exponentials. In general $e^{A+B} \neq e^A e^B$. The trouble is that matrices do not commute, that is, in general $AB \neq BA$. If you try to prove $e^{A+B} \neq e^A e^B$ using the Taylor series, you will see why the lack of commutativity becomes a problem. However, it is still true that if $AB = BA$, that is, if A and B commute, then $e^{A+B} = e^A e^B$. We will find this fact useful. Let us restate this as a theorem to make a point.

Theorem 4.10.2

If $AB = BA$, then $e^{A+B} = e^A e^B$. Otherwise, $e^{A+B} \neq e^A e^B$ in general.

4.10.2 Simple cases

In some instances it may work to just plug into the series definition. Suppose the matrix is diagonal. For example, $D = \left[\begin{smallmatrix} a & 0 \\ 0 & b \end{smallmatrix}\right]$. Then

$$D^k = \begin{bmatrix} a^k & 0 \\ 0 & b^k \end{bmatrix},$$

and

$$\begin{aligned} e^D &= I + D + \frac{1}{2}D^2 + \frac{1}{6}D^3 + \cdots \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} + \frac{1}{2}\begin{bmatrix} a^2 & 0 \\ 0 & b^2 \end{bmatrix} + \frac{1}{6}\begin{bmatrix} a^3 & 0 \\ 0 & b^3 \end{bmatrix} + \cdots = \begin{bmatrix} e^a & 0 \\ 0 & e^b \end{bmatrix}. \end{aligned}$$

So by this rationale

$$e^I = \begin{bmatrix} e & 0 \\ 0 & e \end{bmatrix} \qquad \text{and} \qquad e^{aI} = \begin{bmatrix} e^a & 0 \\ 0 & e^a \end{bmatrix}.$$

This makes exponentials of certain other matrices easy to compute. For example, the matrix $A = \left[\begin{smallmatrix} 5 & 4 \\ -1 & 1 \end{smallmatrix}\right]$ can be written as $3I + B$ where $B = \left[\begin{smallmatrix} 2 & 4 \\ -1 & -2 \end{smallmatrix}\right]$. Notice that $B^2 = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right]$. So $B^k = 0$ for all $k \geq 2$. Therefore, $e^B = I + B$. Suppose we actually want to compute e^{tA}. The matrices $3tI$ and tB commute (exercise: check this) and $e^{tB} = I + tB$, since $(tB)^2 = t^2B^2 = 0$. We write

$$e^{tA} = e^{3tI+tB} = e^{3tI}e^{tB} = \begin{bmatrix} e^{3t} & 0 \\ 0 & e^{3t} \end{bmatrix} (I + tB) =$$
$$= \begin{bmatrix} e^{3t} & 0 \\ 0 & e^{3t} \end{bmatrix} \begin{bmatrix} 1 + 2t & 4t \\ -t & 1 - 2t \end{bmatrix} = \begin{bmatrix} (1 + 2t)\, e^{3t} & 4te^{3t} \\ -te^{3t} & (1 - 2t)\, e^{3t} \end{bmatrix}.$$

We found a fundamental matrix solution for the system $\vec{x}' = A\vec{x}$. Note that this matrix has a repeated eigenvalue with a defect; there is only one eigenvector for the eigenvalue 3. So we found a perhaps easier way to handle this case. In fact, if a matrix A is 2×2 and has an eigenvalue λ of multiplicity 2, then either $A = \lambda I$, or $A = \lambda I + B$ where $B^2 = 0$. This is a good exercise.

***Exercise* 4.10.1:** *Suppose that A is 2×2 and λ is the only eigenvalue. Show that $(A - \lambda I)^2 = 0$, and therefore that we can write $A = \lambda I + B$, where $B^2 = 0$ (and possibly $B = 0$). Hint: First write down what does it mean for the eigenvalue to be of multiplicity 2. You will get an equation for the entries. Now compute the square of B.*

Matrices B such that $B^k = 0$ for some k are called *nilpotent*. Computation of the matrix exponential for nilpotent matrices is easy by just writing down the first k terms of the Taylor series.

4.10.3 General matrices

In general, the exponential is not as easy to compute as above. We usually cannot write a matrix as a sum of commuting matrices where the exponential is simple for each one. But fear not, it is still not too difficult provided we can find enough eigenvectors. First we need the following interesting result about matrix exponentials. For two square matrices A and B, with B invertible, we have

$$e^{BAB^{-1}} = Be^AB^{-1}.$$

This can be seen by writing down the Taylor series. First

$$(BAB^{-1})^2 = BAB^{-1}BAB^{-1} = BAIAB^{-1} = BA^2B^{-1}.$$

And by the same reasoning $(BAB^{-1})^k = BA^kB^{-1}$. Now write the Taylor series for $e^{BAB^{-1}}$:

$$\begin{aligned} e^{BAB^{-1}} &= I + BAB^{-1} + \frac{1}{2}(BAB^{-1})^2 + \frac{1}{6}(BAB^{-1})^3 + \cdots \\ &= BB^{-1} + BAB^{-1} + \frac{1}{2}BA^2B^{-1} + \frac{1}{6}BA^3B^{-1} + \cdots \\ &= B\big(I + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 + \cdots\big)B^{-1} \\ &= Be^AB^{-1}. \end{aligned}$$

Given a square matrix A, we can usually write $A = EDE^{-1}$, where D is diagonal and E invertible. This procedure is called *diagonalization*. If we can do that, the computation of the exponential becomes easy as e^D is just taking the exponential of the entries on the diagonal. Adding t into the mix, we can then compute the exponential

$$e^{tA} = Ee^{tD}E^{-1}.$$

To diagonalize A we need n linearly independent eigenvectors of A. Otherwise, this method of computing the exponential does not work and we need to be trickier, but we will not get into such details. Let E be the matrix with the eigenvectors as columns. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues and let $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n$ be the eigenvectors, then $E = [\,\vec{v}_1 \quad \vec{v}_2 \quad \cdots \quad \vec{v}_n\,]$. Make a diagonal matrix D with the eigenvalues on the diagonal:

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

We compute

$$\begin{aligned} AE &= A[\,\vec{v}_1 \quad \vec{v}_2 \quad \cdots \quad \vec{v}_n\,] \\ &= [\,A\vec{v}_1 \quad A\vec{v}_2 \quad \cdots \quad A\vec{v}_n\,] \\ &= [\,\lambda_1\vec{v}_1 \quad \lambda_2\vec{v}_2 \quad \cdots \quad \lambda_n\vec{v}_n\,] \\ &= [\,\vec{v}_1 \quad \vec{v}_2 \quad \cdots \quad \vec{v}_n\,]D \\ &= ED. \end{aligned}$$

The columns of E are linearly independent as these are linearly independent eigenvectors of A. Hence E is invertible. Since $AE = ED$, we multiply on the right by E^{-1} and we get

$$A = EDE^{-1}.$$

This means that $e^A = Ee^DE^{-1}$. Multiplying the matrix by t we obtain

$$\boxed{e^{tA} = Ee^{tD}E^{-1} = E\begin{bmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n t} \end{bmatrix}E^{-1}.} \tag{4.19}$$

The formula (4.19), therefore, gives the formula for computing a fundamental matrix solution e^{tA} for the system $\vec{x}' = A\vec{x}$, in the case where we have n linearly independent eigenvectors.

This computation still works when the eigenvalues and eigenvectors are complex, though then you have to compute with complex numbers. It is clear from the definition that if A is real, then e^{tA} is real. So you will only need complex numbers in the computation and not for the result. You may need to apply Euler's formula to simplify the result. If simplified properly, the final matrix will not have any complex numbers in it.

Example 4.10.1: Compute a fundamental matrix solution using the matrix exponential for the system

$$\begin{bmatrix} x \\ y \end{bmatrix}' = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

Then compute the particular solution for the initial conditions $x(0) = 4$ and $y(0) = 2$.

Let A be the coefficient matrix $\left[\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right]$. We first compute (exercise) that the eigenvalues are 3 and -1 and corresponding eigenvectors are $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$. Hence the diagonalization of A is

$$\underbrace{\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}}_{E} \underbrace{\begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1}}_{E^{-1}}.$$

We write

$$\begin{aligned} e^{tA} = Ee^{tD}E^{-1} &= \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} e^{3t} & 0 \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} e^{3t} & 0 \\ 0 & e^{-t} \end{bmatrix} \frac{-1}{2} \begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix} \\ &= \frac{-1}{2} \begin{bmatrix} e^{3t} & e^{-t} \\ e^{3t} & -e^{-t} \end{bmatrix} \begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix} \\ &= \frac{-1}{2} \begin{bmatrix} -e^{3t} - e^{-t} & -e^{3t} + e^{-t} \\ -e^{3t} + e^{-t} & -e^{3t} - e^{-t} \end{bmatrix} = \begin{bmatrix} \frac{e^{3t}+e^{-t}}{2} & \frac{e^{3t}-e^{-t}}{2} \\ \frac{e^{3t}-e^{-t}}{2} & \frac{e^{3t}+e^{-t}}{2} \end{bmatrix}. \end{aligned}$$

The initial conditions are $x(0) = 4$ and $y(0) = 2$. Hence, by the property that $e^{0A} = I$ we find that the particular solution we are looking for is $e^{tA}\vec{b}$ where $\vec{b}$ is $\left[\begin{smallmatrix} 4 \\ 2 \end{smallmatrix}\right]$. Then the particular solution we are looking for is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{e^{3t}+e^{-t}}{2} & \frac{e^{3t}-e^{-t}}{2} \\ \frac{e^{3t}-e^{-t}}{2} & \frac{e^{3t}+e^{-t}}{2} \end{bmatrix} \begin{bmatrix} 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 2e^{3t} + 2e^{-t} + e^{3t} - e^{-t} \\ 2e^{3t} - 2e^{-t} + e^{3t} + e^{-t} \end{bmatrix} = \begin{bmatrix} 3e^{3t} + e^{-t} \\ 3e^{3t} - e^{-t} \end{bmatrix}.$$

4.10.4 Fundamental matrix solutions

We note that if you can compute a fundamental matrix solution in a different way, you can use this to find the matrix exponential e^{tA}. A fundamental matrix solution of a system of ODEs is not unique. The exponential is the fundamental matrix solution with the property

that for $t = 0$ we get the identity matrix. So we must find the right fundamental matrix solution. Let X be any fundamental matrix solution to $\vec{x}' = A\vec{x}$. Then we claim

$$e^{tA} = X(t)\,[X(0)]^{-1}\,.$$

Clearly, if we plug $t = 0$ into $X(t)\,[X(0)]^{-1}$ we get the identity. We can multiply a fundamental matrix solution on the right by any constant invertible matrix and we still get a fundamental matrix solution. All we are doing is changing what are the arbitrary constants in the general solution $\vec{x}(t) = X(t)\,\vec{c}$.

4.10.5 Approximations

If you think about it, the computation of any fundamental matrix solution X using the eigenvalue method is just as difficult as the computation of e^{tA}. So perhaps we did not gain much by this new tool. However, the Taylor series expansion actually gives us a way to approximate solutions, which the eigenvalue method did not.

The simplest thing we can do is to just compute the series up to a certain number of terms. There are better ways to approximate the exponential*. In many cases however, few terms of the Taylor series give a reasonable approximation for the exponential and may suffice for the application. For example, let us compute the first 4 terms of the series for the matrix $A = \left[\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right]$.

$$e^{tA} \approx I + tA + \frac{t^2}{2}A^2 + \frac{t^3}{6}A^3 = I + t\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} + t^2\begin{bmatrix} \frac{5}{2} & 2 \\ 2 & \frac{5}{2} \end{bmatrix} + t^3\begin{bmatrix} \frac{13}{6} & \frac{7}{3} \\ \frac{7}{3} & \frac{13}{6} \end{bmatrix} =$$

$$= \begin{bmatrix} 1 + t + \frac{5}{2}\,t^2 + \frac{13}{6}\,t^3 & 2\,t + 2\,t^2 + \frac{7}{3}\,t^3 \\ 2\,t + 2\,t^2 + \frac{7}{3}\,t^3 & 1 + t + \frac{5}{2}\,t^2 + \frac{13}{6}\,t^3 \end{bmatrix}.$$

Just like the scalar version of the Taylor series approximation, the approximation will be better for small t and worse for larger t. For larger t, we will generally have to compute more terms. Let us see how we stack up against the real solution with $t = 0.1$. The approximate solution is approximately (rounded to 8 decimal places)

$$e^{0.1\,A} \approx I + 0.1\,A + \frac{0.1^2}{2}A^2 + \frac{0.1^3}{6}A^3 = \begin{bmatrix} 1.12716667 & 0.22233333 \\ 0.22233333 & 1.12716667 \end{bmatrix}.$$

And plugging $t = 0.1$ into the real solution (rounded to 8 decimal places) we get

$$e^{0.1\,A} = \begin{bmatrix} 1.12734811 & 0.22251069 \\ 0.22251069 & 1.12734811 \end{bmatrix}.$$

Not bad at all! Although if we take the same approximation for $t = 1$ we get

$$I + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 = \begin{bmatrix} 6.66666667 & 6.33333333 \\ 6.33333333 & 6.66666667 \end{bmatrix},$$

*C. Moler and C.F. Van Loan, *Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later*, SIAM Review 45 (1), 2003, 3–49

while the real value is (again rounded to 8 decimal places)

$$e^A = \begin{bmatrix} 10.22670818 & 9.85882874 \\ 9.85882874 & 10.22670818 \end{bmatrix}.$$

So the approximation is not very good once we get up to $t = 1$. To get a good approximation at $t = 1$ (say up to 2 decimal places) we would need to go up to the 11^{th} power (exercise).

4.10.6 Non-Homogeneous Systems

Integrating factor

Now that we have matrix exponentials, we can try to use them to help us solve non-homogeneous systems of differential equations. First, let's recall what we did for first order equations. If we have an equation of the form

$$x'(t) + px(t) = f(t)$$

where will assume that p is constant (even though it doesn't have to be). We would go about solving this problem by multiplying both sides of the equation by e^{pt}, writing the left-hand side as a product rule, integrating both sides, and solving.

With matrix exponentials, we can do exactly the same thing with first order systems. Let us focus on the nonhomogeneous first order equation

$$\vec{x}'(t) = A\vec{x}(t) + \vec{f}(t),$$

where A is a constant matrix. The method we look at here is the *integrating factor method*. For simplicity we rewrite the equation as

$$\vec{x}'(t) + P\vec{x}(t) = \vec{f}(t),$$

where $P = -A$. We multiply both sides of the equation by e^{tP} (being mindful that we are dealing with matrices that may not commute) to obtain

$$e^{tP}\vec{x}'(t) + e^{tP}P\vec{x}(t) = e^{tP}\vec{f}(t).$$

We notice that $Pe^{tP} = e^{tP}P$. This fact follows by writing down the series definition of e^{tP}:

$$\begin{aligned} Pe^{tP} &= P\left(I + tP + \frac{1}{2}(tP)^2 + \cdots\right) = P + tP^2 + \frac{1}{2}t^2P^3 + \cdots = \\ &= \left(I + tP + \frac{1}{2}(tP)^2 + \cdots\right)P = e^{tP}P. \end{aligned}$$

So $\frac{d}{dt}\left(e^{tP}\right) = Pe^{tP} = e^{tP}P$. The product rule says

$$\frac{d}{dt}\Big(e^{tP}\vec{x}(t)\Big) = e^{tP}\vec{x}'(t) + e^{tP}P\vec{x}(t),$$

and so

$$\frac{d}{dt}\Big(e^{tP}\vec{x}(t)\Big) = e^{tP}\vec{f}(t).$$

We can now integrate. That is, we integrate each component of the vector separately

$$e^{tP}\vec{x}(t) = \int e^{tP}\vec{f}(t)\,dt + \vec{c}.$$

In Exercise 4.10.10, you will compute and verify that $(e^{tP})^{-1} = e^{-tP}$. Therefore, we obtain

$$\vec{x}(t) = e^{-tP}\int e^{tP}\vec{f}(t)\,dt + e^{-tP}\vec{c}.$$

Perhaps it is better understood as a definite integral. In this case it will be easy to also solve for the initial conditions. Consider the equation with initial conditions

$$\vec{x}'(t) + P\vec{x}(t) = \vec{f}(t), \qquad \vec{x}(0) = \vec{b}.$$

The solution can then be written as

$$\boxed{\vec{x}(t) = e^{-tP}\int_0^t e^{sP}\vec{f}(s)\,ds + e^{-tP}\vec{b}.} \tag{4.20}$$

Again, the integration means that each component of the vector $e^{sP}\vec{f}(s)$ is integrated separately. It is not hard to see that (4.20) really does satisfy the initial condition $\vec{x}(0) = \vec{b}$.

$$\vec{x}(0) = e^{-0P}\int_0^0 e^{sP}\vec{f}(s)\,ds + e^{-0P}\vec{b} = I\vec{b} = \vec{b}.$$

Example 4.10.2: Suppose that we have the system

$$\begin{aligned} x_1' + 5x_1 - 3x_2 &= e^t, \\ x_2' + 3x_1 - x_2 &= 0, \end{aligned}$$

with initial conditions $x_1(0) = 1, x_2(0) = 0$.

Solution: Let us write the system as

$$\vec{x}' + \begin{bmatrix} 5 & -3 \\ 3 & -1 \end{bmatrix}\vec{x} = \begin{bmatrix} e^t \\ 0 \end{bmatrix}, \qquad \vec{x}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The matrix $P = \left[\begin{smallmatrix} 5 & -3 \\ 3 & -1 \end{smallmatrix}\right]$ has a doubled eigenvalue 2 with defect 1, and we leave it as an exercise to double check we computed e^{tP} correctly. Once we have e^{tP}, we find e^{-tP}, simply by negating t.

$$e^{tP} = \begin{bmatrix} (1+3t)\,e^{2t} & -3te^{2t} \\ 3te^{2t} & (1-3t)\,e^{2t} \end{bmatrix}, \qquad e^{-tP} = \begin{bmatrix} (1-3t)\,e^{-2t} & 3te^{-2t} \\ -3te^{-2t} & (1+3t)\,e^{-2t} \end{bmatrix}.$$

Instead of computing the whole formula at once, let us do it in stages. First

$$\begin{aligned}\int_0^t e^{sP}\vec{f}(s)\,ds &= \int_0^t \begin{bmatrix}(1+3s)\,e^{2s} & -3se^{2s}\\ 3se^{2s} & (1-3s)\,e^{2s}\end{bmatrix}\begin{bmatrix}e^s\\0\end{bmatrix}\,ds\\ &= \int_0^t \begin{bmatrix}(1+3s)\,e^{3s}\\ 3se^{3s}\end{bmatrix}\,ds\\ &= \begin{bmatrix}\int_0^t (1+3s)\,e^{3s}\,ds\\ \int_0^t 3se^{3s}\,ds\end{bmatrix}\\ &= \begin{bmatrix}te^{3t}\\ \frac{(3t-1)\,e^{3t}+1}{3}\end{bmatrix} \qquad \text{(used integration by parts).}\end{aligned}$$

Then

$$\begin{aligned}\vec{x}(t) &= e^{-tP}\int_0^t e^{sP}\vec{f}(s)\,ds + e^{-tP}\vec{b}\\ &= \begin{bmatrix}(1-3t)\,e^{-2t} & 3te^{-2t}\\ -3te^{-2t} & (1+3t)\,e^{-2t}\end{bmatrix}\begin{bmatrix}te^{3t}\\ \frac{(3t-1)\,e^{3t}+1}{3}\end{bmatrix} + \begin{bmatrix}(1-3t)\,e^{-2t} & 3te^{-2t}\\ -3te^{-2t} & (1+3t)\,e^{-2t}\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix}\\ &= \begin{bmatrix}te^{-2t}\\ -\frac{e^t}{3}+\left(\frac{1}{3}+t\right)\,e^{-2t}\end{bmatrix} + \begin{bmatrix}(1-3t)\,e^{-2t}\\ -3te^{-2t}\end{bmatrix}\\ &= \begin{bmatrix}(1-2t)\,e^{-2t}\\ -\frac{e^t}{3}+\left(\frac{1}{3}-2t\right)\,e^{-2t}\end{bmatrix}.\end{aligned}$$

Phew!

Let us check that this really works.

$$x_1' + 5x_1 - 3x_2 = (4te^{-2t} - 4e^{-2t}) + 5(1-2t)\,e^{-2t} + e^t - (1-6t)\,e^{-2t} = e^t.$$

Similarly (exercise) $x_2' + 3x_1 - x_2 = 0$. The initial conditions are also satisfied (exercise). ⌟

For systems, the integrating factor method only works if P does not depend on t, that is, P is constant. The problem is that in general

$$\frac{d}{dt}\left[e^{\int P(t)\,dt}\right] \neq P(t)\,e^{\int P(t)\,dt},$$

because matrix multiplication is not commutative.

4.10.7 Exercises

Exercise 4.10.2: *Using the matrix exponential, find a fundamental matrix solution for the system $x' = 3x + y$, $y' = x + 3y$.*

Exercise 4.10.3: *Find e^{tA} for the matrix $A = \left[\begin{smallmatrix}2 & 3\\ 0 & 2\end{smallmatrix}\right]$.*

Exercise 4.10.4:* *Compute e^{tA} where $A = \left[\begin{smallmatrix}1 & -2\\ -2 & 1\end{smallmatrix}\right]$.*

Exercise 4.10.5:* *Compute e^{tA} where $A = \left[\begin{smallmatrix}1 & -3 & 2\\ -2 & 1 & 2\\ -1 & -3 & 4\end{smallmatrix}\right]$.*

Exercise 4.10.6:*

a) *Compute* e^{tA} *where* $A = \left[\begin{smallmatrix} 3 & -1 \\ 1 & 1 \end{smallmatrix}\right]$. b) *Solve* $\vec{x}\,' = A\vec{x}$ *for* $\vec{x}(0) = \left[\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right]$.

***Exercise* 4.10.7:** *Find a fundamental matrix solution for the system* $x_1' = 7x_1 + 4x_2 + 12x_3$, $x_2' = x_1 + 2x_2 + x_3$, $x_3' = -3x_1 - 2x_2 - 5x_3$. *Then find the solution that satisfies* $\vec{x}(0) = \begin{bmatrix} 0 \\ 1 \\ -2 \end{bmatrix}$.

***Exercise* 4.10.8:** *Compute the matrix exponential* e^A *for* $A = \left[\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right]$.

***Exercise* 4.10.9** (challenging)**:** *Suppose* $AB = BA$. *Show that under this assumption,* $e^{A+B} = e^A e^B$.

***Exercise* 4.10.10:** *Use Exercise 4.10.9 to show that* $(e^A)^{-1} = e^{-A}$. *In particular this means that* e^A *is invertible even if* A *is not.*

***Exercise* 4.10.11:** *Let* A *be a* 2×2 *matrix with eigenvalues* -1, 1, *and corresponding eigenvectors* $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$, $\left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$.

a) *Find matrix* A *with these properties.*

b) *Find a fundamental matrix solution to* $\vec{x}' = A\vec{x}$.

c) *Solve the system in with initial conditions* $\vec{x}(0) = \left[\begin{smallmatrix} 2 \\ 3 \end{smallmatrix}\right]$.

***Exercise* 4.10.12:** *Suppose that* A *is an* $n \times n$ *matrix with a repeated eigenvalue* λ *of multiplicity* n. *Suppose that there are* n *linearly independent eigenvectors. Show that the matrix is diagonal, in particular* $A = \lambda I$. *Hint: Use diagonalization and the fact that the identity matrix commutes with every other matrix.*

***Exercise* 4.10.13:** *Let* $A = \left[\begin{smallmatrix} -1 & -1 \\ 1 & -3 \end{smallmatrix}\right]$.

a) *Find* e^{tA}. b) *Solve* $\vec{x}' = A\vec{x}$, $\vec{x}(0) = \left[\begin{smallmatrix} 1 \\ -2 \end{smallmatrix}\right]$.

***Exercise* 4.10.14:** *Let* $A = \left[\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right]$. *Approximate* e^{tA} *by expanding the power series up to the third order.*

Exercise* 4.10.15: *Compute the first 3 terms (up to the second degree) of the Taylor expansion of* e^{tA} *where* $A = \left[\begin{smallmatrix} 2 & 3 \\ 2 & 2 \end{smallmatrix}\right]$ *(Write as a single matrix). Then use it to approximate* $e^{0.1A}$.

***Exercise* 4.10.16:** *For any positive integer* n, *find a formula (or a recipe) for* A^n *for the following matrices:*

a) $\begin{bmatrix} 3 & 0 \\ 0 & 9 \end{bmatrix}$ b) $\begin{bmatrix} 5 & 2 \\ 4 & 7 \end{bmatrix}$ c) $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ d) $\begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$

Exercise* 4.10.17: *For any positive integer* n, *find a formula (or a recipe) for* A^n *for the following matrices:*

a) $\begin{bmatrix} 7 & 4 \\ -5 & -2 \end{bmatrix}$ b) $\begin{bmatrix} -3 & 4 \\ -6 & -7 \end{bmatrix}$ c) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

***Exercise* 4.10.18:** *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} 5 & -6 \\ 3 & -1 \end{bmatrix} \vec{x} + \begin{bmatrix} t \\ 3 \end{bmatrix} \qquad \vec{x}(0) = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

using matrix exponentials.

***Exercise* 4.10.19:** *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} -4 & 2 \\ -9 & 5 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{3t} \\ e^{t} - 1 \end{bmatrix} \qquad \vec{x}(0) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

using matrix exponentials.

***Exercise* 4.10.20:** *Solve the initial value problem*

$$\vec{x}' = \begin{bmatrix} 3 & 2 \\ 0 & 4 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{4t} \\ e^{3t} - t \end{bmatrix} \qquad \vec{x}(0) = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

using matrix exponentials.

Chapter 5

Nonlinear systems

5.1 Linearization, critical points, and stability

Attribution: [JL], §8.1, 8.2.

Learning Objectives

After this section, you will be able to:

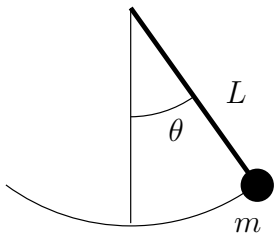
- Find critical points of a non-linear system of differential equations,
- Linearize a non-linear system around a critical point,
- Determine if a critical point of a non-linear system is isolated,
- Use the Jacobian matrix to classify the critical point of a non-linear system, and
- Determine the stability of a critical point from the classification.

Except for a few brief detours in chapter 1, we considered mostly linear equations. Linear equations suffice in many applications, but in reality most phenomena require nonlinear equations. Nonlinear equations, however, are notoriously more difficult to understand than linear ones, and many strange new phenomena appear when we allow our equations to be nonlinear.

Not to worry, we did not waste all this time studying linear equations. Nonlinear equations can often be approximated by linear ones if we only need a solution "locally," for example, only for a short period of time, or only for certain parameters. Understanding specific linear equations can also give us qualitative understanding about a more general nonlinear problem. The idea is similar to what you did in calculus in trying to approximate a function by a line with the right slope.

In § 2.4 we looked at the pendulum of length L. The goal was to solve for the angle $\theta(t)$ as a function of the time t. The equation for the setup is the nonlinear equation

$$\theta'' + \frac{g}{L}\sin\theta = 0.$$



Instead of solving this equation, we solved the rather easier linear

equation

$$\theta'' + \frac{g}{L}\theta = 0.$$

While the solution to the linear equation is not exactly what we were looking for, it is rather close to the original, as long as the angle θ is small and the time period involved is short.

You might ask: Why don't we just solve the nonlinear problem? Well, it might be very difficult, impractical, or impossible to solve analytically, depending on the equation in question. We may not even be interested in the actual solution, we might only be interested in some qualitative idea of what the solution is doing. For example, what happens as time goes to infinity?

5.1.1 Autonomous systems and phase plane analysis

We restrict our attention to a two-dimensional autonomous system

$$x' = f(x, y), \qquad y' = g(x, y),$$

where $f(x, y)$ and $g(x, y)$ are functions of two variables, and the derivatives are taken with respect to time t. Solutions are functions $x(t)$ and $y(t)$ such that

$$x'(t) = f\big(x(t), y(t)\big), \qquad y'(t) = g\big(x(t), y(t)\big).$$

The way we will analyze the system is very similar to § 1.7, where we studied a single autonomous equation. The ideas in two dimensions are the same, but the behavior can be far more complicated.

It may be best to think of the system of equations as the single vector equation

$$\begin{bmatrix} x \\ y \end{bmatrix}' = \begin{bmatrix} f(x, y) \\ g(x, y) \end{bmatrix}. \tag{5.1}$$

As in § 4.1 we draw the *phase portrait* (or *phase diagram*), where each point (x, y) corresponds to a specific state of the system. We draw the *vector field* given at each point (x, y) by the vector $\begin{bmatrix} f(x,y) \\ g(x,y) \end{bmatrix}$. And as before if we find solutions, we draw the trajectories by plotting all points $\big(x(t), y(t)\big)$ for a certain range of t.

Example 5.1.1: Consider the second order equation $x'' = -x + x^2$. Write this equation as a first order nonlinear system

$$x' = y, \qquad y' = -x + x^2.$$

The phase portrait with some trajectories is drawn in Figure 5.1 on the facing page.

From the phase portrait it should be clear that even this simple system has fairly complicated behavior. Some trajectories keep oscillating around the origin, and some go off towards infinity. We will return to this example often, and analyze it completely in this (and the next) section.

If we zoom into the diagram near a point where $\begin{bmatrix} f(x,y) \\ g(x,y) \end{bmatrix}$ is not zero, then nearby the arrows point generally in essentially that same direction and have essentially the same magnitude. In

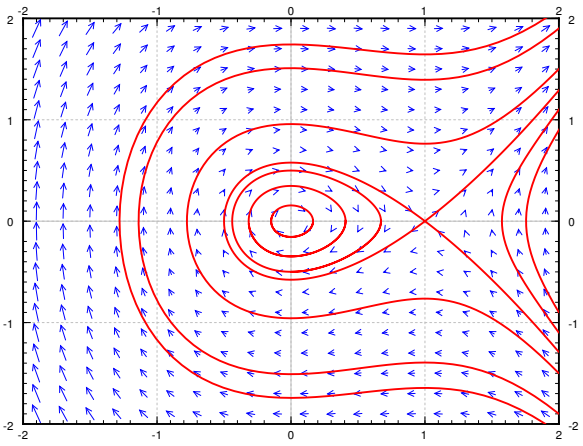


Figure 5.1: Phase portrait with some trajectories of $x' = y$, $y' = -x + x^2$.

other words the behavior is not that interesting near such a point. We are of course assuming that $f(x, y)$ and $g(x, y)$ are continuous.

Let us concentrate on those points in the phase diagram above where the trajectories seem to start, end, or go around. We see two such points: $(0, 0)$ and $(1, 0)$. The trajectories seem to go around the point $(0, 0)$, and they seem to either go in or out of the point $(1, 0)$. These points are precisely those points where the derivatives of both x and y are zero.

Definition 5.1.1

The *critical points* of a system of differential equations

$$\frac{dx}{dt} = f(x, y)$$
$$\frac{dy}{dt} = g(x, y)$$

are the points (x, y) such that

$$\begin{bmatrix} f(x, y) \\ g(x, y) \end{bmatrix} = \vec{0}.$$

In other words, these are the points where both $f(x, y) = 0$ and $g(x, y) = 0$.

The critical points are where the behavior of the system is in some sense the most complicated. If $\left[\begin{smallmatrix} f(x,y) \\ g(x,y) \end{smallmatrix}\right]$ is zero, then nearby, the vector can point in any direction whatsoever. Also, the trajectories are either going towards, away from, or around these points, so if we are looking for long-term qualitative behavior of the system, we should look at what is happening near the critical points.

Critical points are also sometimes called *equilibria*, since we have so-called *equilibrium solutions* at critical points. If (x_0, y_0) is a critical point, then we have the solutions

$$x(t) = x_0, \quad y(t) = y_0.$$

In Example 5.1.1 on page 370, there are two equilibrium solutions:

$$x(t) = 0, \quad y(t) = 0, \qquad \text{and} \qquad x(t) = 1, \quad y(t) = 0.$$

The discussion here should seem a bit familiar; it is the same as how we formulated equilibrium solutions to autonomous differential equations in in § 1.7.

5.1.2 Linearization

How do linear systems fit into this approach? For a linear, homogeneous system of two variables defined by

$$\vec{x}' = A\vec{x}$$

where A is an invertible matrix, the only critical point is the origin $(0, 0)$. Since A is invertible, the only vector that satisfies $A\vec{x} = 0$ is $\vec{x} = 0$, see § 3.5. (This also applies beyond two variables, but we'll stick to that for simplicity.) In § 4.7 we studied the behavior of a homogeneous linear system of two equations near a critical point. Let us put the understanding we gained in that section to good use understanding what happens near critical points of nonlinear systems.

In calculus we learned to estimate a function by taking its derivative and linearizing. We work similarly with nonlinear systems of ODE. Suppose (x_0, y_0) is a critical point. In order to linearize the system of differential equations, we want to linearize the two functions $f(x, y)$ and $g(x, y)$ that define this system. To do so, we will replace f and g by the tangent plane approximation to the functions. That is, if we set $z = f(x, y)$, the tangent plane is given by

$$L_f(x, y) = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0).$$

Since (x_0, y_0) is a critical point, we know that $f(x_0, y_0) = 0$, so the tangent plane is given by

$$L_f(x, y) = f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0).$$

Similarly, the tangent plane for $g(x, y)$ near the critical point (x_0, y_0) is given by

$$L_g(x, y) = g_x(x_0, y_0)(x - x_0) + g_y(x_0, y_0)(y - y_0).$$

The idea of linearization in calculus was that we could use the tangent line or tangent plane to approximate a function near to a given point. For systems of differential equations, the idea is that we can approximate the solutions to the system of differential equations by the solutions to the linearized systems as long as we stay near the critical point. That means that we can approximate the solution to

$$\begin{aligned} \frac{dx}{dt} &= f(x, y) \\ \frac{dy}{dt} &= g(x, y) \end{aligned}$$

near the critical point (x_0, y_0) by the solution to the system

$$\begin{aligned} \frac{dx}{dt} &= f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) \\ \frac{dy}{dt} &= g_x(x_0, y_0)(x - x_0) + g_y(x_0, y_0)(y - y_0) \end{aligned}$$

Next, change variables to (u, v), so that $(u, v) = (0, 0)$ corresponds to (x_0, y_0). That is,

$$u = x - x_0, \qquad v = y - y_0,$$

which is not going to affect our differential equations because x_0 and y_0 are constant.

Since $\frac{dx}{dt} = \frac{du}{dt}$ and $\frac{dy}{dt} = \frac{dv}{dt}$, we can rewrite the approximation system as

$$\begin{aligned} \frac{du}{dt} &= f_x(x_0, y_0)u + f_y(x_0, y_0)v \\ \frac{dv}{dt} &= g_x(x_0, y_0)u + g_y(x_0, y_0)v \end{aligned}$$

In multivariable calculus you may have seen that the several variables version of the derivative is the *Jacobian matrix**. The Jacobian matrix of the vector-valued function $\begin{bmatrix} f(x,y) \\ g(x,y) \end{bmatrix}$ at (x_0, y_0) is

$$\begin{bmatrix} \frac{\partial f}{\partial x}(x_0, y_0) & \frac{\partial f}{\partial y}(x_0, y_0) \\ \frac{\partial g}{\partial x}(x_0, y_0) & \frac{\partial g}{\partial y}(x_0, y_0) \end{bmatrix}.$$

This matrix gives the best linear approximation as u and v (and therefore x and y) vary.

Definition 5.1.2

The *linearization* of the equation (5.1) as the linear system

$$\begin{bmatrix} u \\ v \end{bmatrix}' = \begin{bmatrix} \frac{\partial f}{\partial x}(x_0, y_0) & \frac{\partial f}{\partial y}(x_0, y_0) \\ \frac{\partial g}{\partial x}(x_0, y_0) & \frac{\partial g}{\partial y}(x_0, y_0) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}.$$

Example 5.1.2: Determine the linearization of the system of differential equations in Example 5.1.1: $x' = y$, $y' = -x + x^2$ at all of its critical points.

Solution: There are two critical points, $(0, 0)$ and $(1, 0)$. The Jacobian matrix at any point is

$$\begin{bmatrix} \frac{\partial f}{\partial x}(x, y) & \frac{\partial f}{\partial y}(x, y) \\ \frac{\partial g}{\partial x}(x, y) & \frac{\partial g}{\partial y}(x, y) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 + 2x & 0 \end{bmatrix}.$$

Therefore at $(0, 0)$, we have $u = x$ and $v = y$, and the linearization is

$$\begin{bmatrix} u \\ v \end{bmatrix}' = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}.$$

At the point $(1, 0)$, we have $u = x - 1$ and $v = y$, and the linearization is

$$\begin{bmatrix} u \\ v \end{bmatrix}' = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}.$$

The phase diagrams of the two linearizations at the point $(0, 0)$ and $(1, 0)$ are given in Figure 5.2 on the following page. Note that the variables are now u and v. Compare Figure 5.2 with Figure 5.1 on page 371, and look especially at the behavior near the critical points.

*Named for the German mathematician Carl Gustav Jacob Jacobi (1804–1851).

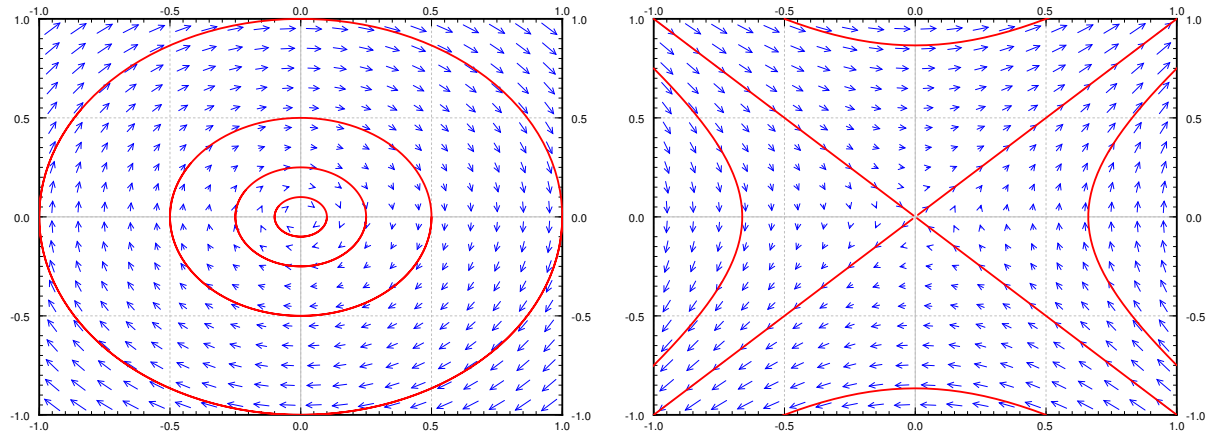


Figure 5.2: Phase diagram with some trajectories of linearizations at the critical points $(0,0)$ *(left) and* $(1,0)$ *(right) of* $x' = y$, $y' = -x + x^2$.

5.1.3 Isolated critical points and almost linear systems

The next step in this process is to try to figure out a way to analyze what is happening to a non-linear system of differential equations near equilibrium solutions *without* using a slope field/phase portrait. We would like to be able to determine this from the equations alone, not any of the pictures that come from them. Thankfully, our ability to analyze linear systems helps us accomplish this goal.

Definition 5.1.3

A critical point is *isolated* if it is the only critical point in some small "neighborhood" of the point.

That is, if we zoom in far enough it is the only critical point we see. In the example above, the critical point was isolated. If on the other hand there would be a whole curve of critical points, then it would not be isolated. For example, the system

$$x' = y(x-1) \qquad y' = (x-2)(x-1)$$

has the entire line $x = 1$ as critical points. Therefore, these are not isolated.

Definition 5.1.4

A system is called *almost linear* at a critical point (x_0, y_0), if the critical point is isolated and the Jacobian matrix at the point is invertible, or equivalently if the linearized system has an isolated critical point.

This is also equivalent to zero not being an eigenvalue of the Jacobian matrix at the critical point. In such a case, the nonlinear terms are very small and the system behaves like its linearization, at least if we are close to the critical point.

For example, the system in Examples 5.1.1 and 5.1.2 has two isolated critical points $(0,0)$ and $(0,1)$, and is almost linear at both critical points as the Jacobian matrices at both points, $\left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, are invertible.

On the other hand, the system $x' = x^2$, $y' = y^2$ has an isolated critical point at $(0,0)$, however the Jacobian matrix

$$\begin{bmatrix} 2x & 0 \\ 0 & 2y \end{bmatrix}$$

is zero when $(x,y) = (0,0)$. So the system is not almost linear. Even a worse example is the system $x' = x$, $y' = x^2$, which does not have isolated critical points; x' and y' are both zero whenever $x = 0$, that is, the entire y-axis.

Fortunately, most often critical points are isolated, and the system is almost linear at the critical points. So if we learn what happens there, we will have figured out the majority of situations that arise in applications.

5.1.4 Stability and classification of isolated critical points

Once we have an isolated critical point, the system is almost linear at that critical point, and we computed the associated linearized system, we can classify what happens to the solutions. The classifications for linear two-variable systems from § 4.7 are generally the same as what we use here, with one minor caveat. Let us list the behaviors depending on the eigenvalues of the Jacobian matrix at the critical point in Table 5.1. This table is very similar to Table 4.1 on page 324, with the exception of missing "center" points. The repeated eigenvalue cases are also missing. They behave similarly to the real eigenvalue descriptions in the table below, but similar to centers, the behavior can change slightly. It can behave like either a spiral or a node, but will be either a source or sink based on the sign of the repeated eigenvalue. We will discuss centers later, as they are more complicated.

Eigenvalues of the Jacobian matrix	Behavior	Stability
real and both positive	source / unstable node	unstable
real and both negative	sink / stable node	asymptotically stable
real and opposite signs	saddle	unstable
complex with positive real part	spiral source	unstable
complex with negative real part	spiral sink	asymptotically stable

Table 5.1: Behavior of an almost linear system near an isolated critical point.

In the third column, we mark points as *asymptotically stable* or *unstable*.

Definition 5.1.5

Let (x_0, y_0) be a critical point for a non-linear system of two differential equations.

1. We say that the critical point is a *stable critical point* if, given any small distance ϵ to (x_0, y_0), and any initial condition within a perhaps smaller radius around (x_0, y_0), the trajectory of the system never goes further away from (x_0, y_0) than ϵ.

2. The critical point is an *unstable critical point* if it is not stable; that is, there are trajectories that start within a distance ϵ of (x_0, y_0) and end up farther than ϵ from that point.

3. The critical point is called *asymptotically stable* if given any initial condition sufficiently close to (x_0, y_0) and any solution $\big(x(t), y(t)\big)$ satisfying that condition, then
$$\lim_{t\to\infty} \big(x(t), y(t)\big) = (x_0, y_0).$$

Informally, a point is stable if we start close to a critical point and follow a trajectory we either go towards, or at least not away from, this critical point. If the point is asymptotically stable, then any trajectory for a sufficiently close initial condition goes towards the critical point (x_0, y_0), and unstable means that, in general, trajectories move away from the critical point.

Example 5.1.3: Find and analyze the critical points of $x' = -y - x^2$, $y' = -x + y^2$.

Solution: See Figure 5.3 for the phase diagram. Let us find the critical points. These are the points where $-y - x^2 = 0$ and $-x + y^2 = 0$. The first equation means $y = -x^2$, and so $y^2 = x^4$. Plugging into the second equation we obtain $-x + x^4 = 0$. Factoring we obtain $x(1 - x^3) = 0$. Since we are looking only for real solutions we get either $x = 0$ or $x = 1$. Solving for the corresponding y using $y = -x^2$, we get two critical points, one being $(0, 0)$ and the other being $(1, -1)$. Clearly the critical points are isolated.

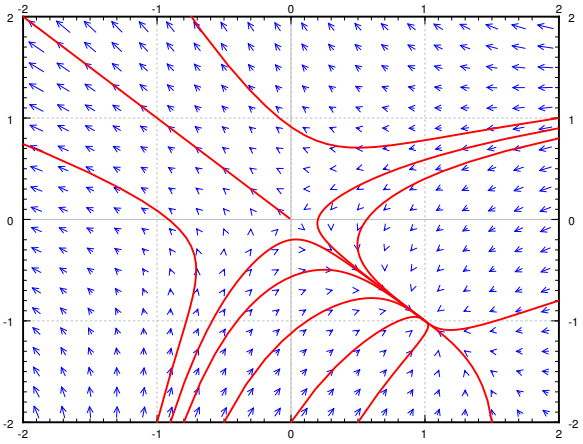


Figure 5.3: The phase portrait with few sample trajectories of $x' = -y - x^2$, $y' = -x + y^2$.

Let us compute the Jacobian matrix:

$$\begin{bmatrix} -2x & -1 \\ -1 & 2y \end{bmatrix}.$$

At the point $(0,0)$ we get the matrix $\left[\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right]$ and so the two eigenvalues are 1 and -1. As the matrix is invertible, the system is almost linear at $(0,0)$. As the eigenvalues are real and of opposite signs, we get a saddle point, which is an unstable equilibrium point. Looking at the phase portrait, we can see trajectories that would start near $(0,0)$ and end up farther away from $(0,0)$. These trajectories may end up at $(1,-1)$, but that is away from $(0,0)$.

At the point $(1,-1)$ we get the matrix $\left[\begin{smallmatrix} -2 & -1 \\ -1 & -2 \end{smallmatrix}\right]$ and computing the eigenvalues we get -1, -3. The matrix is invertible, and so the system is almost linear at $(1,-1)$. As we have real eigenvalues and both negative, the critical point is a sink, and therefore an asymptotically stable equilibrium point. That is, if we start with any point $(x(0),y(0))$ close to $(1,-1)$ as an initial condition and plot a trajectory, it approaches $(1,-1)$. In other words,

$$\lim_{t\to\infty} \big(x(t),y(t)\big) = (1,-1).$$

As you can see from the diagram, this behavior is true even for some initial points quite far from $(1,-1)$, but it is definitely not true for all initial points. ⌟

Example 5.1.4: Find and analyze the critical points of $x' = y + y^2e^x$, $y' = x$.

Solution: First let us find the critical points. These are the points where $y + y^2e^x = 0$ and $x = 0$. Simplifying we get $0 = y + y^2 = y(y+1)$. So the critical points are $(0,0)$ and $(0,-1)$, and hence are isolated. Let us compute the Jacobian matrix:

$$\begin{bmatrix} y^2e^x & 1+2ye^x \\ 1 & 0 \end{bmatrix}.$$

At the point $(0,0)$ we get the matrix $\left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ and so the two eigenvalues are 1 and -1. As the matrix is invertible, the system is almost linear at $(0,0)$. And, as the eigenvalues are real and of opposite signs, we get a saddle point, which is an unstable equilibrium point.

At the point $(0,-1)$ we get the matrix $\left[\begin{smallmatrix} 1 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ whose eigenvalues are $\frac{1}{2} \pm i\frac{\sqrt{3}}{2}$. The matrix is invertible, and so the system is almost linear at $(0,-1)$. As we have complex eigenvalues with positive real part, the critical point is a spiral source, and therefore an unstable equilibrium point.

See Figure 5.4 on the next page for the phase diagram. Notice the two critical points, and the behavior of the arrows in the vector field around these points. ⌟

5.1.5 The trouble with centers

Recall, a linear system with a center means that trajectories travel in closed elliptical orbits in some direction around the critical point. Such a critical point we call a *center* or a *stable center*. It is not an asymptotically stable critical point, as the trajectories never approach the critical point, but at least if you start sufficiently close to the critical point, you stay close to

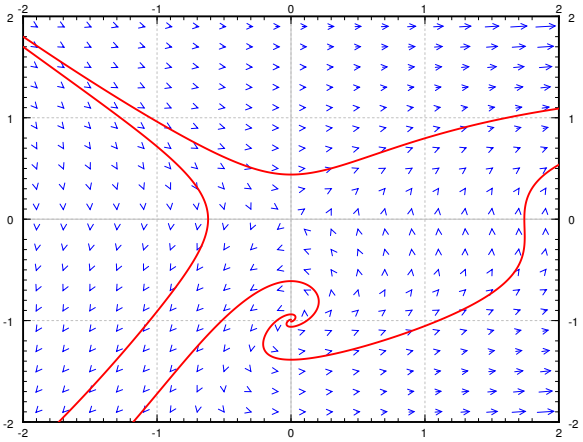


Figure 5.4: The phase portrait with few sample trajectories of $x' = y + y^2e^x$, $y' = x$.

the critical point. The simplest example of such behavior is the linear system with a center. Another example is the critical point $(0, 0)$ in Example 5.1.1 on page 370.

The trouble with a center in a nonlinear system is that whether the trajectory goes towards or away from the critical point is governed by the sign of the real part of the eigenvalues of the Jacobian matrix, and the Jacobian matrix in a nonlinear system changes from point to point. Since this real part is zero at the critical point itself, it can have either sign nearby, meaning the trajectory could be pulled towards or away from the critical point.

Example 5.1.5: Find and analyze the critical point(s) of $x' = y, y' = -x + y^3$.

Solution: The only critical point is the origin $(0, 0)$. The Jacobian matrix is

$$\begin{bmatrix} 0 & 1 \\ -1 & 3y^2 \end{bmatrix}.$$

At $(0, 0)$ the Jacobian matrix is $\left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$, which has eigenvalues $\pm i$. So the linearization has a center.

Using the quadratic equation, the eigenvalues of the Jacobian matrix at any point (x, y) are

$$\lambda = \frac{3}{2}y^2 \pm i\frac{\sqrt{4 - 9y^4}}{2}.$$

At any point where $y \neq 0$ (so at most points near the origin), the eigenvalues have a positive real part (y^2 can never be negative). This positive real part pulls the trajectory away from the origin. A sample trajectory for an initial condition near the origin is given in Figure 5.5 on the next page. ⌟

The same process could be carried out with the system $x' = y, y' = -x - y^3$. This one will also have a center as the linearization at the origin, but the non-linear system will have a spiral sink at the origin. The moral of the example is that further analysis is needed when the linearization has a center. The analysis will in general be more complicated than in the example above, and is more likely to involve case-by-case consideration. Such a complication

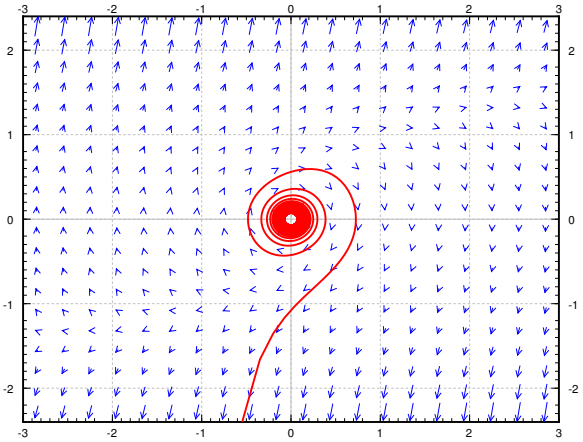


Figure 5.5: An unstable critical point (spiral source) at the origin for $x' = y, y' = -x + y^3$, even if the linearization has a center.

should not be surprising to you. By now in your mathematical career, you have seen many places where a simple test is inconclusive, recall for example the second derivative test for maxima or minima, and requires more careful, and perhaps ad hoc analysis of the situation.

5.1.6 Exercises

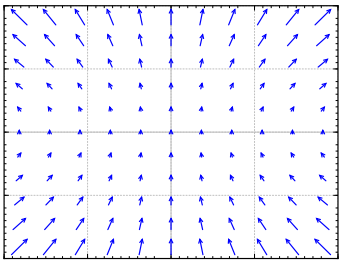
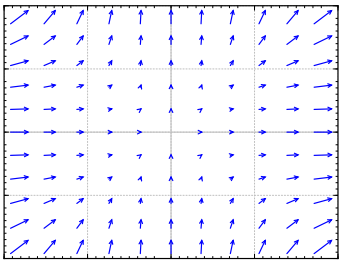
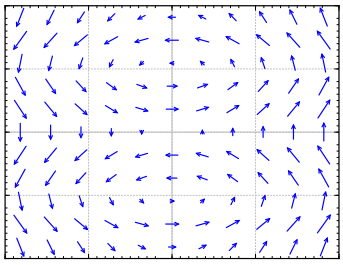
***Exercise* 5.1.1:** *Sketch the phase plane vector field for:*

a) $x' = x^2, \quad y' = y^2,$ b) $x' = (x-y)^2, \quad y' = -x,$ c) $x' = e^y, \quad y' = e^x.$

***Exercise* 5.1.2:** *Match systems*

$$(i)\ x' = x^2, \quad y' = y^2, \quad (ii)\ x' = xy, \quad y' = 1 + y^2, \quad (iii)\ x' = \sin(\pi y), \quad y' = x,$$

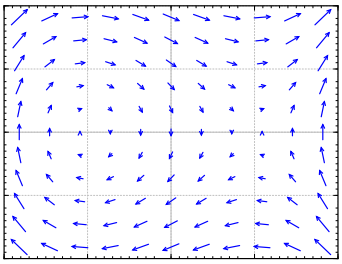
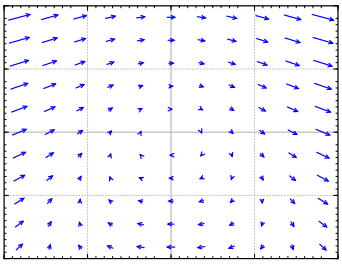
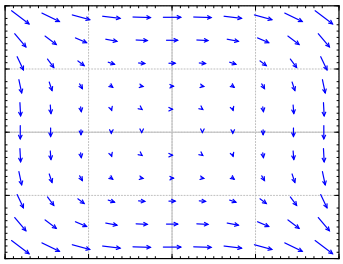
to the vector fields below. Justify.

a)  b)  c) 

Exercise* 5.1.3: *Match systems*

$$(i)\ x' = y^2, \quad y' = -x^2, \quad (ii)\ x' = y, \quad y' = (x-1)(x+1), \quad (iii)\ x' = y + x^2, \quad y' = -x,$$

to the vector fields below. Justify.

a)  b)  c) 

***Exercise* 5.1.4:** *Find the critical points and linearizations of the following systems.*

a) $x' = x^2 - y^2, \quad y' = x^2 + y^2 - 1,$ b) $x' = -y, \quad y' = 3x + yx^2,$

c) $x' = x^2 + y, \quad y' = y^2 + x.$

Exercise* 5.1.5: *Find the critical points and linearizations of the following systems.*

a) $x' = \sin(\pi y) + (x-1)^2, \quad y' = y^2 - y,$ b) $x' = x + y + y^2, \quad y' = x,$

c) $x' = (x-1)^2 + y, \quad y' = x^2 + y.$

***Exercise* 5.1.6:** *For the following systems, verify they have critical point at* $(0,0)$*, and find the linearization at* $(0,0)$*.*

a) $x' = x + 2y + x^2 - y^2, \quad y' = 2y - x^2$ b) $x' = -y, \quad y' = x - y^3$

c) $x' = ax + by + f(x,y)$, $y' = cx + dy + g(x,y)$, *where* $f(0,0) = 0$, $g(0,0) = 0$, *and all first partial derivatives of* f *and* g *are also zero at* $(0,0)$*, that is,* $\frac{\partial f}{\partial x}(0,0) = \frac{\partial f}{\partial y}(0,0) = \frac{\partial g}{\partial x}(0,0) = \frac{\partial g}{\partial y}(0,0) = 0.$

***Exercise* 5.1.7:** *Take the system* $x' = (x-2)(x+y), \quad y' = (y+3)(x-y).$

a) *Find all critical points.*

b) *Determine the linearization of this system around each of the critical points.*

c) *For each of the critical points, determine the behavior and classify the type of solution that the* linearized *system will have around that critical point.*

***Exercise* 5.1.8:** *Take the system* $x' = (x^2 - y)(x+3), \quad y' = (y-1)(x+y+1).$

a) *Find all critical points.*

b) *Determine the linearization of this system around each of the critical points.*

c) *For each of the critical points, determine the behavior and classify the type of solution that the* linearized *system will have around that critical point.*

***Exercise* 5.1.9:** *Take* $x' = (x-y)^2, \quad y' = (x+y)^2.$

a) *Find the set of critical points.*

b) *Sketch a phase diagram and describe the behavior near the critical point(s).*

c) *Find the linearization. Is it helpful in understanding the system?*

***Exercise* 5.1.10:** *Take* $x' = x^2$, $y' = x^3$.

a) *Find the set of critical points.*

b) *Sketch a phase diagram and describe the behavior near the critical point(s).*

c) *Find the linearization. Is it helpful in understanding the system?*

Exercise* 5.1.11: *The idea of critical points and linearization works in higher dimensions as well. You simply make the Jacobian matrix bigger by adding more functions and more variables. For the following system of 3 equations find the critical points and their linearizations:*

$$x' = x + z^2, \qquad y' = z^2 - y, \qquad z' = z + x^2.$$

Exercise* 5.1.12: *Any two-dimensional non-autonomous system* $x' = f(x, y, t)$, $y' = g(x, y, t)$ *can be written as a three-dimensional autonomous system (three equations). Write down this autonomous system using the variables* u, v, w.

***Exercise* 5.1.13:** *For the systems below, find and classify the critical points, also indicate if the equilibria are stable, asymptotically stable, or unstable.*

a) $x' = -x + 3x^2, y' = -y$ b) $x' = x^2 + y^2 - 1$, $y' = x$

c) $x' = ye^x$, $y' = y - x + y^2$

Exercise* 5.1.14: *For the systems below, find and classify the critical points.*

a) $x' = -x + x^2$, $y' = y$ b) $x' = y - y^2 - x$, $y' = -x$ c) $x' = xy$, $y' = x + y - 1$

***Exercise* 5.1.15:** *Find and classify all critical points of the system*

$$\frac{dx}{dt} = (x + 1)(x - y + 3) \qquad \frac{dy}{dt} = (x - 2)(x - y).$$

***Exercise* 5.1.16:** *Find and classify all critical points of the system*

$$\frac{dx}{dt} = x^2 - y^2 \qquad \frac{dy}{dt} = (x + 4)(y - 2).$$

***Exercise* 5.1.17:** *Find and classify the critical point(s) of* $x' = -x^2$, $y' = -y^2$.

***Exercise* 5.1.18:** *Suppose* $x' = -xy$, $y' = x^2 - 1 - y$.

a) *Show there are two spiral sinks at* $(-1, 0)$ *and* $(1, 0)$.

b) *For any initial point of the form* $(0, y_0)$, *find the trajectory.*

c) *Can a trajectory starting at* (x_0, y_0) *where* $x_0 > 0$ *spiral into the critical point at* $(-1, 0)$*? Why or why not?*

***Exercise* 5.1.19:** *In the example $x' = y$, $y' = y^3 - x$ show that for any trajectory, the distance from the origin is an increasing function. Conclude that the origin behaves like is a spiral source. Hint: Consider $f(t) = \big(x(t)\big)^2 + \big(y(t)\big)^2$ and show it has positive derivative.*

***Exercise* 5.1.20:** *Find and analyze all critical points of the system $x' = y$, $y' = -x - y^3$. Use the ideas from Exercise 5.1.19 to show that the solutions to this problem move towards the origin as t grows.*

Exercise* 5.1.21: *Derive an analogous classification of critical points for equations in one dimension, such as $x' = f(x)$ based on the derivative. A point x_0 is critical when $f(x_0) = 0$ and almost linear if in addition $f'(x_0) \neq 0$. Figure out if the critical point is stable or unstable depending on the sign of $f'(x_0)$. Explain. Hint: see § 1.7.*

5.2 Behavior of non-linear systems

Attribution: [JL], §8.2.

Learning Objectives

After this section, you will be able to:

- Find the trajectories for a non-linear system,
- Determine if a system is Hamiltonian and use that fact to find the general solution,
- Use nullclines to help analyze a non-linear system, and
- Identify basins of attraction and separatrices for a non-linear system.

5.2.1 Conservative equations

An equation of the form

$$x'' + f(x) = 0$$

for an arbitrary function $f(x)$ is called a *conservative equation*. For example the pendulum equation is a conservative equation. The equations are conservative as there is no friction in the system so the energy in the system is "conserved." Let us write this equation as a system of nonlinear ODE.

$$x' = y, \qquad y' = -f(x).$$

These types of equations have the advantage that we can solve for their trajectories easily.

Definition 5.2.1

Assume that we have an autonomous system of differential equations defining x and y,

$$x' = f(x, y) \qquad y' = g(x, y).$$

A *trajectory* for this system is a curve in the xy-plane that the solution curve $(x(t), y(t))$ will stay on for all t. This curve will generally be given with y as a function of x, or the level curve of some function $F(x, y)$.

For conservative equations, we want to first think of y as a function of x for a moment. Then use the chain rule

$$x'' = y' = \frac{dy}{dx}x' = y\frac{dy}{dx},$$

where the prime indicates a derivative with respect to t. We obtain $y\frac{dy}{dx} + f(x) = 0$. We integrate with respect to x to get $\int y\frac{dy}{dx}\,dx + \int f(x)\,dx = C$. In other words

$$\frac{1}{2}y^2 + \int f(x)\,dx = C.$$

We obtained an implicit equation for the trajectories, with different C giving different trajectories. The value of C is conserved on any trajectory. This expression is sometimes

called the *Hamiltonian* or the energy of the system. If you look back to § 1.9, you will notice that $y\frac{dy}{dx} + f(x) = 0$ is an exact equation, and we just found a potential function.

Another approach we could use in this case is separable equations, if it works out. The idea is that we have the system

$$x' = y, \qquad y' = -f(x)$$

and want to develop a differential equation for y in terms of x. We can write this differential equation using some principles from implicit differentiation and parametric equations as

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt},$$

which, for this case, is

$$\frac{dy}{dx} = \frac{-f(x)}{y}.$$

This equation is separable as

$$y\ dy = -f(x)\ dx$$

and we can get to the same implicit equation for the trajectories as before.

Example 5.2.1: Find the trajectories for the equation $x'' + x - x^2 = 0$, which is the equation from Example 5.1.1 on page 370.

Solution: The corresponding first order system is

$$x' = y, \qquad y' = -x + x^2.$$

Trajectories satisfy

$$\frac{1}{2}y^2 + \frac{1}{2}x^2 - \frac{1}{3}x^3 = C.$$

We solve for y

$$y = \pm\sqrt{-x^2 + \frac{2}{3}x^3 + 2C}.$$

Plotting these graphs we get exactly the trajectories in Figure 5.1 on page 371. In particular we notice that near the origin the trajectories are *closed curves*: they keep going around the origin, never spiraling in or out. Therefore we discovered a way to verify that the critical point at $(0, 0)$ is a stable center. The critical point at $(0, 1)$ is a saddle as we already noticed. This example is typical for conservative equations. □

Consider an arbitrary conservative equation $x'' + f(x) = 0$. All critical points occur when $y = 0$ (the x-axis), that is when $x' = 0$. The critical points are those points on the x-axis where $f(x) = 0$. The trajectories are given by

$$y = \pm\sqrt{-2\int f(x)\,dx + 2C}.$$

So all trajectories are mirrored across the x-axis. In particular, there can be no spiral sources nor sinks. The Jacobian matrix is

$$\begin{bmatrix} 0 & 1 \\ -f'(x) & 0 \end{bmatrix}.$$

The critical point is almost linear if $f'(x) \neq 0$ at the critical point. Let J denote the Jacobian matrix. The eigenvalues of J are solutions to

$$0 = \det(J - \lambda I) = \lambda^2 + f'(x).$$

Therefore $\lambda = \pm\sqrt{-f'(x)}$. In other words, either we get real eigenvalues of opposite signs (if $f'(x) < 0$), or we get purely imaginary eigenvalues (if $f'(x) > 0$). There are only two possibilities for critical points, either an *unstable saddle point*, or a *stable center*. There are never any sinks or sources.

5.2.2 Hamiltonian Systems

A generalization of conservative equations to systems is a Hamiltonian system. This type of system has all of the nice properties of conservative equations when converted into systems, but allows for more general interactions between x and y. For these systems, the point is that the equation has a conserved quantity called a Hamiltonian, which does not change as the system evolves in time, which generally represents the energy of the system. Calling this function $H(x, y)$, this means that

$$\frac{d}{dt}H(x, y) = 0.$$

By the chain rule, this is equivalent to

$$\frac{\partial H}{\partial x}\frac{dx}{dt} + \frac{\partial H}{\partial y}\frac{dy}{dt} = 0.$$

One way to satisfy this is with

$$\begin{aligned} \frac{dx}{dt} &= -\frac{\partial H}{\partial y} \\ \frac{dy}{dt} &= \frac{\partial H}{\partial x} \end{aligned} \tag{5.2}$$

and this gives the definition of a *Hamiltonian system*.

Definition 5.2.2

The system

$$\begin{aligned} \frac{dx}{dt} &= f(x, y) \\ \frac{dy}{dt} &= g(x, y) \end{aligned} \tag{5.3}$$

is *Hamiltonian* if there is a function $H(x, y)$ so that $f(x, y) = -\frac{\partial H}{\partial y}$ and $g(x, y) = \frac{\partial H}{\partial x}$.

For solving these sorts of systems, we know that

$$\frac{d}{dt}H(x,y) = 0,$$

since that's how we defined the system. This means that the trajectories of this system are given by

$$H(x,y) = C$$

for a constant C determined by initial conditions. So if we can find the function H that expresses the system in the form (5.2), then we are done.

Finding this H is a lot similar to finding solutions to exact equations in § 1.9. First, we need to determine if the system is Hamiltonian. Since we want to have that

$$f(x,y) = -\frac{\partial H}{\partial y} \qquad g(x,y) = \frac{\partial H}{\partial x}$$

we know that

$$f_x(x,y) = -\frac{\partial^2 H}{\partial x \partial y} \qquad g_y(x,y) = \frac{\partial^2 H}{\partial x \partial y}$$

which shows that

$$f_x + g_y = 0.$$

This is what we can use to check if a system is Hamiltonian; compare to Theorem 1.9.1 for exact equations.

Once we know that a system is Hamiltonian, we can integrate the different components of the equation to find the function H. Since $f = -\frac{\partial H}{\partial y}$, then we can write

$$H(x,y) = -\int f(x,y)\, dy + A(x)$$

where $A(x)$ is an unknown function, which can be determined by differentiating this in x and setting equal to $g(x,y)$.

Example 5.2.2: Consider the system of differential equations given by

$$x' = -4x + 3y \qquad y' = 2x + 4y.$$

Determine if this system is Hamiltonian and, if so, find the trajectories of the solution.

Solution: We first check if $f_x + g_y = 0$ to see if the system is Hamiltonian. Since $f_x = -4$ and $g_y = 4$, this means we have a Hamiltonian system. In order to find the function H, we use that

$$\frac{\partial H}{\partial y} = -f(x,y) = 4x - 3y.$$

Integrating both sides in y gives that

$$H(x,y) = 4xy - \frac{3}{2}y^2 + A(x)$$

for an unknown function $A(x)$. Differentiating this in x gives

$$\frac{\partial H}{\partial x} = 4y + A'(x)$$

which we want to equal $2x + 4y$. This gives that $A'(x) = 2x$ so $A(x) = x^2$. Thus, the Hamiltonian is given by

$$H(x, y) = x^2 + 4xy - \frac{3}{2}y^2$$

so that the trajectories are defined by

$$x^2 + 4xy - \frac{3}{2}y^2 = C$$

for any constant C. These are sketched in Figure 5.6.

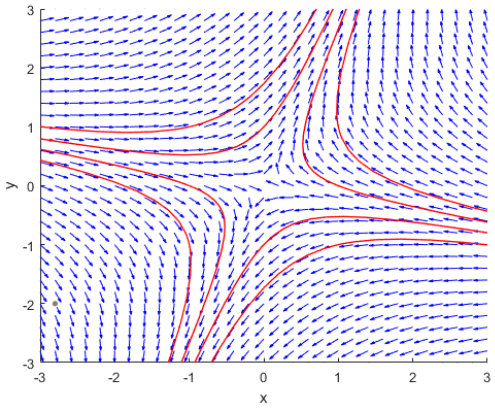


Figure 5.6: Vector field and trajectories for a Hamiltonian System

Note that this system is linear and autonomous. Therefore, we could have solved this using those methods as well. For this, we have the coefficient matrix

$$A = \begin{bmatrix} -4 & 3 \\ 2 & 4 \end{bmatrix}$$

and we can find the eigenvalues of this matrix as the roots of

$$\det(A - \lambda I) = (-4 - \lambda)(4 - \lambda) - (3)(2) = \lambda^2 - 22$$

whose roots are $\pm\sqrt{22}$ which have opposite signs. Therefore, this will be a saddle point, which we see represented in the plot in Figure 5.6.

5.2.3 Separatrices and Basins of Attractions

If we have an asymptotically stable critical point (x_0, y_0) for an autonomous system of differential equation, we know that solutions that start "near" this point will converge to it as $t \to \infty$. That's what it means for the point to be asymptotically stable. However, for applications, it may be important to know exactly which initial conditions $(x(0), y(0))$ will end up converging to (x_0, y_0). This can be particularly relevant when there are multiple asymptotically stable equilibrium solutions and we need to determine which one a given initial condition will converge to.

Definition 5.2.3

Let (x_0, y_0) be an asymptotically stable equilibrium solution for the autonomous system $\vec{x}\,' = \vec{F}(x, y)$. The *basin of attraction* for (x_0, y_0) is the set of all points (a, b) where the solution to

$$\vec{x}' = \vec{F}(x, y) \qquad \vec{x}(0) = \begin{bmatrix} a \\ b \end{bmatrix}$$

converges to (x_0, y_0) as $t \to \infty$.

In general, the basin of attraction for an asymptotically stable critical point is difficult, if not impossible, to find analytically. The main approach here is to use a direction field to approximate the basin of attraction.

Example 5.2.3: Find all asymptotically stable critical points for the autonomous system

$$\frac{dx}{dt} = x(7 - 2x - 5y) \qquad \frac{dy}{dt} = (y + 1)(5 - 3x - 2y)$$

and determine an approximate basin of attraction for each.

Solution: We want to start by finding the critical points for this system, classifying them to determine if they are asymptotically stable, and then use a slope field to try to find the basin of attraction. In order to have a critical point, we need to have both $\frac{dx}{dt}$ and $\frac{dy}{dt}$ equal to zero. This means that we need

$$[x = 0 \text{ or } 7 - 2x - 5y = 0] \quad \text{and} \quad [y = -1 \text{ or } 5 - 3x - 2y = 0]\,.$$

This results in the points $(0, -1)$, $(0, 5/2)$, $(6, -1)$, and $(1, 1)$. In order to classify each of these critical points, we need to find the Jacobian matrix for this system, which is

$$J(x, y) = \begin{bmatrix} F_x & F_y \\ G_x & G_y \end{bmatrix} = \begin{bmatrix} 7 - 4x - 5y & -5x \\ -3(y + 1) & 3 - 3x - 4y \end{bmatrix},$$

and then we want to plug each critical point into this matrix in turn.

Plugging in $(0, -1)$ gives $\left[\begin{smallmatrix} 12 & 0 \\ 0 & 7 \end{smallmatrix}\right]$, which has eigenvalues of 12 and 7, and so is a nodal source, which is unstable. Plugging in $(0, 5/2)$ gives $\left[\begin{smallmatrix} -11/2 & 0 \\ -21/2 & -7 \end{smallmatrix}\right]$, which has eigenvalues of $-11/2$ and -7, which is a nodal sink, and so asymptotically stable. Plugging in $(6, -1)$ gives $\left[\begin{smallmatrix} -12 & 30 \\ 0 & -11 \end{smallmatrix}\right]$, which has eigenvalues at -11 and -12, giving an asymptotically stable nodal sink.

The last point at $(1, 1)$ gives $\left[\begin{smallmatrix} -2 & -5 \\ -6 & -4 \end{smallmatrix}\right]$, which is not triangular, and so does not have easily identifiable eigenvalues. We could use trace-determinant analysis to classify (§ 4.7) or we can just compute the eigenvalues. Those are found by the roots of

$$\det(A - \lambda I) = (-2 - \lambda)(-4 - \lambda) - (-5)(-2) = \lambda^2 + 6\lambda - 2$$

and since the last term is negative, we know we are going to get roots of opposite signs, so this is an unstable saddle point. The actual eigenvalues are

$$\frac{-6 \pm \sqrt{36 - 4(1)(-2)}}{2} = -3 \pm \frac{\sqrt{44}}{2} = -3 \pm \sqrt{11}.$$

So, this means we have two asymptotically stable critical points, $(0, 5/2)$ and $(6, -1)$. We need to look at a slope field to determine the approximate basin of attraction for each of these points.

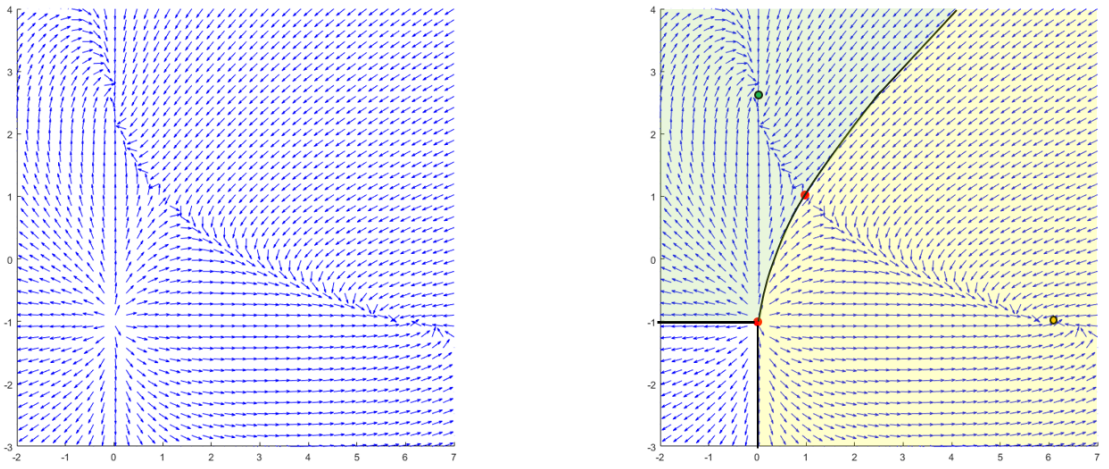


Figure 5.7: Plots showing the slope field for Example 5.2.3 (left) and a slope field combined with approximate basins of attraction (right).

From Figure 5.7, we can see that there is a sort of dividing line between the two nodal sinks. If the solution starts on one side of the line, it funnels into one critical point, and on the other side, it heads to the other one. This dividing line also seems to pass through the saddle point at $(1, 1)$, which is not a coincidence, as we will see later.

Another interesting feature of these regions is the boundary of them. This is a curve that separates solutions that converge to the asymptotically stable equilibrium solution and those that don't. This leads to another definition.

Definition 5.2.4

Consider the autonomous system $\vec{x}' = \vec{F}(x, y)$. A *separatrix* (plural separatrices) is a curve in the plane that separates trajectories that have different long-term behaviors of solutions to $\vec{x}\,' = \vec{F}(x, y)$.

The boundary of a basin of attraction is a separatrix because the long-term behavior inside the curve (converging to the asymptotically stable critical point) is different from the behavior outside the curve (going somewhere else). These dividing curves also show up in other contexts.

Example 5.2.4: Analyze the system

$$\frac{dx}{dt} = -x - y$$
$$\frac{dy}{dt} = -2x$$

in the context of separatrices.

Solution: This is a linear, homogeneous system, so we can analyze it via that approach. For the coefficient matrix $A = \begin{bmatrix} -1 & -1 \\ -2 & 0 \end{bmatrix}$, we have eigenvalues as the roots of $(-1-\lambda)(-\lambda) - 2$, or $\lambda^2 + \lambda - 2$. Therefore, the eigenvalues here are 1 and -2. For 1 we have eigenvector $\begin{bmatrix} 1 \\ -2 \end{bmatrix}$ and for -2, an eigenvector is $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. We can see what this looks like on a slope field in Figure 5.8.

There are no asymptotically stable critical points here, so there are no basins of attraction. However, there are two distinct behaviors of the solution curve. It is going away from the origin, but it could go to the top left, or to the bottom right. Both of those make sense based on the slope field here. So how do we know which way it goes? The line drawn on the right side of Figure 5.8 seems to divide these two regions up. If the solution starts above the line, it goes to the top left, otherwise, it goes to the bottom right. This is the separatrix for this saddle point.

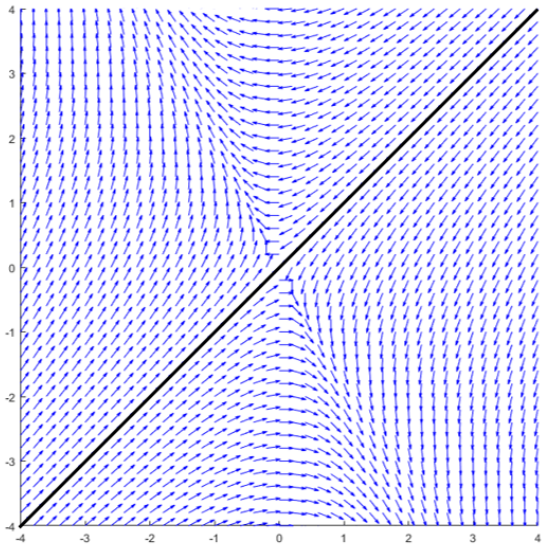


Figure 5.8: Plot showing the slope field for Example 5.2.4 with a sketch of the separatrix.

But what is that line? If we inspect the graph more closely, the separatrix here is the straight-line solution that converges to zero over time; the one particular solution that does not go off to infinity because it only has the e^{-2t} term in it. So the straight-line solutions that flow into saddle points divide what happens on the two sides of it. This is a very common fact in looking at separatrices: if they go through a critical point, they generally do so as the in-flowing solution from a saddle point. All of the examples we have seen so far with separatrices have done exactly this. ⌟

5.2.4 Nullclines

When trying to find critical points for a non-linear, autonomous system, we need all (both, in the case of two component systems) of the equations to be zero. What happens if only one of the equations is zero? This is a lot easier to find, and can also give us a fair bit of information.

Definition 5.2.5

Consider the autonomous two-component system

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

A *nullcline* for this system is a curve where either $f(x, y) = 0$ or $g(x, y) = 0$. We can also be more specific and use the term *x-nullcline* for the curve(s) where $\frac{dx}{dt} = 0$ and *y-nullcline* for where $\frac{dy}{dt} = 0$.

The way we can use these nullclines is to know in general which direction the solution curve will move in different regions of the plane. Assuming that all of the functions involved are continuous, if we know that the solution at a given point will move to the right, that is, if $\frac{dx}{dt} > 0$, then we know that the solution will continue to move to the right until we cross an x-nullcline. If the solution starts going back to the left, this means that $\frac{dx}{dt}$ becomes negative, and so must cross zero, which is where a nullcline is.

In addition, we know that along an x-nullcline, $\frac{dx}{dt} = 0$, so the solution can only be moving in the y direction, that is, vertically. If we can determine in which direction the solution graph will cross the nullclines, this can also be helpful and useful. It doesn't give as much information as a full slope field or trajectory plot, but it can give a general idea of what is going to happen to the solution over time.

Example 5.2.5: Use a nullcline analysis to determine the overall behavior of solutions to the system

$$\frac{dx}{dt} = (y - 1)(x - y) \qquad \frac{dy}{dt} = (x + 3)(x - 2).$$

Solution: We can get the equations for nullclines from the factors of each of the differential equations here. For x-nullclines, we get $y = 1$ and $y = x$, and for y-nullclines, we get $x = -3$ and $x = 2$. Once we have these lines, we need to determine what happens within each of the regions on the resulting graph. For example, if we look in the region above $y = 1$, above $y = x$ and right of $x = 2$, we can plug in, for example $(3, 4)$. At this point $\frac{dx}{dt} = (4 - 1)(3 - 4) = -3 < 0$ and $\frac{dy}{dt} = (3 + 3)(3 - 2) = 6 > 0$. Therefore, the solution here moves up and to the left. We can fill in all of the other regions in a similar manner. This is shown on the left of Figure 5.9.

Based on the nullcline diagram here, we can see that there seems to be some sort of spiraling behavior around both $(2, 2)$ and $(-3, -3)$, which we know are critical points because the two different nullclines intersect there. From this alone, we can't really tell if they are sources, sinks, or centers, but we do get a general idea of the behavior. We also see what looks like saddles at $(-3, 1)$ and $(2, 1)$, since these critical points have two opposite arrows pointing

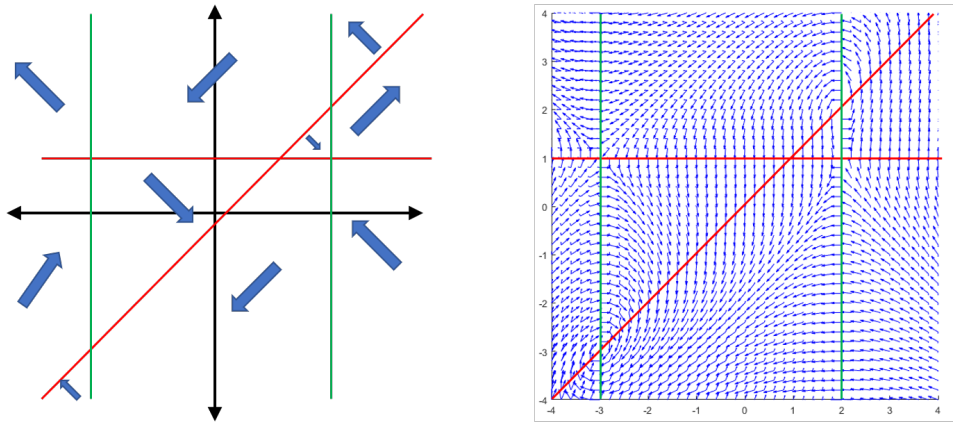


Figure 5.9: Plots showing the nullcline diagram for Example 5.2.5 (left) and a slope field for the same differential equation (right).

towards this point (corresponding to the negative eigenvalue of the linearized system), and two opposite arrows pointing away from the point (corresponding to the positive eigenvalue. The slope field seems to validate all of these general discussions from the nullcline diagram.

Example 5.2.6: Use a nullcline analysis to determine the overall behavior of solutions to the system

$$\frac{dx}{dt} = (x-3)(y+1) \qquad \frac{dy}{dt} = (y-2)(x+y).$$

Solution: As with the previous example, we can find the nullcline equations from the factors above. The x-nullclines are at $x = 3$ and $y = -1$, and the y-nullclines are at $y = 2$ and $x = -y$. We can plug in points to fill in the nullcline diagram like before, and compare to the slope field, shown in Figure 5.10.

In this diagram, we see some of the other types of critical points and what they look like through nullclines. The point at $(3, 2)$ looks like a nodal source, because all of the arrows point away from that point, and $(3, -3)$ looks like a nodal source. Finally, we see a potential saddle at $(1, -1)$ because of the patterns of the arrows, all of which are also shown in the slope field for this system.

An extra point with this type of result is that we know we can not cross the nullclines at $x = 3$ and $y = 2$. For instance, the line $x = 3$ is an x-nullcline. This means that the solution must cross the line moving vertically. However, it is a vertical line, and there is no way to cross a vertical line moving vertically. The same argument applies to $y = 2$. We can also see this by the fact that the arrows on either side of the line both point into or away from these nullclines.

5.2.5 Exercises

***Exercise* 5.2.1:** *Find the implicit equations of the trajectories of the following conservative systems. Next find their critical points (if any) and classify them.*

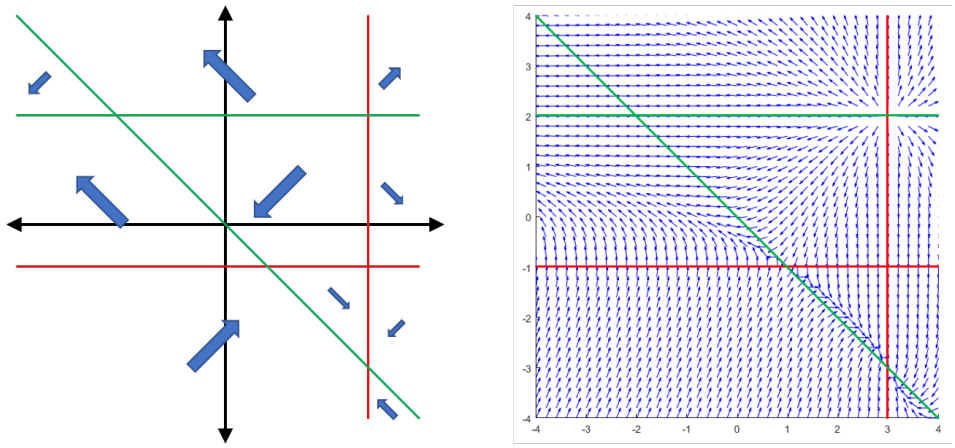


Figure 5.10: Plots showing the nullcline diagram for Example 5.2.6 (left) and a slope field for the same differential equation (right).

a) $x'' + x + x^3 = 0$ b) $\theta'' + \sin\theta = 0$

c) $z'' + (z-1)(z+1) = 0$ d) $x'' + x^2 + 1 = 0$

Exercise* 5.2.2: *Find the implicit equations of the trajectories of the following conservative systems. Next find their critical points (if any) and classify them.*

a) $x'' + x^2 = 4$ b) $x'' + e^x = 0$ c) $x'' + (x+1)e^x = 0$

Exercise* 5.2.3: *The conservative system $x'' + x^3 = 0$ is not almost linear. Classify its critical point(s) nonetheless.*

***Exercise* 5.2.4:** *Determine if the following system is Hamiltonian. If it is, find the general solution in the form $H(x, y) = C$ and sketch some of the trajectories.*

$$\frac{dx}{dt} = x - 2y \qquad \frac{dy}{dt} = 3x - y.$$

***Exercise* 5.2.5:** *Determine if the following system is Hamiltonian. If it is, find the general solution in the form $H(x, y) = C$ and sketch some of the trajectories.*

$$\frac{dx}{dt} = 4x - 2y + 2 \qquad \frac{dy}{dt} = -5x + y - 1.$$

***Exercise* 5.2.6:** *Determine if the following system is Hamiltonian. If it is, find the general solution in the form $H(x, y) = C$ and sketch some of the trajectories.*

$$\frac{dx}{dt} = x^2 - 2xy + 3y^2 \qquad \frac{dy}{dt} = y^2 - 2xy + e^x.$$

***Exercise* 5.2.7:** *Determine if the following system is Hamiltonian. If it is, find the general solution in the form $H(x, y) = C$ and sketch some of the trajectories.*

$$\frac{dx}{dt} = 3x - 2xy \qquad \frac{dy}{dt} = 2xy - 3y.$$

***Exercise* 5.2.8:** *Consider a generic thing on a spring, with displacement u and velocity v. Assume that*

$$mu'' + ku = 0,$$

where m and k are some positive constants.

a) Rewrite this equation as a first-order system in u and v.

b) Find a Hamiltonian function for this system (in terms of u and v).

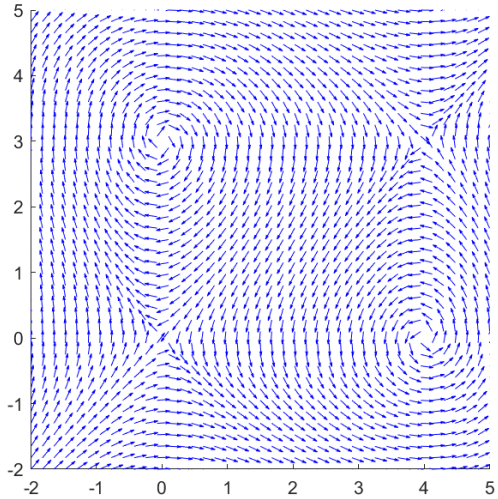
c) What shapes are the level curves of the Hamiltonian function?

d) Does this system have a basin of attraction? Explain briefly.

***Exercise* 5.2.9:** *Suppose f is always positive. Find the trajectories of $x'' + f(x') = 0$. Are there any critical points?*

***Exercise* 5.2.10:** *Suppose that $x' = f(x, y)$, $y' = g(x, y)$. Suppose that $g(x, y) > 1$ for all x and y. Are there any critical points? What can we say about the trajectories at t goes to infinity?*

***Exercise* 5.2.11:** *Here is the direction field for the system $\frac{dx}{dt} = y^2 - 3y, \frac{dy}{dt} = x^2 - 4x$. The critical points are $(0, 0), (4, 0), (0, 3)$, and $(4, 3)$. Draw the nullclines on the plot. What do the nullclines tell us about the critical points?*



***Exercise* 5.2.12:** *Nullclines apply to linear systems as well, although since we can often solve those explicitly they're less necessary. Construct the nullcline diagram for the system* $\frac{dx}{dt} = -3x + y$, $\frac{dy}{dt} = 6x + 2y$, *and use it to classify (by type) the equilibrium point at the origin. What is the linearization of this system at* $(0, 0)$*?*

***Exercise* 5.2.13:** *Consider the system* $\frac{dx}{dt} = -2x + y, \frac{dy}{dt} = -y + x^2$.

a) *Find all equilibrium solutions.*

b) *Sketch all nullclines for this system on a single diagram. Label each region, and use these results to classify each equilibrium point.*

***Exercise* 5.2.14:** *Nullclines need not be lines. Consider the system*

$$\begin{bmatrix} \dfrac{dx}{dt} = 4 - y^2 \\ \dfrac{dy}{dt} = 8 - x^2 - y^2 \end{bmatrix}.$$

a) *Find all critical points of this system.*

b) *Sketch the nullcline diagram and label all regions DL, DR, UL, or UR. Classify (according to type) any critical point(s) that can be classified using this analysis.*

c) *Two critical points cannot be classified using the nullcline analysis. Classify these (again according to type) using the Jacobian.*

***Exercise* 5.2.15:** *Consider the system*

$$\begin{bmatrix} \frac{dx}{dt} = x - y^2 + 2 \\ \frac{dy}{dt} = x^2 - y^2 \end{bmatrix} \tag{5.4}$$

a) *Find all critical points of (5.4).*

b) *Create the nullcline diagram for the system, labelling each region as one of UL, UR, DL, or DR. Use this information to classify two critical points according to type.*

c) *Use the Jacobian matrix to classify any remaining critical points.*

d) *Is there a conserved quantity (Hamiltonian function) for this system? If so, find one. If not, explain why not.*

***Exercise* 5.2.16:** *For a conflict between two armies, Lanchester's Law asserts that* $\frac{dx}{dt} = -\alpha y$ *and* $\frac{dy}{dt} = -\beta x$, *where* x *and* y *are the two populations, and* α *and* β *are some positive constants.*

a) *Find a Hamiltonian function for this system satisfying* $H(0,0) = 0$.

b) *Classify the critical point at the origin according to type* **and** *stability.*

c) *Assume that we are just looking at the first quadrant, since the populations are non-negative. Find the curve along which the Hamiltonian function is zero, and explain its significance in terms of who wins the conflict.*

***Exercise* 5.2.17:** *Consider the non-linear system*

$$\frac{dx}{dt} = 4x - 3y - x(x^2 + y^2), \quad \frac{dy}{dt} = 3x + 4y - y(x^2 + y^2). \tag{5.5}$$

a) (5.5) *has a critical point at the origin. What is the linearization of* (5.5) *at the origin?*

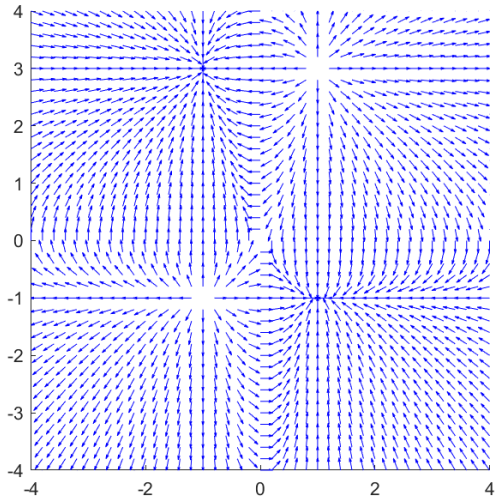
b) *Demonstrate that* (5.5) *is locally linear in a neighborhood of the origin.*

c) *Classify the origin according to its type and stability.*

***Exercise* 5.2.18:** *Consider the system of differential equations*

$$\frac{dx}{dt} = (x^2 - 1)y \qquad \frac{dy}{dt} = (y - 3)(y - 1)x \tag{5.6}$$

which has slope field sketched below.



a) *Find and classify all critical points of the system* (5.6).

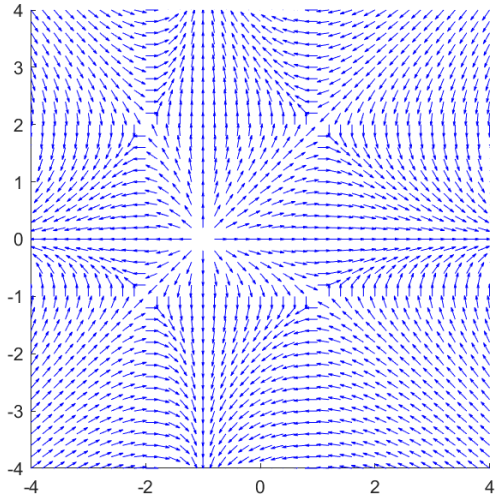
b) *Draw any separatrices that you can spot on the slope field.*

c) *Do any of these critical points have a basin of attraction? If so, sketch out what regions of the plane correspond to a basin of attraction for those critical points.*

***Exercise* 5.2.19:** *Consider the system of differential equations*

$$\frac{dx}{dt} = (2-y)(y+1)(x+1) \qquad \frac{dy}{dt} = -(x+2)(x-1)y \tag{5.7}$$

which has slope field sketched below.



a) *Find and classify all critical points of the system* (5.7).

b) *Draw any separatrices that you can spot on the slope field.*

c) *Do any of these critical points have a basin of attraction? If so, sketch out what regions of the plane correspond to a basin of attraction for those critical points.*

5.3 Applications of nonlinear systems

Attribution: [JL], §8.3.

Learning Objectives

After this section, you will be able to:

- Use non-linear systems to model the motion of a pendulum, and
- Use non-linear systems to model population dynamics like predator-prey and competing species models.

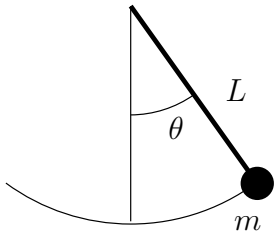
In this section we study two very standard examples of nonlinear systems. First, we look at the nonlinear pendulum equation. We saw the pendulum equation's linearization before, but we noted it was only valid for small angles and short times. Now we find out what happens for large angles. Next, we look at the predator-prey equation, which finds various applications in modeling problems in biology, chemistry, economics, and elsewhere.

5.3.1 Pendulum

The first example we study is the pendulum equation $\theta'' + \frac{g}{L}\sin\theta = 0$. Here, θ is the angular displacement, g is the gravitational acceleration, and L is the length of the pendulum. In this equation we disregard friction, so we are talking about an idealized pendulum.

This equation is a conservative equation, so we can use our analysis of conservative equations from the previous section. Let us change the equation to a two-dimensional system in variables (θ, ω) by introducing the new variable ω:

$$\begin{bmatrix}\theta\\ \omega\end{bmatrix}' = \begin{bmatrix}\omega\\ -\frac{g}{L}\sin\theta\end{bmatrix}.$$



The critical points of this system are when $\omega = 0$ and $-\frac{g}{L}\sin\theta = 0$, or in other words if $\sin\theta = 0$. So the critical points are when $\omega = 0$ and θ is a multiple of π. That is, the points are $\ldots(-2\pi, 0), (-\pi, 0), (0, 0), (\pi, 0), (2\pi, 0)\ldots$. While there are infinitely many critical points, they are all isolated. Let us compute the Jacobian matrix:

$$\begin{bmatrix}\frac{\partial}{\partial\theta}\Big(\omega\Big) & \frac{\partial}{\partial\omega}\Big(\omega\Big)\\ \frac{\partial}{\partial\theta}\Big(-\frac{g}{L}\sin\theta\Big) & \frac{\partial}{\partial\omega}\Big(-\frac{g}{L}\sin\theta\Big)\end{bmatrix} = \begin{bmatrix}0 & 1\\ -\frac{g}{L}\cos\theta & 0\end{bmatrix}.$$

For conservative equations, there are two types of critical points. Either stable centers, or saddle points. The eigenvalues of the Jacobian matrix are $\lambda = \pm\sqrt{-\frac{g}{L}\cos\theta}$.

The eigenvalues are going to be real when $\cos\theta < 0$. This happens at the odd multiples of π. The eigenvalues are going to be purely imaginary when $\cos\theta > 0$. This happens at the even multiples of π. Therefore the system has a stable center at the points $\ldots(-2\pi, 0), (0, 0), (2\pi, 0)\ldots$, and it has an unstable saddle at the points $\ldots(-3\pi, 0), (-\pi, 0), (\pi, 0), (3\pi, 0)\ldots$ Look at the phase diagram in Figure 5.11 on the facing page, where for simplicity we let $\frac{g}{L} = 1$.

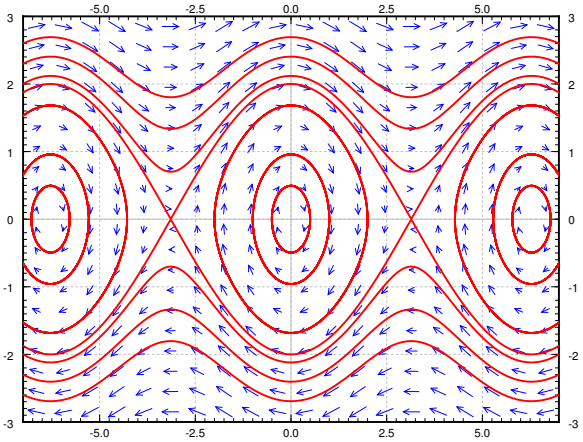


Figure 5.11: Phase plane diagram and some trajectories of the nonlinear pendulum equation.

Since this is a pendulum without friction, we can characterize the two differnet types of trajectories here. There are the curves running along the top and bottom of the phase portrait that look somewhat like sine waves. These graphs never cross the x-axis, which is the line $\omega = 0$. Therefore, these are trajectories where the pendulum never stops moving; it just keeps spinning around in full circles forever, crossing through all possible values of θ. The other type of trajectory are the ellipses around each of the stable equilibrium solutions. In these cases, the graph only spans a specific range of θ values, represented by the reduced x range of the ellipse, and cycles there forever. This represents a pendulum that does not have enough energy to make a full circle, and just oscillates back-and-forth to a fixed height forever.

In the linearized equation we have only a single critical point, the center at $(0, 0)$. Now we see more clearly what we meant when we said the linearization is good for small angles. The horizontal axis is the deflection angle. The vertical axis is the angular velocity of the pendulum. Suppose we start at $\theta = 0$ (no deflection), and we start with a small angular velocity ω. Then the trajectory keeps going around the critical point $(0, 0)$ in an approximate circle. This corresponds to short swings of the pendulum back and forth. When θ stays small, the trajectories really look like circles and hence are very close to our linearization.

When we give the pendulum a big enough push, it goes across the top and keeps spinning about its axis. This behavior corresponds to the wavy curves that do not cross the horizontal axis in the phase diagram. Let us suppose we look at the top curves, when the angular velocity ω is large and positive. Then the pendulum is going around and around its axis. The velocity is going to be large when the pendulum is near the bottom, and the velocity is the smallest when the pendulum is close to the top of its loop.

At each critical point, there is an equilibrium solution. Consider the solution $\theta = 0$; the pendulum is not moving and is hanging straight down. This is a stable place for the pendulum to be, hence this is a *stable* equilibrium.

The other type of equilibrium solution is at the unstable point, for example $\theta = \pi$. Here the pendulum is upside down. Sure you can balance the pendulum this way and it will stay,

but this is an *unstable* equilibrium. Even the tiniest push will make the pendulum start swinging wildly.

See Figure 5.12 for a diagram. The first picture is the stable equilibrium $\theta = 0$. The second picture corresponds to those "almost circles" in the phase diagram around $\theta = 0$ when the angular velocity is small. The next picture is the unstable equilibrium $\theta = \pi$. The last picture corresponds to the wavy lines for large angular velocities.

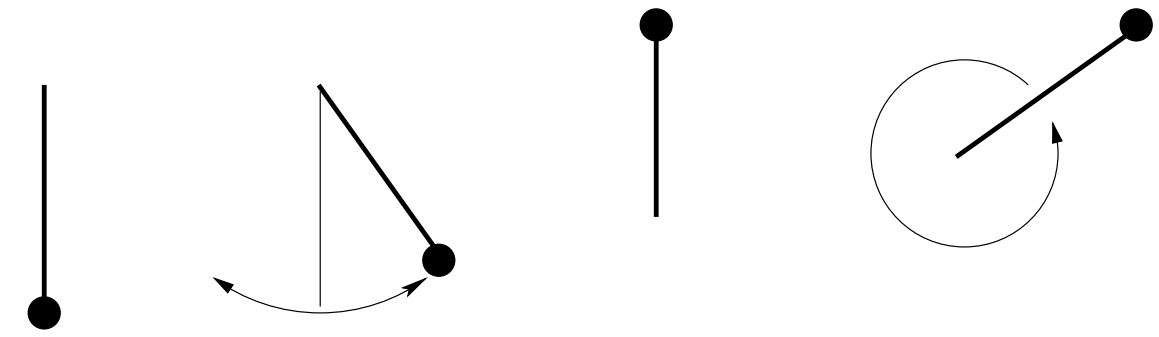


Figure 5.12: Various possibilities for the motion of the pendulum.

The quantity

$$\frac{1}{2}\omega^2 - \frac{g}{L}\cos\theta$$

is conserved by any solution. This is the energy or the Hamiltonian of the system.

We have a conservative equation and so (exercise) the trajectories are given by

$$\omega = \pm\sqrt{\frac{2g}{L}\cos\theta + C},$$

for various values of C. Let us look at the initial condition of $(\theta_0, 0)$, that is, we take the pendulum to angle θ_0, and just let it go (initial angular velocity 0). We plug the initial conditions into the above and solve for C to obtain

$$C = -\frac{2g}{L}\cos\theta_0.$$

Thus the expression for the trajectory is

$$\omega = \pm\sqrt{\frac{2g}{L}}\sqrt{\cos\theta - \cos\theta_0}.$$

Let us figure out the period. That is, the time it takes for the pendulum to swing back and forth. We notice that the trajectory about the origin in the phase plane is symmetric about both the θ and the ω-axis. That is, in terms of θ, the time it takes from θ_0 to $-\theta_0$ is the same as it takes from $-\theta_0$ back to θ_0. Furthermore, the time it takes from $-\theta_0$ to 0 is

the same as to go from 0 to θ_0. Therefore, let us find how long it takes for the pendulum to go from angle 0 to angle θ_0, which is a quarter of the full oscillation and then multiply by 4.

We figure out this time by finding $\frac{dt}{d\theta}$ and integrating from 0 to θ_0. The period is four times this integral. Let us stay in the region where ω is positive. Since $\omega = \frac{d\theta}{dt}$, inverting we get

$$\frac{dt}{d\theta} = \sqrt{\frac{L}{2g}} \frac{1}{\sqrt{\cos\theta - \cos\theta_0}}.$$

Therefore the period T is given by

$$T = 4\sqrt{\frac{L}{2g}} \int_0^{\theta_0} \frac{1}{\sqrt{\cos\theta - \cos\theta_0}}\, d\theta.$$

The integral is an improper integral, and we cannot in general evaluate it symbolically. We must resort to numerical approximation if we want to compute a particular T.

Recall from § 2.4, the linearized equation $\theta'' + \frac{g}{L}\theta = 0$ has period

$$T_{\text{linear}} = 2\pi\sqrt{\frac{L}{g}}.$$

We plot T, T_{linear}, and the relative error $\frac{T - T_{\text{linear}}}{T}$ in Figure 5.13. The relative error says how far is our approximation from the real period percentage-wise. Note that T_{linear} is simply a constant, it does not change with the initial angle θ_0. The actual period T gets larger and larger as θ_0 gets larger. Notice how the relative error is small when θ_0 is small. It is still only 15% when $\theta_0 = \frac{\pi}{2}$, that is, a 90 degree angle. The error is 3.8% when starting at $\frac{\pi}{4}$, a 45 degree angle. At a 5 degree initial angle, the error is only 0.048%.

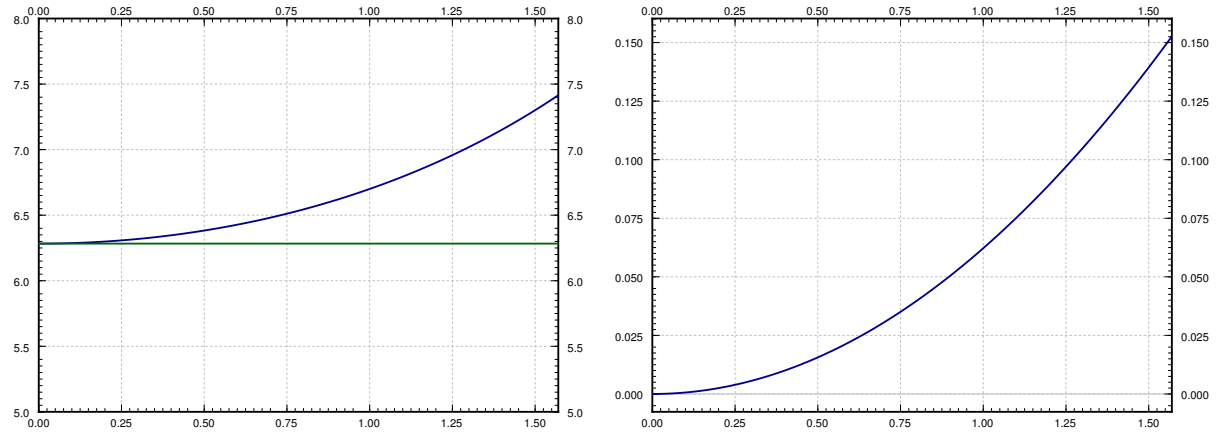


Figure 5.13: The plot of T and T_{linear} with $\frac{g}{L} = 1$ (left), and the plot of the relative error $\frac{T - T_{linear}}{T}$ (right), for θ_0 between 0 and $\pi/2$.

While it is not immediately obvious from the formula, it is true that

$$\lim_{\theta_0 \uparrow \pi} T = \infty.$$

That is, the period goes to infinity as the initial angle approaches the unstable equilibrium point. So if we put the pendulum almost upside down it may take a very long time before it gets down. This is consistent with the limiting behavior, where the exactly upside down pendulum never makes an oscillation, so we could think of that as infinite period.

5.3.2 Predator-prey or Lotka–Volterra systems

One of the most common simple applications of nonlinear systems are the so-called *predator-prey* or *Lotka–Volterra*[*] systems. For example, these systems arise when two species interact, one as the prey and one as the predator. It is then no surprise that the equations also see applications in economics. The system also arises in chemical reactions. In biology, this system of equations explains the natural periodic variations of populations of different species in nature. Before the application of differential equations, these periodic variations in the population baffled biologists.

We keep with the classical example of hares and foxes in a forest, it is the easiest to understand.

$$x = \#\text{ of hares (the prey)},$$
$$y = \#\text{ of foxes (the predator)}.$$

When there are a lot of hares, there is plenty of food for the foxes, so the fox population grows. However, when the fox population grows, the foxes eat more hares, so when there are lots of foxes, the hare population should go down, and vice versa. The Lotka–Volterra model proposes that this behavior is described by the system of equations

$$x' = (a - by)x,$$
$$y' = (cx - d)y,$$

where a, b, c, d are some parameters that describe the interaction of the foxes and hares[†]. In this model, these are all positive numbers.

Let us analyze the idea behind this model. The model is a slightly more complicated idea based on the exponential population model. First expand,

$$x' = (a - by)x = ax - byx.$$

The hares are expected to simply grow exponentially in the absence of foxes, that is where the ax term comes in, the growth in population is proportional to the population itself. We are assuming the hares always find enough food and have enough space to reproduce. However, there is another component $-byx$, that is, the population also is decreasing proportionally to the number of foxes. Together we can write the equation as $(a - by)x$, so it is like exponential growth or decay but the constant depends on the number of foxes.

The equation for foxes is very similar, expand again

$$y' = (cx - d)y = cxy - dy.$$

[*]Named for the American mathematician, chemist, and statistician Alfred James Lotka (1880–1949) and the Italian mathematician and physicist Vito Volterra (1860–1940).

[†]This interaction does not end well for the hare.

The foxes need food (hares) to reproduce: the more food, the bigger the rate of growth, hence the cxy term. On the other hand, there are natural deaths in the fox population, and hence the $-dy$ term.

Without further delay, let us start with an explicit example. Suppose the equations are

$$x' = (0.4 - 0.01y)x, \qquad y' = (0.003x - 0.3)y.$$

See Figure 5.14 for the phase portrait. In this example it makes sense to also plot x and y as graphs with respect to time. Therefore the second graph in Figure 5.14 is the graph of x and y on the vertical axis (the prey x is the thinner blue line with taller peaks), against time on the horizontal axis. The particular solution graphed was with initial conditions of 20 foxes and 50 hares.

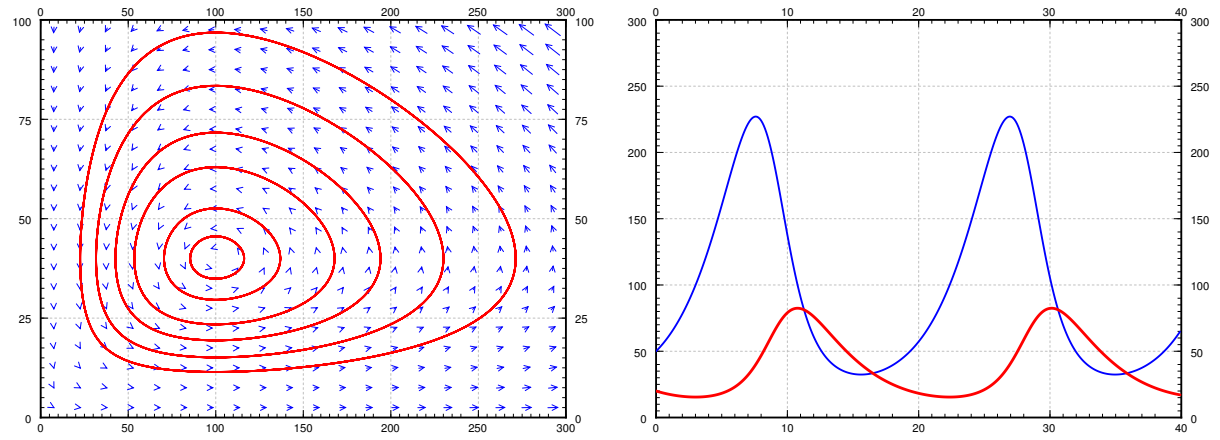


Figure 5.14: The phase portrait (left) and graphs of x and y for a sample solution (right).

Let us analyze what we see on the graphs. We work in the general setting rather than putting in specific numbers. We start with finding the critical points. Set $(a - by)x = 0$, and $(cx - d)y = 0$. The first equation is satisfied if either $x = 0$ or $y = a/b$. If $x = 0$, the second equation implies $y = 0$. If $y = a/b$, the second equation implies $x = d/c$. There are two equilibria: at $(0, 0)$ when there are no animals at all, and at $(d/c, a/b)$. In our specific example $x = d/c = 100$, and $y = a/b = 40$. This is the point where there are 100 hares and 40 foxes.

We compute the Jacobian matrix:

$$\begin{bmatrix} a - by & -bx \\ cy & cx - d \end{bmatrix}.$$

At the origin $(0, 0)$ we get the matrix $\left[\begin{smallmatrix} a & 0 \\ 0 & -d \end{smallmatrix}\right]$, so the eigenvalues are a and $-d$, hence real and of opposite signs. So the critical point at the origin is a saddle. This makes sense. If you started with some foxes but no hares, then the foxes would go extinct, that is, you would approach the origin. If you started with no foxes and a few hares, then the hares would keep multiplying without check, and so you would go away from the origin.

OK, how about the other critical point at $(d/c, a/b)$. Here the Jacobian matrix becomes

$$\begin{bmatrix} 0 & -\frac{bd}{c} \\ \frac{ac}{b} & 0 \end{bmatrix}.$$

The eigenvalues satisfy $\lambda^2 + ad = 0$. In other words, $\lambda = \pm i\sqrt{ad}$. The eigenvalues being purely imaginary, we are in the case where we cannot quite decide using only linearization. We could have a stable center, spiral sink, or a spiral source. That is, the equilibrium could be asymptotically stable, stable, or unstable. Of course I gave you a picture above that seems to imply it is a stable center. But never trust a picture only. Perhaps the oscillations are getting larger and larger, but only *very* slowly. Of course this would be bad as it would imply something will go wrong with our population sooner or later. And I only graphed a very specific example with very specific trajectories.

How can we be sure we are in the stable situation? As we said before, in the case of purely imaginary eigenvalues, we have to do a bit more work. The main approach that can be used here is to directly solve for the trajectories. We can determine a differential equation that relates x to y by writing

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{(cx-d)y}{(a-by)x}.$$

This is a separable first order equation, which we can rewrite as

$$\frac{a-by}{y}dy = \frac{cx-d}{x}dx.$$

After simplifying the fractions, we can integrate this to obtain the implicit solution

$$a\ln(y) - by = cx - d\ln(x) + C$$

or

$$C = a\ln(y) + d\ln(x) - cx - by. \tag{5.8}$$

Since we ended up finding a trajectory here that sketches out a closed curve, and we know that our solutions must lie on these trajectories, that tells us that, for a fact, we do have closed loops here, and the critical point is stable.

However, we can go a bit farther than this with our discussion here. If we let $D = e^C$ in (5.8), we can rearrange the expression to get that

$$D = \frac{y^a x^d}{e^{cx+by}} = y^a x^d e^{-cx-by},$$

and based on how our trajectory setup works, we know that this D will be conserved along the flow of the solution. That is, if the initial condition has a specific value of D, the solution will continue to have that same value for all t. This idea came up before in the idea of conservative or Hamiltonian systems in § 5.2. Such a quantity is called the *constant of motion*, and this forces the trajectory to go in closed loops. Let us check D really is a constant of

motion. How do we check, you say? Well, a constant is something that does not change with time, so let us compute the derivative with respect to time:

$$D' = ay^{a-1}y'x^d e^{-cx-by} + y^a dx^{d-1}x' e^{-cx-by} + y^a x^d e^{-cx-by}(-cx' - by').$$

Our equations give us what x' and y' are so let us plug those in:

$$\begin{aligned} D' &= ay^{a-1}(cx-d)yx^d e^{-cx-by} + y^a dx^{d-1}(a-by)xe^{-cx-by} \\ &\quad + y^a x^d e^{-cx-by}\big(-c(a-by)x - b(cx-d)y\big) \\ &= y^a x^d e^{-cx-by}\Big(a(cx-d) + d(a-by) + \big(-c(a-by)x - b(cx-d)y\big)\Big) \\ &= 0. \end{aligned}$$

So along the trajectories D is constant. In fact, the expression $D = \frac{y^a x^d}{e^{cx+by}}$ gives us an implicit equation for the trajectories. In any case, once we have found this constant of motion, it must be true that the trajectories are simple curves, that is, the level curves of $\frac{y^a x^d}{e^{cx+by}}$. It turns out, the critical point at $(d/c, a/b)$ is a maximum for D (left as an exercise). So $(d/c, a/b)$ is a stable equilibrium point, and we do not have to worry about the foxes and hares going extinct or their populations exploding.

One blemish on this wonderful model is that the number of foxes and hares are discrete quantities and we are modeling with continuous variables. Our model has no problem with there being 0.1 fox in the forest for example, while in reality that makes no sense. The approximation is a reasonable one as long as the number of foxes and hares are large, but it does not make much sense for small numbers. One must be careful in interpreting any results from such a model.

An interesting consequence (perhaps counterintuitive) of this model is that adding animals to the forest might lead to extinction, because the variations will get too big, and one of the populations will get close to zero. For example, suppose there are 20 foxes and 50 hares as before, but now we bring in more foxes, bringing their number to 200. If we run the computation, we find the number of hares will plummet to just slightly more than 1 hare in the whole forest. In reality that most likely means the hares die out, and then the foxes will die out as well as they will have nothing to eat.

Example 5.3.1: Consider the system

$$x' = (2y-6)x \qquad y' = (2-x)y.$$

This fits the description of a predator-prey model. Which species is the predator? Find and analyze the critical points of this system, and draw a sketch of the phase portrait, with arrows to indicate the direction of flow around this portrait.

Solution: If we expand out the equations in the model, we get

$$x' = 2xy - 6x \qquad y' = 2y - xy.$$

These equations show that, if $y = 0$, x would decay away in time, and if $x = 0$, y would grow indefinitely. This means that x is the predator and y is the prey in this relationship. For

critical points, we can look back at the factored version of the equations to see that we get one critical point at $(0, 0)$ and one critical point at $(2, 3)$. Since this is a predator-prey model, we know that we will have cycles around the critical point at $(2, 3)$.

The direction of these cycles is determined by the predator-prey relationship. If we start with x large (greater than 2) and y small (less than 3), then there are a lot of predators and few prey. This implies that the next thing to happen is that the predator population will decrease because there is not enough prey. We can also see this from the equations; if $x \geq 2$ and $y \leq 3$, then both $\frac{dx}{dt}$ and $\frac{dy}{dt}$ will be negative. Similarly, if y is large and x is small, there are a lot of prey and few predators, so the prey population will continue to grow, while the predators also grow because of the excess of food. This means that the populations will follow these trajectories in a clockwise direction.

For the actual trajectories, we can solve for them in the same way as the calculations before this example. We can rewrite this system to give a differential equation for the trajectories as

$$\frac{dy}{dx} = \frac{(2-x)y}{(2y-6)x}$$

which can be rearranged as a separable equation to

$$\left(2 - \frac{6}{y}\right) dy = \left(\frac{2}{x} - 1\right) dx.$$

Solving this gives

$$2y - 6\ln(y) + C = 2\ln(x) - x$$

or

$$C = 2\ln(x) + 6\ln(y) - x - 2y.$$

This will be used to draw the trajectories in Figure 5.15.

This can also be seen using a nullcline analysis. The nullclines we need to draw are $x = 0$, $y = 0$, $x = 2$ and $y = 3$. Our discussion previously shows that the arrow in the bottom-right quadrant should point to the lower left, and the arrow in the top left should point up and right. We can fill in the other two quadrants to see that the solution should move around the circle in a clockwise direction. Figure 5.15 shows the nullcline image and trajectory curves for this example. ⌟

5.3.3 Competing Species systems

Another application of non-linear systems that also works with population models is a competing species interaction. The setup is that there are two species that live in the same environment, and need to compete over resources. This means that both species will grow on their own, but when the two species interact, it is negative for both species. This gives rise to a system of differential equations of the form

$$\frac{dx}{dt} = ax - bxy$$
$$\frac{dy}{dt} = cy - dxy$$

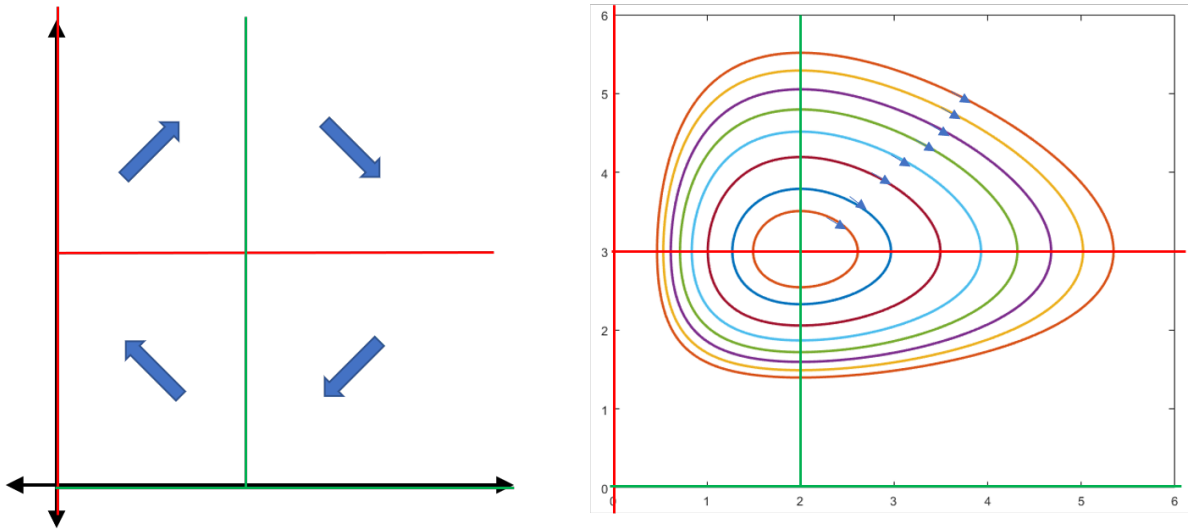


Figure 5.15: The nullcline analysis image (left) and graphs of several sample trajectories with arrows (right) for Example 5.3.1

if both populations grow exponentially, or

$$\frac{dx}{dt} = ax(K - x) - bxy$$
$$\frac{dy}{dt} = cy(M - y) - dxy$$

if both species grow logistically. The numbers here are all positive constants that explain how the different populations affect growth rates. For the logistic model, let's look at the equilibrium solutions. For this, we need

$$x(aK - ax - by) = 0 \qquad y(cM - cy - dx) = 0$$

which gives equilibrium solutions at $(0, 0)$, $(0, M)$, and $(K, 0)$, all of which result in one (or both) of the species being extinct. The other equilibrium solution is more interesting, because it involves both species coexisting. This happens when

$$by = aK - ax \qquad cy = cM - dx.$$

Solving this gives a critical point with $x > 0$ and $y > 0$.

The Jacobian matrix for this system is

$$J(x, y) = \begin{bmatrix} aK - 2ax - by & -bx \\ -dy & cM - 2cy - dx \end{bmatrix}.$$

Unlike the predator-prey system that always had the same type of equilibriums solution every time, there are multiple options for how this system can behave based on the values of

a, b, c, d, K, and M. It is possible that the coexistence equilibrium solution will be a nodal sink, so that all nearby solutions will converge to it over time, and the species will continue to exist in harmony. However, it is also possible that the coexistence solution is a saddle and the solutions at $(K, 0)$ and $(0, M)$ are sinks. This means that coexistence is unstable, and that over time, the populations will converge to one of the other two equilibrium solutions, meaning that one of the species will die out as time goes on. Determining which will survive will require a numerical model since these equations can not be solved analytically.

Example 5.3.2: Analyze the competing species model given by the system of differential equations

$$x' = x(4 - x - 2y) \qquad y' = y(7 - y - 3x).$$

Is the coexistence solution stable or unstable? What will happen to the populations over time?

Solution: Solving for the equilibrium solutions gives $(0, 0)$, $(4, 0)$, $(0, 7)$, and the coexistence solution where

$$4 - x = 2y \qquad y = 7 - 3x.$$

Simplifying this gives

$$4 - x = 14 - 6x$$

or $x = 2$. The second equation then implies that $y = 1$.

The Jacobian for this system is

$$J(x, y) = \begin{bmatrix} 4 - 2x - 2y & -2x \\ -3y & 7 - 2y - 3x \end{bmatrix}.$$

Evaluating this matrix at the point $(2, 1)$ gives

$$\begin{bmatrix} -2 & -4 \\ -3 & -1 \end{bmatrix},$$

which we need to find the eigenvalues to classify what type of linearized solution we have here. These are determined by

$$(-2 - \lambda)(-1 - \lambda) - 12 = \lambda^2 + 3\lambda - 9 = 0.$$

Thus, the eigenvalues are given by

$$\lambda = \frac{-3 \pm \sqrt{9 + 36}}{2}$$

which will be real with opposite signs. Therefore, this equilibrium solution is a saddle, and unstable. To confirm this, we can also check the equilibrium solutions at $(4, 0)$ and $(0, 7)$. For $(4, 0)$, we get the matrix

$$\begin{bmatrix} -8 & -8 \\ 0 & -1 \end{bmatrix}$$

which is a nodal sink. For $(0, 7)$, we get

$$\begin{bmatrix} -10 & 0 \\ -21 & -7 \end{bmatrix}$$

which is also a nodal sink. Thus, we see that the coexistence equilibrium solution is unstable, and both of the equilibrium solutions with one species extinct are stable. Therefore, over time, one of the two species will die off depending on the initial population. ⌟

Showing that a system of equations has a stable solution can be a very difficult problem. When Isaac Newton put forth his laws of planetary motions, he proved that a single planet orbiting a single sun is a stable system. But any solar system with more than 1 planet proved very difficult indeed. In fact, such a system behaves chaotically (see § 5.5), meaning small changes in initial conditions lead to very different long-term outcomes. From numerical experimentation and measurements, we know the earth will not fly out into the empty space or crash into the sun, for at least some millions of years or so. But we do not know what happens beyond that.

5.3.4 Exercises

***Exercise* 5.3.1:** *Take the* damped nonlinear pendulum equation $\theta'' + \mu\theta' + (g/L)\sin\theta = 0$ *for some* $\mu > 0$ *(that is, there is some friction).*

a) Suppose $\mu = 1$ *and* $g/L = 1$ *for simplicity, find and classify the critical points.*

b) Do the same for any $\mu > 0$ *and any* g *and* L*, but such that the damping is small, in particular,* $\mu^2 < 4(g/L)$*.*

c) Explain what your findings mean, and if it agrees with what you expect in reality.

Exercise* 5.3.2: *Take the damped nonlinear pendulum equation* $\theta'' + \mu\theta' + (g/L)\sin\theta = 0$ *for some* $\mu > 0$ *(that is, there is friction). Suppose the friction is large, in particular* $\mu^2 > 4(g/L)$*.*

a) Find and classify the critical points.

b) Explain what your findings mean, and if it agrees with what you expect in reality.

***Exercise* 5.3.3:** *Suppose the hares do not grow exponentially, but logistically. In particular consider*

$$x' = (0.4 - 0.01y)x - \gamma x^2, \qquad y' = (0.003x - 0.3)y.$$

For the following two values of γ*, find and classify all the critical points in the positive quadrant, that is, for* $x \geq 0$ *and* $y \geq 0$*. Then sketch the phase diagram. Discuss the implication for the long term behavior of the population.*

a) $\gamma = 0.001$*,* *b)* $\gamma = 0.01$*.*

Exercise* 5.3.4: *Suppose we have the system predator-prey system where the foxes are also killed at a constant rate* h *(*h *foxes killed per unit time):* $x' = (a - by)x$*,* $y' = (cx - d)y - h$*.*

a) *Find the critical points and the Jacobian matrices of the system.*

b) *Put in the constants $a = 0.4$, $b = 0.01$, $c = 0.003$, $d = 0.3$, $h = 10$. Analyze the critical points. What do you think it says about the forest?*

Exercise* 5.3.5** (challenging):**** *Suppose the foxes never die. That is, we have the system $x' = (a - by)x$, $y' = cxy$. Find the critical points and notice they are not isolated. What will happen to the population in the forest if it starts at some positive numbers. Hint: Think of the constant of motion.*

***Exercise* 5.3.6:** *The following system of differential equations models a pair of populations interacting.*

$$\frac{dx}{dt} = 4x - 2xy \qquad \frac{dy}{dt} = 3xy - y$$

a) *Does this system of differential equations better fit with a competing species model or a predator-prey model? If it is predator-prey, which species is the predator?*

b) *Find and classify the critical point (if it exists) with both $x > 0$ and $y > 0$.*

c) *Describe what is going to happen to the population of these species over time. It this depends on the initial condition, say so.*

***Exercise* 5.3.7:** *The following system of differential equations models a pair of populations interacting.*

$$\frac{dx}{dt} = x(6 - 3y - 2x) \qquad \frac{dy}{dt} = y(4 - y - 3x)$$

a) *Does this system of differential equations better fit with a competing species model or a predator-prey model? If it is predator-prey, which species is the predator?*

b) *Find and classify the critical point (if it exists) with both $x > 0$ and $y > 0$.*

c) *Describe what is going to happen to the population of these species over time. It this depends on the initial condition, say so.*

***Exercise* 5.3.8:** *The following system of differential equations models a pair of populations interacting.*

$$\frac{dx}{dt} = x(5 - x - 2y) \qquad \frac{dy}{dt} = y(7 - x - 3y)$$

a) *Does this system of differential equations better fit with a competing species model or a predator-prey model? If it is predator-prey, which species is the predator?*

b) *Find and classify the critical point (if it exists) with both $x > 0$ and $y > 0$.*

c) *Describe what is going to happen to the population of these species over time. It this depends on the initial condition, say so.*

***Exercise* 5.3.9:**

a) *Suppose x and y are positive variables. Show $\frac{yx}{e^{x+y}}$ attains a maximum at $(1,1)$.*

b) *Suppose a, b, c, d are positive constants, and also suppose x and y are positive variables. Show $\frac{y^a x^d}{e^{cx+by}}$ attains a maximum at $(d/c, a/b)$.*

***Exercise* 5.3.10:** *Suppose that for the pendulum equation we take a trajectory giving the spinning-around motion, for example $\omega = \sqrt{\frac{2g}{L}\cos\theta + \frac{2g}{L} + \omega_0^2}$. This is the trajectory where the lowest angular velocity is ω_0^2. Find an integral expression for how long it takes the pendulum to go all the way around.*

***Exercise* 5.3.11:** *Consider a predator-prey interaction where humans have gotten involved. The idea is that at least one of the species is valuable for food or another resource, and the two species still intact in their normal predator-prey manner. The first version of this will deal with "constant effort harvesting," which means that humans will remove animals from the populations are a rate proportional to the population. This results in equations of the form*

$$\frac{dx}{dt} = x(a - by - E_1) \qquad \frac{dy}{dt} = y(-d + cx - E_2)$$

where E_1 and E_2 denote the amount of harvesting done.

a) *There is a single equilibrium solution with $x > 0$ and $y > 0$ in the case of no harvesting, that is, $E_1 = E_2 = 0$. Find this equilibrium solution.*

b) *Without doing any mathematical work, what do you think will happen to the equilibrium solution if just the prey is harvested? What if just the predator is harvested? What if both are harvested?*

c) *Find the location of the equilibrium system in each of the three cases in the previous part. Do this in terms of the constants E_1 and E_2 for all three cases.*

***Exercise* 5.3.12:** *The second version of this will deal with "constant yield harvesting," which means that humans will remove animals from the populations at a fixed rate, no matter their population. This results in equations of the form*

$$\frac{dx}{dt} = x(a - by) - H_1 \qquad \frac{dy}{dt} = y(-d + cx) - H_2$$

where H_1 and H_2 denote the amount of harvesting done.

a) *There is a single equilibrium solution with $x > 0$ and $y > 0$ in the case of no harvesting, that is, $H_1 = H_2 = 0$. Find this equilibrium solution.*

b) *Without doing any mathematical work, what do you think will happen to the equilibrium solution if just the prey is harvested? What if just the predator is harvested? What if both are harvested?*

c) *Find the location of the equilibrium system in each of the three cases in the previous part. Do this in terms of the constants H_1 and H_2 for all three cases.*

***Exercise* 5.3.13:** *The general competing species model has the form*

$$\frac{dx}{dt} = x(\rho_1 - \gamma_1 y - M_1 x) \qquad \frac{dy}{dt} = y(\rho_2 - \gamma_2 x - M_2 y)$$

where ρ indicates the growth rate, M is related to the carrying capacity, and γ is connected to the interaction term. Assume that this model is being used to represent species A and B of fish living in a pond at time t, which is initially stocked with both species of fish. We want to analyze the behavior of this equation under different sets of coefficients.

a) *If $\rho_2/\gamma_2 > \rho_1/M_1$ and $\rho_2/M_2 > \rho_1/\gamma_1$, show that the only equilibrium populations in the pond are no fish, no fish of species A, or no fish of species B. What happens for large values of t?*

b) *If $\rho_1/M_1 > \rho_2/\gamma_2$ and $\rho_1/\gamma_1 > \rho_2/M_2$, show that the only equilibrium populations in the pond are no fish, no fish of species A, or no fish of species B. What happens for large values of t?*

c) *Suppose that $\rho_2/\gamma_2 > \rho_1/M_1$ and $\rho_1/\gamma_1 > \rho_2/M_2$. Show that there is a stable equilibrium where both species coexist.*

***Exercise* 5.3.14** (challenging)**:** *Take the pendulum, suppose the initial position is $\theta = 0$.*

a) *Find the expression for ω giving the trajectory with initial condition $(0, \omega_0)$. Hint: Figure out what C should be in terms of ω_0.*

b) *Find the crucial angular velocity ω_1, such that for any higher initial angular velocity, the pendulum will keep going around its axis, and for any lower initial angular velocity, the pendulum will simply swing back and forth. Hint: When the pendulum doesn't go over the top the expression for ω will be undefined for some θs.*

c) *What do you think happens if the initial condition is $(0, \omega_1)$, that is, the initial angle is 0, and the initial angular velocity is exactly ω_1.*

5.4 Limit cycles

Attribution: [JL], §8.4.

Learning Objectives

After this section, you will be able to:

- Identify differential equations that have limit cycles from slope fields and
- Find and classify limit cycles of systems of differential equations by converting the system to depend on radius.

For nonlinear systems, trajectories do not simply need to approach or leave a single point. They may in fact approach a larger set, such as a circle or another closed curve.

Example 5.4.1: The *Van der Pol oscillator** is the following equation

$$x'' - \mu(1 - x^2)x' + x = 0,$$

where μ is some positive constant. The Van der Pol oscillator originated with electrical circuits, but finds applications in diverse fields such as biology, seismology, and other physical sciences.

For simplicity, let us use $\mu = 1$. A phase diagram is given in the left-hand plot in Figure 5.16. Notice how the trajectories seem to very quickly settle on a closed curve. On the right-hand side is the plot of a single solution for $t = 0$ to $t = 30$ with initial conditions $x(0) = 0.1$ and $x'(0) = 0.1$. The solution quickly tends to a periodic solution.

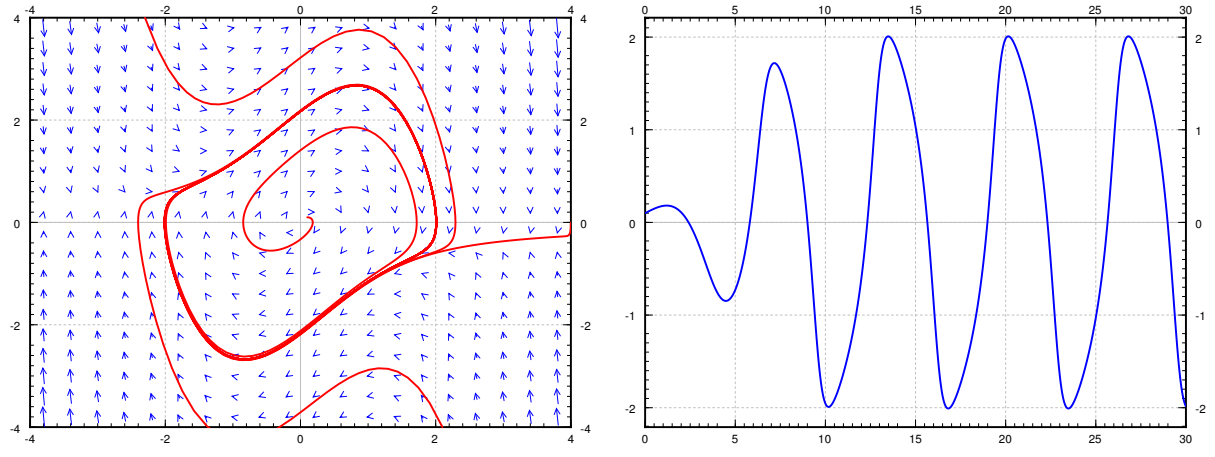


Figure 5.16: The phase portrait (left) and a graph of a sample solution of the Van der Pol oscillator.

The Van der Pol oscillator is an example of so-called *relaxation oscillation*. The word relaxation comes from the sudden jump (the very steep part of the solution). For larger μ

*Named for the Dutch physicist Balthasar van der Pol (1889–1959).

the steep part becomes even more pronounced, for small μ the limit cycle looks more like a circle. In fact, setting $\mu = 0$, we get $x'' + x = 0$, which is a linear system with a center and all trajectories become circles.

What we see in this example is a curve to which many solution seem to head towards as t gets larger. This motivates the following definition.

Definition 5.4.1

1. A trajectory in the phase portrait that is a closed curve (a curve that is a loop) is called a *closed trajectory*.
2. A *limit cycle* is a closed trajectory such that at least one other trajectory spirals into it.
3. If all trajectories that start near the limit cycle spiral into it, the limit cycle is called *asymptotically stable*.

For example, the closed curve in the phase portrait for the Van der Pol equation is a limit cycle, and the limit cycle in the Van der Pol oscillator is asymptotically stable.

Given a closed trajectory on an autonomous system, any solution that starts on it is periodic. Such a curve is called a *periodic orbit*. More precisely, if $\big(x(t), y(t)\big)$ is a solution such that for some t_0 the point $\big(x(t_0), y(t_0)\big)$ lies on a periodic orbit, then both $x(t)$ and $y(t)$ are periodic functions (with the same period). That is, there is some number P such that $x(t) = x(t + P)$ and $y(t) = y(t + P)$.

We would like to be able to identify when these sorts of periodic orbits can or can't happen to understand more about these systems. Thankfully, we have a theorem that gives us some help here.

Theorem 5.4.1 (Poincaré–Bendixson)

Consider the system

$$x' = f(x, y), \qquad y' = g(x, y), \tag{5.9}$$

where the functions f and g have continuous derivatives in some region R in the plane. Suppose R is a closed bounded region (a region in the plane that includes its boundary and does not have points arbitrarily far from the origin). Suppose $\big(x(t), y(t)\big)$ is a solution of (5.9) in R that exists for all $t \geq t_0$. Then either the solution is a periodic function, or the solution tends towards a periodic solution in R.

The main point of the theorem*is that if you find one solution that exists for all t large enough (that is, as t goes to infinity) and stays within a bounded region, then you have found either a periodic orbit, or a solution that spirals towards a limit cycle or tends to a critical point. That is, in the long term, the behavior is very close to a periodic function. Note that a constant solution at a critical point is periodic (with any period). The theorem is more a qualitative statement rather than something to help us in computations. In practice it is hard

*Ivar Otto Bendixson (1861–1935) was a Swedish mathematician.

to find analytic solutions and so hard to show rigorously that they exist for all time. But if we think the solution exists we numerically solve for a large time to approximate the limit cycle. Another caveat is that the theorem only works in two dimensions. In three dimensions and higher, there is simply too much room.

The theorem applies to all solutions in the Van der Pol oscillator. Solutions that start at any point except the origin $(0,0)$ will tend to the periodic solution around the limit cycle, and if the initial condition of $(0,0)$ will lead to the constant solution $x = 0$, $y = 0$.

Example 5.4.2: Consider

$$x' = y + (x^2 + y^2 - 1)^2 x, \qquad y' = -x + (x^2 + y^2 - 1)^2 y.$$

A vector field along with solutions with initial conditions $(1.02, 0)$, $(0.9, 0)$, and $(0.1, 0)$ are drawn in Figure 5.17. Analyze this system to determine what will happen to the solution for a variety of initial conditions.

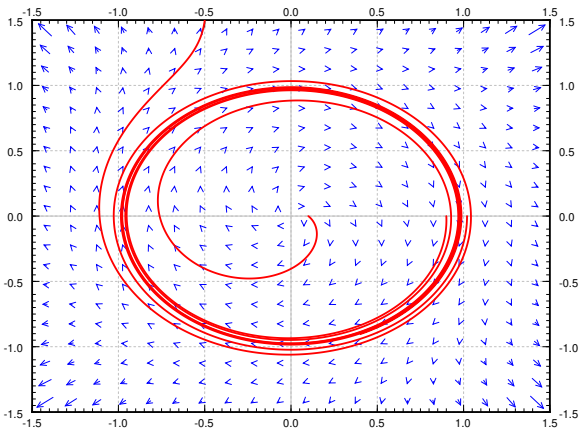


Figure 5.17: Semistable limit cycle example.

Solution: Notice that points on the unit circle (distance one from the origin) satisfy $x^2 + y^2 - 1 = 0$. And $x(t) = \sin(t)$, $y = \cos(t)$ is a solution of the system. Therefore we have a closed trajectory. For points off the unit circle, the second term in x' pushes the solution further away from the y-axis than the system $x' = y$, $y' = -x$, and y' pushes the solution further away from the x-axis than the linear system $x' = y$, $y' = -x$. In other words for all other initial conditions the trajectory will spiral out.

This means that for initial conditions inside the unit circle, the solution spirals out towards the periodic solution on the unit circle, and for initial conditions outside the unit circle the solutions spiral off towards infinity. Therefore the unit circle is a limit cycle, but not an asymptotically stable one. In relation to the terms used for autonomous equations in § 1.7, we could refer to this as a semistable limit cycle, since on one side (inside) the solutions spiral towards the periodic orbit, while on the other side (outside) the solutions move away. The Poincaré–Bendixson Theorem applies to the initial points inside the unit circle, as those solutions stay bounded, but not to those outside, as those solutions go off to infinity. ⌟

A very similar analysis applies to the system

$$x' = y + (x^2 + y^2 - 1)x, \qquad y' = -x + (x^2 + y^2 - 1)y.$$

We still obtain a closed trajectory on the unit circle, and points outside the unit circle spiral out to infinity, but now points inside the unit circle spiral towards the critical point at the origin. So this system does not have a limit cycle, even though it has a closed trajectory.

One way to see this more explicitly is by trying to write this all in terms of

$$r = \sqrt{x^2 + y^2}.$$

For simplicity here, we will determine everything in terms of

$$s = r^2 = x^2 + y^2$$

because as long as $r > 0$, r and s always have the same behavior (in terms of increasing and decreasing), and it is easier to compute with s.

Using the first example

$$x' = y + (x^2 + y^2 - 1)^2 x, \qquad y' = -x + (x^2 + y^2 - 1)^2 y.$$

we see that

$$\begin{aligned} s' &= 2xx' + 2yy' \\ &= 2x(y + (x^2 + y^2 - 1)^2 x + 2y(-x + (x^2 + y^2 - 1)^2 y) \\ &= 2xy + 2x^2(x^2 + y^2 - 1)^2 - 2xy + 2y^2(x^2 + y^2 - 1)^2 \\ s' &= 2s(s-1)^2 \end{aligned}$$

Thus, we are left with the equation

$$\frac{ds}{dt} = 2s(s-1)^2$$

which is an autonomous first-order equation that we can analyze. We have two equilibrium solutions in terms of s at $s = 0$, which corresponds to the origin, and $s = 1$, which corresponds to the unit circle. We can then plug in values to see that for $s = \frac{1}{2}$, $\frac{ds}{dt} > 0$, so that the solutions will increase out to the unit circle. For $s > 1$, $\frac{ds}{dt} > 0$ as well, so solutions move away from the circle outside it. This is the same as the result we obtained in the first example.

For the second example, we end up with the autonomous equation

$$\frac{ds}{dt} = 2s(s-1)$$

which is negative for $0 < s < 1$ and positive for $1 < s$, giving that solutions that start inside the unit circle will converge to the origin, and solutions that start outside the circle will move away from it.

Due to the Picard theorem (Theorem 4.1.1 on page 278) we find that no matter where we are in the plane we can always find a solution a little bit further in time, as long as f and g

have continuous derivatives. So if we find a closed trajectory in an autonomous system, then for every initial point inside the closed trajectory, the solution will exist for all time and it will stay bounded (it will stay inside the closed trajectory). Since the closed trajectory is a solution, we can not cross it (by Picard theorem), and so we have to stay trapped inside. So the moment we found the solution above going around the unit circle, we knew that for every initial point inside the circle, the solution exists for all time and the Poincaré–Bendixson theorem applies.

Let us next look for conditions when limit cycles (or periodic orbits) do not exist. We assume the equation (5.9) is defined on a *simply connected region*, that is, a region with no holes we can go around. For example the entire plane is a simply connected region, and so is the inside of the unit disc. However, the entire plane minus a point is not a simply connected domain as it has a "hole" at the origin.

Theorem 5.4.2 (Bendixson–Dulac)

Suppose R is a simply connected region, and the expression[a]

$$\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y}$$

is either always positive or always negative on R (except perhaps a small set such as on isolated points or curves) then the system (5.9) has no closed trajectory inside R.

[a] Usually the expression in the Bendixson–Dulac Theorem is $\frac{\partial(\varphi f)}{\partial x} + \frac{\partial(\varphi g)}{\partial y}$ for some continuously differentiable function φ. For simplicity, let us just consider the case $\varphi = 1$.

The theorem* gives us a way of ruling out the existence of a closed trajectory, and hence a way of ruling out limit cycles. The exception about points or curves means that we can allow the expression to be zero at a few points, or perhaps on a curve, but not on any larger set.

Example 5.4.3: Let us look at $x' = y + y^2 e^x$, $y' = x$ in the entire plane (see Example 5.1.4 on page 377) and try to apply Theorem 5.4.2.

Solution: The entire plane is simply connected and so we can apply the theorem. We compute $\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = y^2 e^x + 0$. The function $y^2 e^x$ is always positive except on the line $y = 0$. Therefore, via the theorem, the system has no closed trajectories. □

In some books (or the internet) the theorem is not stated carefully and it concludes there are no periodic solutions. That is not quite right. The example above has two critical points and hence it has constant solutions, and constant functions are periodic. The conclusion of the theorem should be that there exist no trajectories that form closed curves. Another way to state the conclusion of the theorem would be to say that there exist no nonconstant periodic solutions that stay in R.

Let us look at a somewhat more complicated example.

Example 5.4.4: Take the system $x' = -y - x^2$, $y' = -x + y^2$ (see Example 5.1.3 on page 376) and look at how Theorem 5.4.2 works here.

*Henri Dulac (1870–1955) was a French mathematician.

Solution: We compute $\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = -2x + 2y = 2(-x+y)$. This expression takes on both signs, so if we are talking about the whole plane we cannot simply apply the theorem. However, we could apply it on the set where $-x + y \geq 0$. Via the theorem, there is no closed trajectory in that set. Similarly, there is no closed trajectory in the set $-x + y \leq 0$. We cannot conclude (yet) that there is no closed trajectory in the entire plane. Perhaps half of it is in the set where $-x + y \geq 0$ and the other half is in the set where $-x + y \leq 0$.

The key is to look at the line where $-x + y = 0$, or $x = y$. On this line $x' = -y - x^2 = -x - x^2$ and $y' = -x + y^2 = -x + x^2$. In particular, when $x = y$ then $x' \leq y'$. That means that the arrows, the vectors (x', y'), always point into the set where $-x + y \geq 0$. There is no way we can start in the set where $-x + y \geq 0$ and go into the set where $-x + y \leq 0$. Once we are in the set where $-x + y \geq 0$, we stay there. So no closed trajectory can have points in both sets. ⌟

Example 5.4.5: Consider $x' = y + (x^2 + y^2 - 1)x$, $y' = -x + (x^2 + y^2 - 1)y$, and consider the region R given by $x^2 + y^2 > \frac{1}{2}$. That is, R is the region outside a circle of radius $\frac{1}{\sqrt{2}}$ centered at the origin. Then there is a closed trajectory in R, namely $x = \cos(t)$, $y = \sin(t)$. Furthermore,

$$\frac{\partial f}{\partial x} + \frac{\partial g}{\partial x} = 4x^2 + 4y^2 - 2,$$

which is always positive on R. So what is going on? The Bendixson–Dulac theorem does not apply since the region R is not simply connected—it has a hole, the circle we cut out!

5.4.1 Exercises

***Exercise* 5.4.1:** *Consider the two-dimensional system of differential equation written in polar coordinates as*

$$\frac{dr}{dt} = r(r-1)(r-4)^2 \qquad \frac{d\theta}{dt} = 1.$$

Determine all limit cycles, periodic solutions, and classify the stability of each of these solutions.

***Exercise* 5.4.2:** *Consider the two-dimensional system of differential equation written in polar coordinates as*

$$\frac{dr}{dt} = r^2(r-1)^2(r-3) \qquad \frac{d\theta}{dt} = -1.$$

Determine all limit cycles, periodic solutions, and classify the stability of each of these solutions.

Exercise* 5.4.3: *Consider the system of differential equation given by*

$$\frac{dx}{dt} = x(3 - 2y^2 - x^2) \qquad \frac{dy}{dt} = y(3 - y^2).$$

Find and classify all limit cycles by converting to an autonomous equation in $r = \sqrt{x^2 + y^2}$ or $s = x^2 + y^2$.

Exercise 5.4.4:* *Consider the system of differential equation given by*

$$\frac{dx}{dt} = -x(x^2+y^2)^2 + 6x(x^2+y^2) - 8x + 6y \qquad \frac{dy}{dt} = -y(x^2+y^2)^2 + 6y(x^2+y^2) - 8y - 6x.$$

Find and classify all limit cycles by converting to an autonomous equation in $r = \sqrt{x^2+y^2}$ *or* $s = x^2+y^2$.

Exercise 5.4.5: *Consider the system*

$$\left[\begin{array}{l} \dfrac{dx}{dt} = x + 2y + x(x^2+y^2-2\sqrt{x^2+y^2}) \\ \dfrac{dy}{dt} = -2x + y + y(x^2+y^2-2\sqrt{x^2+y^2}) \end{array}\right]. \tag{5.10}$$

a) *Use polar coordinates to write* $\dfrac{dr}{dt}$ *as a function of* r.

b) *Draw the phase line of the DE* $\dfrac{dr}{dt} = f(r)$, *where* $f(r)$ *is the function from part a.*

c) *Does the system (5.10) have a limit cycle? If so, find it. If not, explain why not. For each positive root of* $f(r)$, *decide whether the corresponding trajectory one is stable, unstable, or semistable.*

Exercise 5.4.6: *Show that the following systems have no closed trajectories.*

a) $x' = x^3 + y, \quad y' = y^3 + x^2,$ b) $x' = e^{x-y}, \quad y' = e^{x+y},$

c) $x' = x + 3y^2 - y^3, \quad y' = y^3 + x^2.$

Exercise 5.4.7:* *Show that the following systems have no closed trajectories.*

a) $x' = x + y^2, \quad y' = y + x^2,$ b) $x' = -x\sin^2(y), \quad y' = e^x,$

c) $x' = xy, \quad y' = x + x^2.$

Exercise 5.4.8:* *Suppose an autonomous system in the plane has a solution* $x = \cos(t) + e^{-t}$, $y = \sin(t) + e^{-t}$. *What can you say about the system (in particular about limit cycles and periodic solutions)?*

Exercise 5.4.9: *Formulate a condition for a 2-by-2 linear system* $\vec{x}' = A\vec{x}$ *to not be a center using the Bendixson–Dulac theorem. That is, the theorem says something about certain elements of* A.

Exercise 5.4.10: *Explain why the Bendixson–Dulac Theorem does not apply for any conservative system* $x'' + h(x) = 0$.

Exercise 5.4.11: *A system such as* $x' = x, y' = y$ *has solutions that exist for all time* t, *yet there are no closed trajectories. Explain why the Poincaré–Bendixson Theorem does not apply.*

Exercise* 5.4.12*: *Show that the limit cycle of the Van der Pol oscillator (for $\mu > 0$) must not lie completely in the set where $-1 < x < 1$. Compare with Figure 5.16 on page 413.*

Exercise* 5.4.13*: *Differential equations can also be given in different coordinate systems. Suppose we have the system $r' = 1 - r^2$, $\theta' = 1$ given in polar coordinates. Find all the closed trajectories and check if they are limit cycles and if so, if they are asymptotically stable or not.*

Exercise* 5.4.14*: *Suppose we have the system $r' = \sin(r)$, $\theta' = 1$ given in polar coordinates. Find all the closed trajectories.*

5.5 Chaos

Attribution: [JL], §8.5.

Learning Objectives

After this section, you will be able to:

- Identify chaotic behavior and how it is distinct from other types of equations.

You have surely heard the idea of the "butterfly effect," that the flap of a butterfly wing in the Amazon can cause hurricanes in the North Atlantic. In a prior section, we mentioned that a small change in initial conditions of the planets can lead to very different configuration of the planets in the long term. These are examples of *chaotic systems*. Mathematical chaos is not really chaos, there is precise order behind the scenes. Everything is still deterministic. However a chaotic system is extremely sensitive to initial conditions. This also means even small errors induced via numerical approximation create large errors very quickly, so it is almost impossible to numerically approximate for long times. This is a large part of the trouble, as chaotic systems cannot be in general solved analytically.

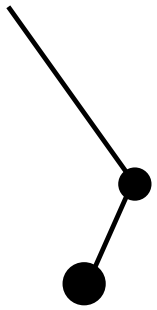
Take the weather, the most well-known chaotic system. A small change in the initial conditions (the temperature at every point of the atmosphere for example) produces drastically different predictions in relatively short time, and so we cannot accurately predict weather. And we do not actually know the exact initial conditions. We measure temperatures at a few points with some error, and then we somehow estimate what is in between. There is no way we can accurately measure the effects of every butterfly wing. Then we solve the equations numerically introducing new errors. You should not trust weather prediction more than a few days out.

Chaotic behavior was first noticed by Edward Lorenz* in the 1960s when trying to model thermally induced air convection (movement). Lorentz was looking at the relatively simple system:

$$x' = -10x + 10y, \qquad y' = 28x - y - xz, \qquad z' = -\frac{8}{3}z + xy.$$

A small change in the initial conditions yields a very different solution after a reasonably short time.

A simple example the reader can experiment with, and which displays chaotic behavior, is a double pendulum. The equations for this setup are somewhat complicated, and their derivation is quite tedious, so we will not bother to write them down. The idea is to put a pendulum on the end of another pendulum. The movement of the bottom mass will appear chaotic. This type of chaotic system is a basis for a whole number of office novelty desk toys. It is simple to build a version. Take a piece of a string. Tie two heavy nuts at different points of the string; one at the end, and one a bit above. Now give the bottom nut a little push. As long as the swings are not too big and the string stays tight, you have a double pendulum system.



*Edward Norton Lorenz (1917–2008) was an American mathematician and meteorologist.

5.5.1 Duffing equation and strange attractors

Let us study the so-called *Duffing equation*:

$$x'' + ax' + bx + cx^3 = C\cos(\omega t).$$

Here a, b, c, C, and ω are constants. Except for the cx^3 term, this equation looks like a forced mass-spring system. The cx^3 means the spring does not exactly obey Hooke's law (which no real-world spring actually does obey exactly). When c is not zero, the equation does not have a closed form solution, so we must resort to numerical solutions, as is usual for nonlinear systems. Not all choices of constants and initial conditions exhibit chaotic behavior. Let us study

$$x'' + 0.05x' + x^3 = 8\cos(t).$$

The equation is not autonomous, so we cannot draw the vector field in the phase plane. We can still draw the trajectories. In Figure 5.18 we plot trajectories for t going from 0 to 15, for two very close initial conditions $(2, 3)$ and $(2, 2.9)$, and also the solutions in the (x, t) space. The two trajectories are close at first, but after a while diverge significantly. This sensitivity to initial conditions is precisely what we mean by the system behaving chaotically.

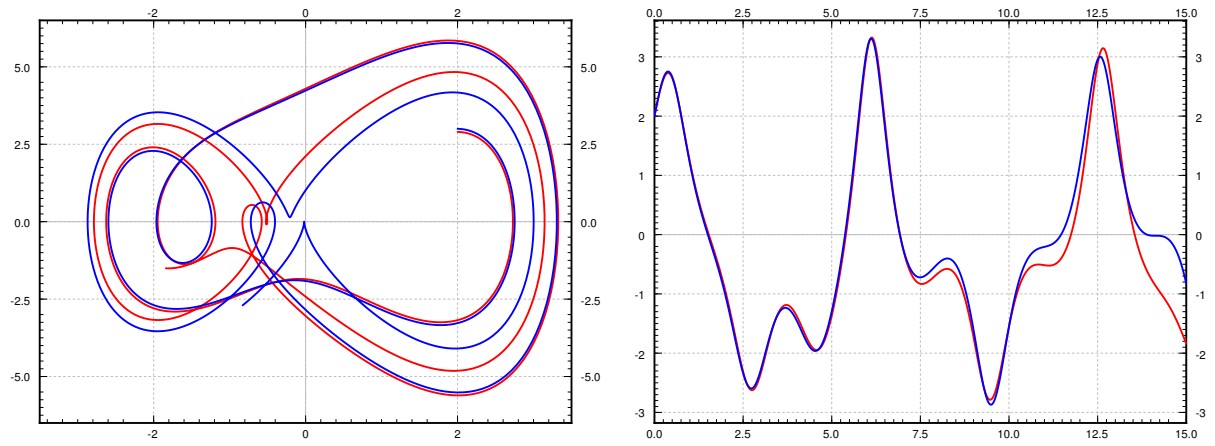


Figure 5.18: On left, two trajectories in phase space for $0 \le t \le 15$, for the Duffing equation one with initial conditions $(2, 3)$ and the other with $(2, 2.9)$. On right the two solutions in (x, t)-space.

Let us see the long term behavior. In Figure 5.19 on the next page, we plot the behavior of the system for initial conditions $(2, 3)$ for a longer period of time. It is hard to see any particular pattern in the shape of the solution except that it seems to oscillate, but each oscillation appears quite unique. The oscillation is expected due to the forcing term. We mention that to produce the picture accurately, a ridiculously large number of steps* had to be used in the numerical algorithm, as even small errors quickly propagate in a chaotic system.

*In fact for reference, 30,000 steps were used with the Runge–Kutta algorithm, see exercises in § 1.6.

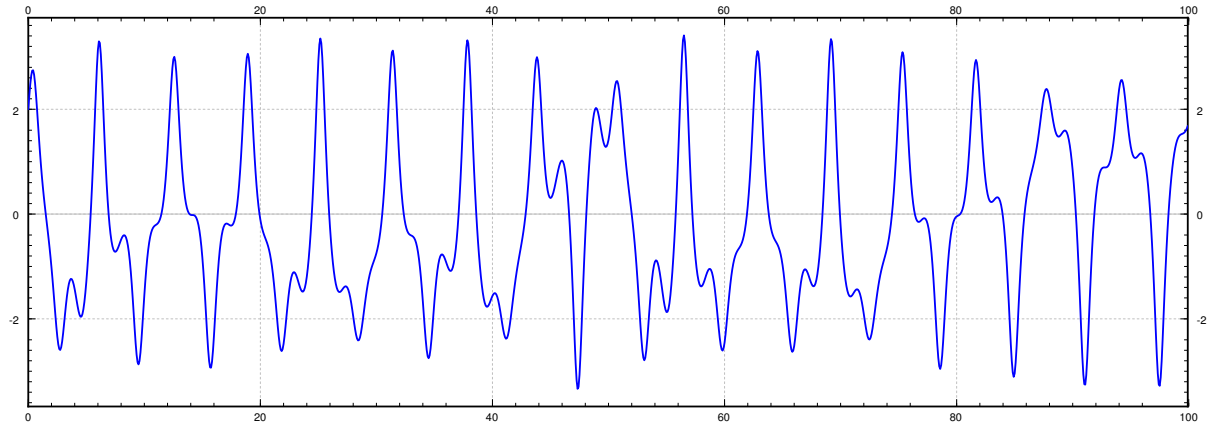


Figure 5.19: The solution to the given Duffing equation for t from 0 to 100.

It is very difficult to analyze chaotic systems, or to find the order behind the madness, but let us try to do something that we did for the standard mass-spring system. One way we analyzed the system is that we figured out what was the long term behavior (not dependent on initial conditions). From the figure above, it is clear that we will not get a nice exact description of the long term behavior for this chaotic system, but perhaps we can find some order to what happens on each "oscillation" and what do these oscillations have in common.

The concept we explore is that of a *Poincaré section**. Instead of looking at t in a certain interval, we look at where the system is at a certain sequence of points in time. Imagine flashing a strobe at a fixed frequency and drawing the points where the solution is during the flashes. The right strobing frequency depends on the system in question. The correct frequency for the forced Duffing equation (and other similar systems) is the frequency of the forcing term. For the Duffing equation above, find a solution $\big(x(t), y(t)\big)$, and look at the points

$$\big(x(0), y(0)\big), \quad \big(x(2\pi), y(2\pi)\big), \quad \big(x(4\pi), y(4\pi)\big), \quad \big(x(6\pi), y(6\pi)\big), \quad \ldots$$

As we are really not interested in the transient part of the solution, that is, the part of the solution that depends on the initial condition, we skip some number of steps in the beginning. For example, we might skip the first 100 such steps and start plotting points at $t = 100(2\pi)$, that is

$$\big(x(200\pi), y(200\pi)\big), \quad \big(x(202\pi), y(202\pi)\big), \quad \big(x(204\pi), y(204\pi)\big), \quad \ldots$$

The plot of these points is the Poincaré section. After plotting enough points, a curious pattern emerges in Figure 5.20 on the following page (the left-hand picture), a so-called *strange attractor*.

Given a sequence of points, an *attractor* is a set towards which the points in the sequence eventually get closer and closer to, that is, they are attracted. The Poincaré section is not really the attractor itself, but as the points are very close to it, we see its shape. The strange

*Named for the French polymath Jules Henri Poincaré (1854–1912).

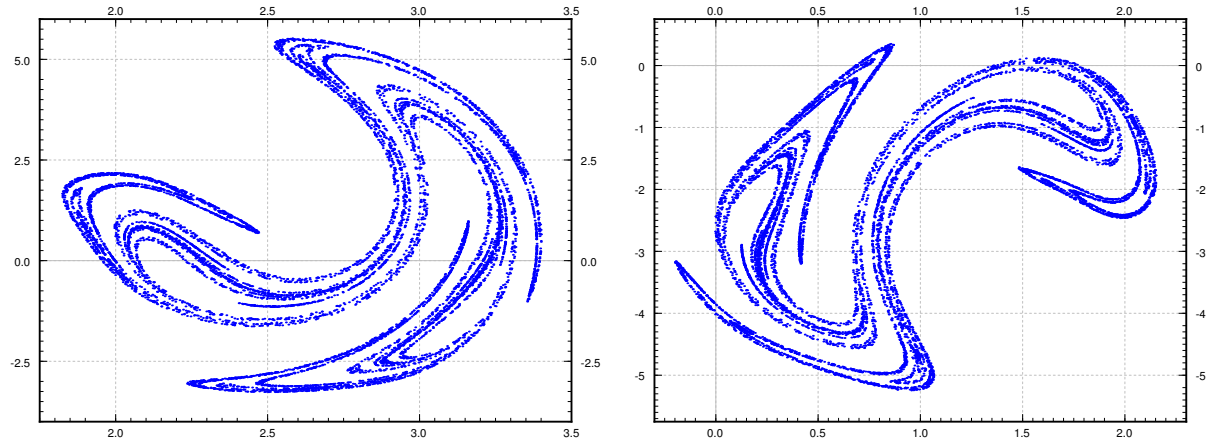


Figure 5.20: Strange attractor. The left plot is with no phase shift, the right plot has phase shift $\pi/4$.

attractor is a very complicated set. It has fractal structure, that is, if you zoom in as far as you want, you keep seeing the same complicated structure.

The initial condition makes no difference. If we start with a different initial condition, the points eventually gravitate towards the attractor, and so as long as we throw away the first few points, we get the same picture. Similarly small errors in the numerical approximations do not matter here.

An amazing thing is that a chaotic system such as the Duffing equation is not random at all. There is a very complicated order to it, and the strange attractor says something about this order. We cannot quite say what state the system will be in eventually, but given the fixed strobing frequency we narrow it down to the points on the attractor.

If we use a phase shift, for example $\pi/4$, and look at the times

$$\pi/4, \quad 2\pi + \pi/4, \quad 4\pi + \pi/4, \quad 6\pi + \pi/4, \quad \ldots$$

we obtain a slightly different attractor. The picture is the right-hand side of Figure 5.20. It is as if we had rotated, moved, and slightly distorted the original. For each phase shift you can find the set of points towards which the system periodically keeps coming back to.

Study the pictures and notice especially the scales—where are these attractors located in the phase plane. Notice the regions where the strange attractor lives and compare it to the plot of the trajectories in Figure 5.18 on page 422.

Let us compare this section to the discussion in § 2.6 about forced oscillations. Take the equation

$$x'' + 2px' + \omega_0^2 x = \frac{F_0}{m}\cos(\omega t).$$

This is like the Duffing equation, but with no x^3 term. The steady periodic solution is of the form

$$x = C\cos(\omega t + \gamma).$$

Strobing using the frequency ω, we obtain a single point in the phase space. The attractor in this setting is a single point—an expected result as the system is not chaotic. It was the opposite of chaotic: Any difference induced by the initial conditions dies away very quickly, and we settle into always the same steady periodic motion.

5.5.2 The Lorenz system

In two dimensions to find chaotic behavior, we must study forced, or non-autonomous, systems such as the Duffing equation. The Poincaré–Bendixson Theorem says that a solution to an autonomous two-dimensional system that exists for all time in the future and does not go towards infinity is periodic or tends towards a periodic solution. Hardly the chaotic behavior we are looking for.

In three dimensions even autonomous systems can be chaotic. Let us very briefly return to the Lorenz system

$$x' = -10x + 10y, \qquad y' = 28x - y - xz, \qquad z' = -\frac{8}{3}z + xy.$$

The Lorenz system is an autonomous system in three dimensions exhibiting chaotic behavior. See the Figure 5.21 for a sample trajectory, which is now a curve in three-dimensional space.

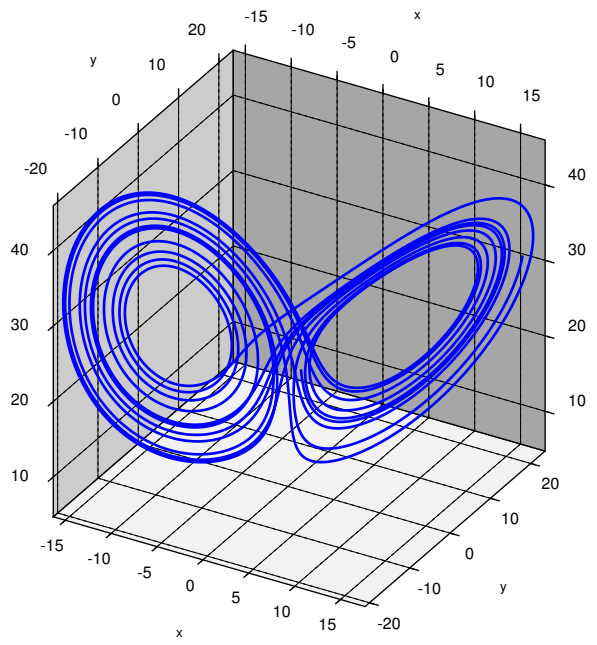


Figure 5.21: A trajectory in the Lorenz system.

The solutions tend to an *attractor* in space, the so-called *Lorenz attractor*. In this case no strobing is necessary. Again we cannot quite see the attractor itself, but if we try to follow a solution for long enough, as in the figure, we get a pretty good picture of what the attractor looks like. The Lorenz attractor is also a strange attractor and has a complicated fractal

structure. And, just as for the Duffing equation, what we want to draw is not the whole trajectory, but start drawing the trajectory after a while, once it is close to the attractor.

The path of the trajectory is not simply a repeating figure-eight. The trajectory spins some seemingly random number of times on the left, then spins a number of times on the right, and so on. As this system arose in weather prediction, one can perhaps imagine a few days of warm weather and then a few days of cold weather, where it is not easy to predict when the weather will change, just as it is not really easy to predict far in advance when the solution will jump onto the other side. See Figure 5.22 for a plot of the x component of the solution drawn above. A negative x corresponds to the left "loop" and a positive x corresponds to the right "loop".

Most of the mathematics we studied in this book is quite classical and well understood. On the other hand, chaos, including the Lorenz system, continues to be the subject of current research. Furthermore, chaos has found applications not just in the sciences, but also in art.

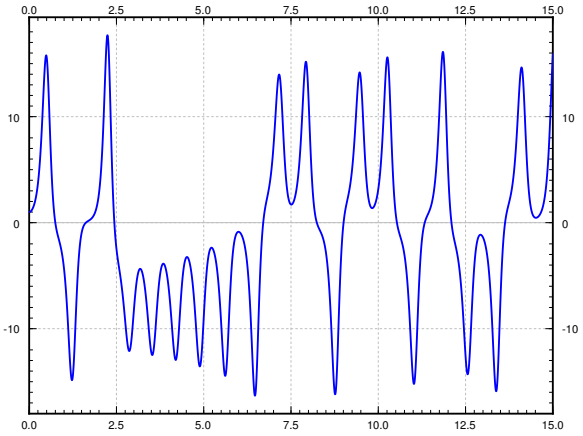


Figure 5.22: Graph of the $x(t)$ component of the solution.

5.5.3 Exercises

Exercise* 5.5.1** (*):*** *Find critical points of the Lorenz system and the associated linearizations.*

Exercise* 5.5.2:* *For the non-chaotic equation $x'' + 2px' + \omega_0^2 x = \frac{F_0}{m}\cos(\omega t)$, suppose we strobe with frequency ω as we mentioned above. Use the known steady periodic solution to find precisely the point which is the attractor for the Poincaré section.*

Exercise* 5.5.3** (project):*** *Construct the double pendulum described in the text with a string and two nuts (or heavy beads). Play around with the position of the middle nut, and perhaps use different weight nuts. Describe what you find.*

Exercise* 5.5.4** (project):*** *A simple fractal attractor can be drawn via the following chaos game. Draw the three vertices of a triangle and label them, say p_1, p_2 and p_3. Draw some random point p (it does not have to be one of the three points above). Roll a die to pick of the p_1, p_2, or p_3 randomly (for example 1 and 4 mean p_1, 2 and 5 mean p_2, and 3 and 6 mean p_3). Suppose we picked p_2, then let p_{new} be the point exactly halfway between p and p_2. Draw this point and let p now refer to this new point p_{new}. Rinse, repeat. Try to be precise and draw as many iterations as possible. Your points will be attracted to the so-called* Sierpinski triangle*. A computer was used to run the game for 10,000 iterations to obtain the picture in Figure 5.23.*

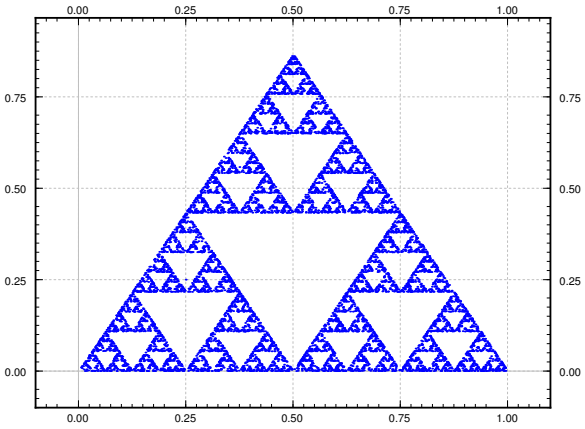


Figure 5.23: 10,000 iterations of the chaos game producing the Sierpinski triangle.

Exercise* 5.5.5** (computer project):*** *Use a computer software (such as Matlab, Octave, or perhaps even a spreadsheet), plot the solution of the given forced Duffing equation with Euler's method. Plotting the solution for t from 0 to 100 with several different (small) step sizes. Discuss.*

Appendix A

Introduction to MATLAB

This document is meant to provide a review of some of the main skills and techniques in MATLAB that are necessary to complete the various MATLAB assignments throughout the course. In addition, these skills will be useful when attempting to use MATLAB, both for illustrating problems in differential equations and for solving other types of problems that can be analyzed using this software.

A.1 The MATLAB Interface

There are many components to the MATLAB interface, and the way that the window is organized can be fully customized. There are four main components of this interface.

1. Current Folder window. This shows the current folder in which MATLAB is running. This determines what files that MATLAB currently has access to and what functions and methods can be called.

2. Editor window. This is the main code-editing window, where script files can be written, edited, saved, and run.

3. Command window. This is where individual lines of code can be entered to see how they work.

4. Workspace window. This shows a list of all variables that currently exist, as well as their values or sizes.

All four of these components are very useful in organizing thoughts and programming practices while using MATLAB. Both the Default layout and Two-Column layout (as of MATLAB R2019b) contain all four of these windows in different locations. Either of these will work for programming in MATLAB, as well as any modifications of them. The current format can be saved using Layout - Save Layout if needed.

A.1.1 File Structure

The main type of file used in MATLAB is the Script file. These are saved as '*.m' files and can represent both stand-alone executable files and functions that can be called from

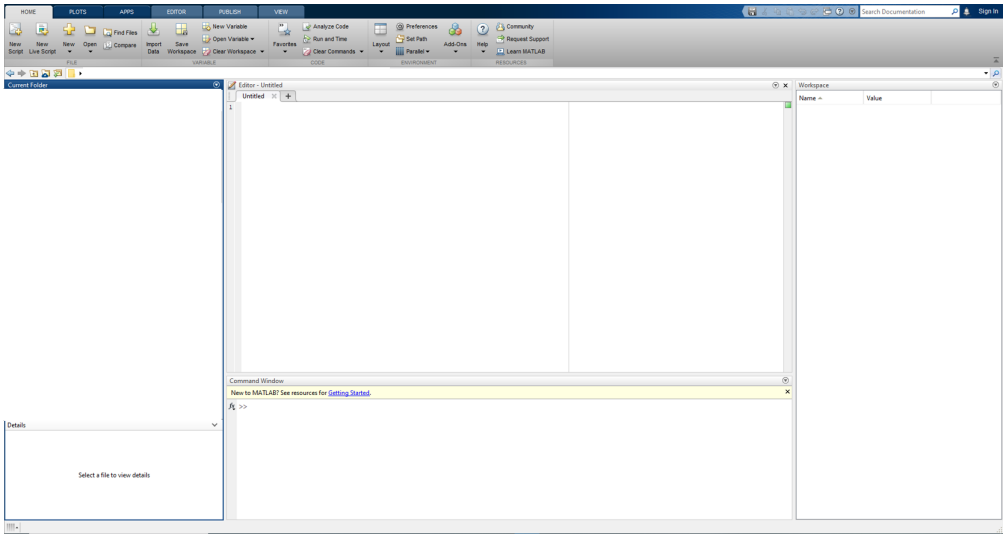


Figure A.1: The default layout provided by MATLAB.

other scripts. For running simple, one-line expressions or debugging code, the Command Window and the command line prompt can be useful. However, for anything more involved and complicated than that, the script editor should be used instead.

In writing a script file or using the Command window, the Current Folder window shows all of the files in the current directory. These are all of the files that MATLAB has access to while running a MATLAB file that it saved in that folder. This means that if a script wants to call a method, it either needs to be a built-in method or a function file that is contained within the same script file or the Current Folder. For more information about writing functions, see Section A.4.

To use script files, multiple lines of code can be entered in a row, and MATLAB will execute them in sequence when the "Run" button is clicked. This button is in the "Editor" tab at the top of the screen.

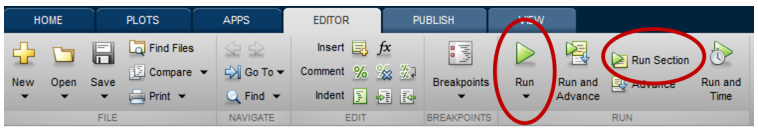


Figure A.2: Location of the Run buttons on the MATLAB interface.

MATLAB Live Scripts can also be used to do very similar things, with some additional benefits. These allow the MATLAB code to be viewed side-by-side with the output, as well as an easy export to PDF functionality. These are saved as '*.mlx' files. These work the same way as scripts in terms of how code is written, and allow the user to mix between text (which can be resized and formatted) and code. For more information on Live Scripts, see the website https://www.mathworks.com/help/matlab/matlab_prog/what-is-a-live-script-or-function.html.

Live Scripts also have the ability to put section breaks between different pieces of code and then run individual sections using the "Run Section" button at the top of the editor.

With Live Scripts, it is necessary to run the entire code (by clicking the run button) before exporting as a PDF in order to get the correct images and outputs in the final PDF. To export, go to Save at the top of the screen, click the down arrow under it, and select "Export to PDF" **after** running the code to regenerate all of the images.

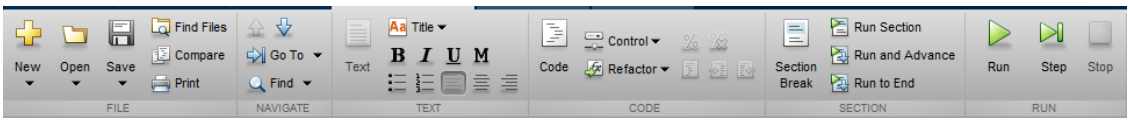


Figure A.3: Header Bar for the MATLAB Live Script Interface.

A.2 Computation in MATLAB

MATLAB can do many of the simple computational operations that would be expected from a calculator. It is easiest to see these operations by using the Command Window, but they can also be implemented in scripts if desired. Addition and subtraction work in standard ways. In the command line, typing

```
2 + 3
```

and pressing ENTER will give an output of

```
ans =
    5
```

showing the answer of this computation. For any computation or line of code, putting a semi-colon (;) at the end will suppress the output, in that typing

```
2 + 3;
```

will not show any output. However, MATLAB did do the computation, which can be shown by storing this output in a variable and doing something with it later.

Multiplication and division, and by extension powers, can work differently in MATLAB. As MATLAB is built around using matrices for calculations and is optimized for this approach, the program interprets all multiplication, division, and exponentiation in terms of matrices as a default. Both components of the multiplication are simple scalars (numbers), then this is fine. The '*' symbol works for multiplication in this context:

```
>> 4*6
ans =
    24
```

as well as using '/' for division and '^' for exponentiation. Issues may arise when the code wants to compute products or powers of multiple values at the same time. Many MATLAB

built-in functions will automatically combine multiple of the same type of calculation into a 'vectorized' calculation, where if the code wanted to compute the sum of two numbers a bunch of times, it would put all of these numbers into arrays and then add the two vectors together. This completes the task of adding all of the different pairs of numbers together, but saves time by not doing them all individually. This works great for addition and subtraction, because addition and subtraction of arrays or matrices is done element-wise, which is the exact operation we wanted to compute in the first place.

However, mutliplication is different. Matrix multiplication is a different operation that, in particular, is not element-wise multiplication. Beyond that, even if two matrices are the same size, it is possible that their product, in the normal matrix sense, is not defined. In MATLAB, the product

```
[1 2 3] * [4 3 2];
```

will return an error because the matrices are not the correct size. From a human point of view, the output desired from this code was likely `[4 6 6]`, the product of each term individually. To obtain this in MATLAB, we need the elementwise operations '.*', './' and '.^' for multipication, division, and exponentiation, respectively. Thus, the following computations can be made in MATLAB

```
>> [1 2 3] .* [4 3 2]
ans =
   [4 6 6]
>> [1 4 6].^2
ans =
    [1 16 36]
>> [5 4 2] ./ [10 2 6]
ans =
    [0.5 2 0.3333]
```

There are many built-in functions in MATLAB that can help with computation and algebra.

- `sqrt(x)` will compute the square root of a number x.
- `exp(x)` will compute e^x for e the base of the natural logarithm, and x any number. Note that MATLAB does not know the definition of e built-in, so it will either need to be defined (using `exp(1)`) or just use `exp()` whenever it is needed.
- `abs(x)` computes the absolute value of a number x.
- `log(x)` computes the natural logarithm of a number x. The functions `log2` and `log10` compute the log base 2 and log base 10 respectively.
- Trigonometric functions can also be computed with `sin(x)`, `cos(x)`, and `tan(x)`.

A.3 Variables and Arrays

As with other programming languages, MATLAB utilizes variables to store information and use it later. The name of variables in MATLAB must start with a letter, but the rest of the name can consist of letters, digits, or underscores. Variables should be named suggestively corresponding to what this information is or the way it will be used. Variables do not need to be created in advance, they are created when something is stored in the variable by putting the name on the left side of an equals sign, with the computation that gives rise to that variable on the right. Even though the output is suppressed, the line

```
val = 2+3;
```

will store the value 5 in the variable `val`, where it can be used later. For example,

```
>> val * 4
ans =
    20

>> val^2 + 2
ans =
    27
```

However, trying to use a variable name without defining it first will cause MATLAB to give an error:

```
>> r
Undefined function or variable 'r'.
```

As variables do not need to be created or instantiated before they are used, any variable can store any type of information. Two of the most common ones are numbers (double precision) or strings.

```
numVar = sqrt(15);
strVar = ``Hello World!'';
```

Strings can be stored using either single or double quotes. Strings also have a lot of useful operations that can be used to make some MATLAB programs run more simply, but they are beyond the scope of this introduction. For information about what can be done with strings, see the MATLAB documentation https://www.mathworks.com/help/matlab/ref/string.html.

Another common variable data type that MATLAB is very comfortable with is arrays. As described previously, MATLAB defaults to matrices when considering multiplication and exponentiation operations. Arrays can be created using square brackets, with either spaces or commas between the entries.

```
A = [2,4,6];
B = [1 3 5];
```

These create horizontal arrays. Vertical arrays can also be created using semi-colons between each entry, and these can be combined with horizontal arrays to create a matrix, or rectangular array of values.

```
C = [5;7;8];
M = [1,2,3;5,6,7];
```

In these examples, A and B will be row arrays (or row vectors) with 3 elements, C will be a column vector with 3 elements, and M will be a matrix with two rows and three columns. For most situations that don't involve matrices, row and column vectors will work equivalently, so either one can be used. Once matrices are involved, it matters which one is chosen, because MATLAB will multiply matrices and vectors in the same way that would be carried out mathematically, which means the dimensions need to match.

To access elements of a matrix, parentheses are used. Unlike other programming languages, MATLAB starts indexing elements at 1, not zero. That is, with the above variables `C(2) = 7`, since 7 is the second element of the array C. In terms of accessing elements of matrices, the first index is the row and the second is the column.

```
>> M = [1,2,3;5,6,7];
>> M(1,1)
ans =
     1

>> M(1,3)
ans =
     3

>> M(2,1)
ans =
     5
```

The matrix (and vectors) do have limits on how big they are, and attempting to access an element outside of that range will cause MATLAB to give an error.

```
>> M(3,1)
Index in position 1 exceeds array bounds (must not exceed 2).
```

Among many other possible variables, another type that can be stored is a handle to a function. How to use functions will be described in Section A.4. The fact that all of these different data types can be stored in variables, with no real indication as to which type a given variable is, means it is critical to name variables carefully with what they correspond to.

A.4 Functions and Anonymous Functions

A key component to programming in MATLAB is the idea of functions. These are programming objects that will accept a number of inputs (called *arguments*) and perform a given set of operations on those arguments, returning some set of ouputs back to the main program. These are mainly used to group code together that has a given purpose and can be called to carry out that purpose on a variety of outputs. An example of a built-in function like this is `sum(V)`. This function takes in a linear array and will return the number that is the sum of all of the elements in the array (if the array is multi-dimensional, it will only sum along one dimension). This is a piece of code that could be written fairly easily; it would just involve taking the array, looping through it and adding up the value at each index. However, putting it into a function allows it to be called more simply in one line, allowing the main script to focus on the task at hand.

There are two main ways that functions can be written in MATLAB. Functions can either be written at the bottom of the MATLAB script where they will be used or they can be written in their own separate script file. If written in a separate file, there can only be one function in each file, and the name of the file (once saved) must match the name given to the function. To write a function, the reserved word 'function' is used:

```
function [a,b] = testFunction(x, y, z)
    % Code here
end
```

Note: If this is done in a script by itself, the function line must be the first line of the code. There can be no code or comments above this line.

In this case, the function takes in three inputs and returns two outputs. When writing the code inside the function, the three inputs will be called x, y, and z, and in order to tell the program what to send back to wherever this function was called, those outputs should be stored in variables a and b. For example, a function that takes in three numbers and returns their sum in the first output and the product in the second would look like

```
function [a,b] = testFunction(x, y, z)
    a = x+y+z;
    b = x*y*z;
end
```

and that would work just fine. However, if any other MATLAB methods were going to use this function, there is a chance they would try to pass in array inputs. If so, then there would be an error in computing `b`, because those products would not be defined. The easiest way to fix this would be to use element-wise products, giving a function that looks like

```
function [a,b] = testFunction(x, y, z)
    a = x+y+z;
    b = x.*y.*z;
end
```

These functions can be as complicated as necessary, including graphs, loops, calls to other functions, and many different components. However, if the function needed is a simple mathematical function, then this can be written in an shorter way with anonymous functions. For example, if the function $f(x, y) = x^2 + 4xy + y^2$ needed to be coded, it could be written as

```
f = @(x,y) x.^2 + 4.*x.*y + y.^2;
```

and this will now make `f` a handle to the function that does exactly what is desired. If a later line of code is

```
>> f(2,1)
ans =
    13
```

the function value will be computed at the desired point. Notice the use of element-wise operations again in this function definition to ensure that it will also work on array inputs. This works for these simple kinds of functions, and can be easier than adding an entire new function to the script file.

Overall, the following two function definitions are *almost* equivalent.

```
fShort = @(x,y) x.^2 + y.^2;
```

```
function z = fLong(x,y)
z = x.^2 + y.^2;
end
```

The only difference arises when trying to use these functions in built-in or written methods that require a handle to a function. The '@' symbol at the beginning of the anonymous function indicates that the thing being defined (`fShort`) is a handle to a function that takes two inputs and computes an output from it. On the other hand, the definition of `fLong` is a function that does this, and is not a handle to that function. To fix this, an '@' symbol needs to be put in-front of `fLong` before using it in one of these methods. As an example `ode45` is a method that numerically computes the solution to a differential equation, and it requires a function handle in the first argument. So, the code

```
ode45(fShort, [0, 3], 1)
```

runs fine. However,

```
ode45(fLong, [0, 3], 1)
```

throws an error about there being not enough inputs for `fLong`. This is because whenever MATLAB sees `fLong`, it is expecting to see two inputs next to it. This is not the case for `fShort` because of the way it was defined. To remedy this, the code needs to be written

```
ode45(@fLong, [0, 3], 1)
```

and then it will execute the same as the first line.

With any of these functions, it is possible to restrict variables and get new functions. This can be fairly easily done with the same setup as for anonymous functions. The line of code

```
fNew = @(y) fShort(1,y)
```

will create a new handle for a function of one variable that is `fShort` when the x value is fixed to be 1. The exact same code will work for `fLong` as you are giving it two inputs.

A.5 Loops and Branching Statements

The code written in a MATLAB script will always proceed in order from one line to the next unless there is some alteration to the flow using loops or branching (if) statements.

A.5.1 For Loops

For loops are a form of iterative programming, where MATLAB will run the same bit of code multiple times with an iterative parameter that can change certain things about the code. If there is an element of the program that needs to carry out a process several times in a row, particularly using the previous step to compute the one after it, a for loop might be the best structure to use. A sample for loop has the following form:

```
for counter = 1:1:10
    % CODE HERE
end
```

In this line, `counter` is the variable that is getting incremented over the list. The rest of that line says that counter starts at 1, increments by 1 each loop, and stops after 10. A line of the form `counter = 2:5:34` will start at 2, increment by 5 each loop, and stop once the counter gets above 34, so after the iteration when `counter = 32`.

In order to loop through an array of values, it is useful to figure out the size of the array and use that to determine how many times the loop should be run. This sort of programming will allow your code to work for a variety of different inputs, no matter the size. This can be done with code like this.

```
v = [1,2,3,4,5]; % This will be your list of values
for counter = 1:1:length(v)
    x = v(counter)^2
end
```

To find how many elements are in an array, the `length` function will work for a linear array. If the array is more complicated, the `size` function can be used. This will give a list of values saying how large the array is in each dimension.

MATLAB also has `while` loops, which allow a loop to run up until a condition becomes false. This is better than for loops in specific situations, but either one can be used. For the code developed here, for loops will be just as easy to write as while loops.

A.5.2 If Statements

If statements, or conditional statements, allow certain parts of code to be executed only if a certain condition is met. For instance, something like

```
if counter < 5
    % CODE HERE
end
```

will only execute if the counter is less than 5, and

```
if mod(counter,2) == 0
    % CODE HERE
end
```

will only run if counter is even, that is, if the remainder when dividing counter by 2 is zero. Notice that `==` is used for comparison here to check if two things are equal, while `=` is used for variable assignment. The condition part of an if statement can be anything that gives back a true or false result. For math operations, these can be any inequalities ($\leq$, $<$, $\geq$, $>$) or `==` for testing inequality. The operator $\sim$ is used for "not", in that $a \sim= b$ will be true if a is not equal to b, and false if they are the same. Outside of numbers, there are other MATLAB methods that will give true or false answers. These can be things like comparing strings, but this is beyond the code developed here.

A.6 Plotting in MATLAB

Graphing in MATLAB always involves plotting a set of points, but these can be fairly easily generated from functions as well. For example

```
xPts = [1,2,3,4,5];
fx = @(x) x.^2 + 2;
yPts = [2,3,2,3,1];
figure(1);
plot(xPts, yPts);
figure(2);
plot(xPts, fx(xPts));
```

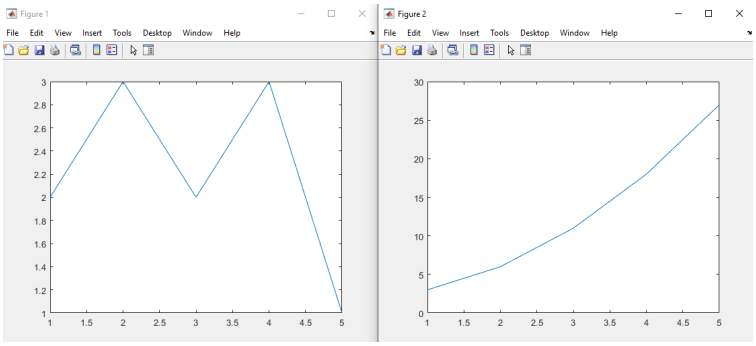


Figure A.4: Output from MATLAB plotting two graphs.

will generate two figures, referred to by the lines `figure(1)` and `figure(2)`, and allow the two graphs to be simultaneously drawn without overlapping each other. Any time MATLAB draws a plot (with the `plot` command) it will overwrite any plot that is already on the target figure. In order to put multiple plots on the same figure, the `hold on;` and `hold off;` commands can be used.

```
xPts = linspace(1,5,100);
fx = @(x) x.^2 + 2;
gx = @(x) x.^2 - 3*x + 7;
figure(1);
hold on;
plot(xPts, fx(xPts));
plot(xPts, gx(xPts));
hold off;
```

The `linspace` generates a list of 100 equally spaced values between 1 and 5 for plotting purposes. It gives an easy way to generate a lot of input values for plotting a smooth-looking graph. It also emphasizes the need to use the element-wise operations in these functions to make sure they all compute correctly.

There are many additional options that can be passed to the `plot` method in order to change the color, shape, and size of the plot. For these options, refer to the MATLAB documentation on the plot function at https://www.mathworks.com/help/matlab/ref/plot.html.

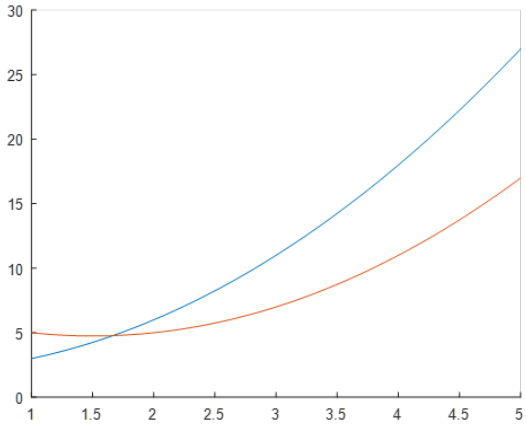


Figure A.5: Output from MATLAB plotting two functions on the same axes.

A.7 Supplemental Code Files

There are eleven supplemental code files provided. In order to use these files in a script or a Live Script, they must be placed in the same folder as the script file, so that the Current Folder window contains both the file being executed and all of these function files. Another option would be to store all of these function files in a single folder, navigating to that folder in the MATLAB Current Folder window, right-clicking on the folder, and selecting "Add to Path." The first of these is more recommended, but the second can also work if there is a common repository to store all of the users custom MATLAB functions. The function headers are given below along with a brief description of their use.

```
function quiver244(f, t_min, t_max, y_min, y_max, col)
% quiver244.m
% Author: Matt Charnley
%
% This function draws a quiver plot for the ODE dy/dt = f(t,y) for
% t_min <= t <= t_max and y_min <= y <= y_max. The function f should be
% passed in as an anonymous function, of two variables or as a function
% handle
%
% The function draws this quiver plot in color col and saves it on the
% current figure, and generates a normalized version
% (all vectors are the same length) as the next figure,
% so that it can be accessed outside of this function.
% For this second figure, the magnitude of the arrows does not mean
% anything, but it is easier to see the direction of them.
% so that it can be accessed outside of this function. It will start with
% hold on; and end with hold off;, so the figure needs to be cleared in the
% main file if needed.
```

The main point of this function is to simplify the process of drawing quiver plots. The code here takes care of the difficulties that arise from the built-in `quiver` function in MATLAB and allows the user to input the right-hand side of a first order ODE and generate quiver plots. It will draw a quiver plot in the first figure, and a normalized quiver plot (all vectors the same length) in the second figure. It can sometimes be easier to see the general trajectory of solutions from the normalized figure, so both graphs are provided. All of the plotting commands use the `hold` commands so that they will not overwrite anything on the desired figures. This allows the overlaying of multiple plots, but means that the code calling this method must clear the figure if it needs to be cleared.

This code can be used as

```
f = @(t,y) t - exp(y);
quiver244(f, 0, 5, -6, 6, 'b');
```

```
quiver244(@f2, 0, 5, -6, 6, 'b');

function z = f2(t,y)
    z = t - exp(y);
end
```

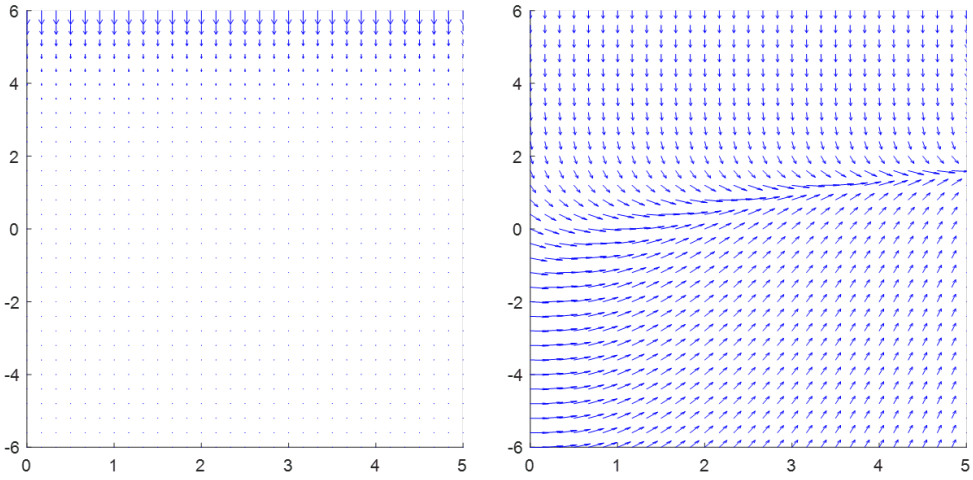


Figure A.6: Sample output from the `quiver244` *function.*

In each case, the `'b'` indicates that the quiver plot will be drawn in blue, and the 1 before that indicates that the two plots will be drawn on figures 1 and 2.

```
function samplePlots244(f, t_min, t_max, y_min, y_max, t_0, y_0, col)
% This function takes the ODE dy/dt = f(t,y) and plots sample solutions
% with initial value (t_0, y_0). It uses ode45 to sketch out the solutions.
% t_0 must be between t_min and t_max. It also truncates the function f so
% that functions will not go off to infinity, causing this to work properly
% on vector inputs for initial conditions in y. The input y_0 can be a
↪  vector
% of initial values, and this function will plot a curve
% for each of those values. If using a vector of initial
% conditions, the function must be written with vector element-wise
% operations.
```

This function follows the same setup as `quiver244`, but draws sample trajectories of the solution instead of the quiver plot. It will take initial conditions as (t_0, y_0). For a single t_0, a vector of initial y_0 values can be passed in and the function will work correctly. This function can be used as

```
f = @(t,y) y.*(y-5).*(y+6);
samplePlots244(f, -1, 6, -7, 6, 0, [-1,0.5,4,5], 'r')
```

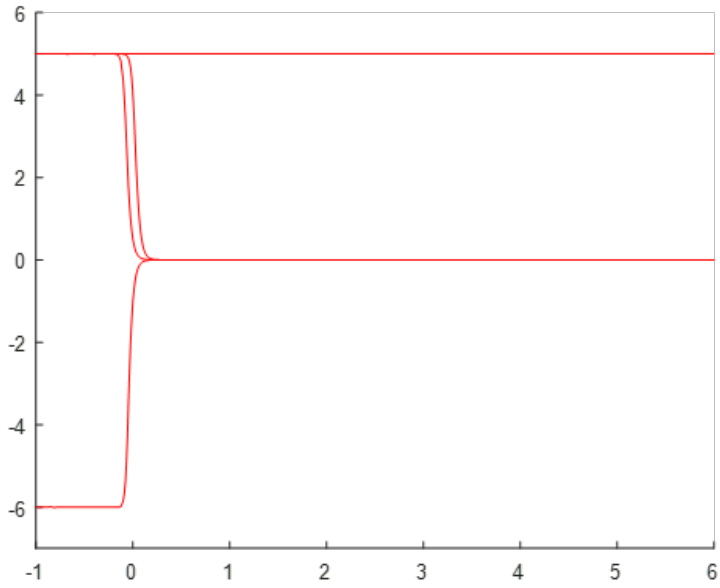


Figure A.7: Sample output from the `samplePlots244` function.

The `'r'` here indicates that this plot will be drawn in red and put on figure 2. If this is combined with the `quiver244` method, then it will overlay these red curves on top of the quiver plot drawn on figure 2.

```
function bifDiag244(f, a_min, a_max, y_min, y_max)
% This function draws a bifurcation diagram for the ode dy/dt = f(alpha, y)
% with parameter alpha running from a_min to a_max. The axes are
% constrained to be from a_min to a_max in the horizontal direction and
% y_min to y_max in the vertical direction.
%
% The black marks are for equilibrium solutions, the blue regions are where
% the solution will tend upwards, and the red region is where it will tend
% downwards.
```

This function will draw a bifurcation diagram for the given differential equation. **Note:** This function will need the optimization tool-box add-on for MATLAB in order to run correctly. As with the previous methods, it will not overwrite the figure. Example implementation:

```
f = @(a,y) y.^2 - a.^2;
bifDiag244(f, -3, 3, -5, 5, 3);
```

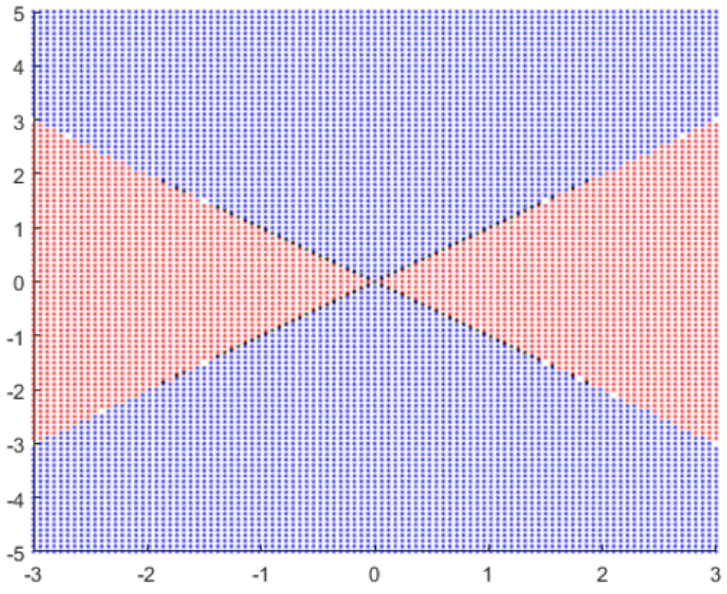


Figure A.8: Sample output from the `bifDiag244` *function.*

```
function quiver2D244(f,g, x_min, x_max, y_min, y_max, col)
% quiver2D244.m
% Author: Matt Charnley
%
% This function draws a quiver plot for the ODE dx/dt = f(x,y), dy/dt =
↪  g(x,y) for
% x_min <= x <= x_max and y_min <= y <= y_max. The functions f and g should
↪  be
% passed in as an anonymous functions, f = @(x,y) ...
%
% The function draws this quiver plot in color col in the current figure
% and generates a normalized version (all vectors are the same length)
% as the next figure, so that it can be accessed outside of this function.
% For this second figure, the magnitude of the arrows does not mean
% anything, but it is easier to see the direction of them.
%
% It will start with
% hold on; and end with hold off;, so the figure needs to be cleared in the
% main file if needed.
```

This function does the same concept as `quiver244` but for the autonomous system of differential equations

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

Example implementation:

```
f = @(x,y) 3.*x - 2.*x.*y;
g = @(x,y) 2.*y - 3.*x.*y;
quiver2D244(f,g, 0, 5, 0, 5, 'g');
```

```
function phaseLine(f, ymin, ymax)
% This function draws a representation of the phaseline for the
% differential equation dy/dt = f(y). The graph is drawn from ymin to ymax,
% and looks for solutions to f(y) = 0 in that region to find equilibrium
% solutions. This requires the Optimization Toolbox fsolve to run
% correctly.
```

This function draws a representation of the phase line for an autonomous first order differential equation $\frac{dy}{dt} = f(y)$ from y_{min} to y_{max}. Example implementation:

```
f = @(y) y.*(y-3).*(y+2);
phaseLine(f, -4, 5);
```

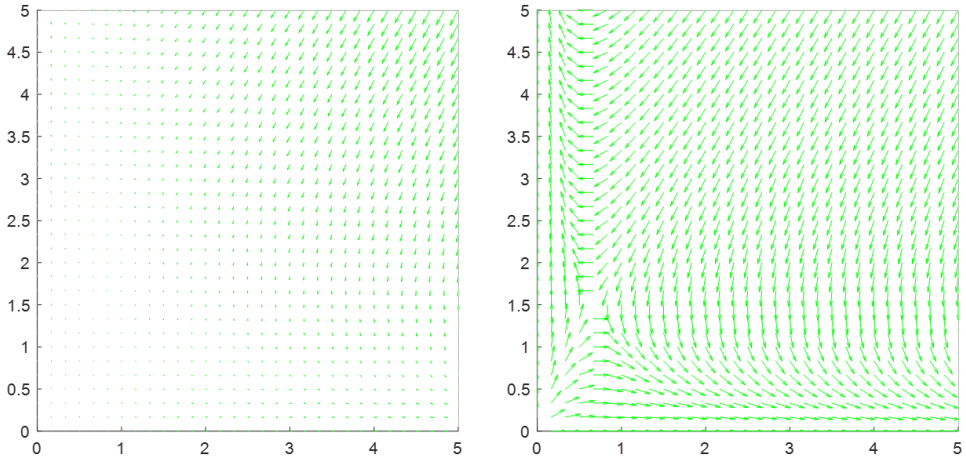


Figure A.9: Sample output from the `quiver2D244` *function.*

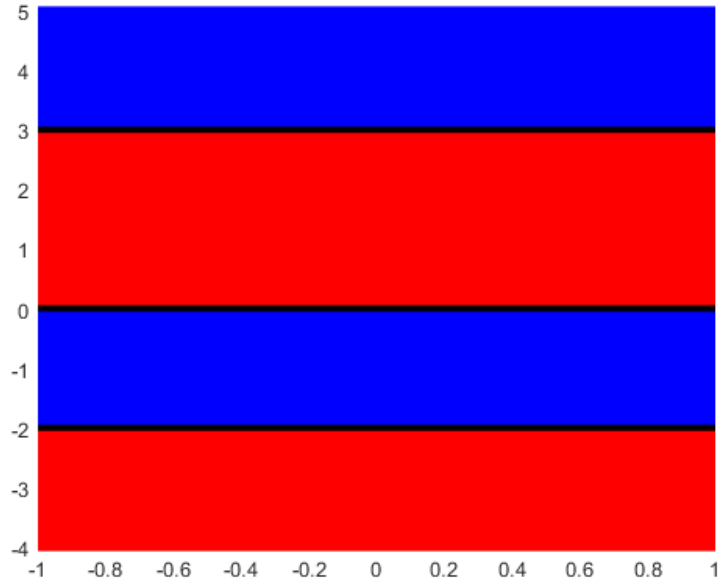


Figure A.10: Sample output from the `phaseLine` *function.*

```
function phasePortrait244(F, G, xmin, xmax, ymin, ymax, tmin, tmax, x0, y0)
% This function draws a 2 dimensional phase portrait for the system dx/dt =
% F(x,y) and dy/dt = G(x,y). The phase portrait will be draw with x bounds
% xmin <= x <= xmax and ymin <= y <= ymax. It is assumed that the initial
% conditions x0 and y0 are at £t=0£, with tmin <= 0 and tmax >=0. x0 and y0
% can be inputted as vectors that are the same length, and a sample curve
% will be drawn for each of them. The black dot will always be plotted at
↪  tmin.
```

This function draws a phase portrait for the two-component autonomous system $\frac{dx}{dt} = F(x, y)$ and $\frac{dy}{dt} = G(x, y)$. The axes are fixed at $x_{min} \leq x \leq x_{max}$ and $y_{min} \leq y \leq y_{max}$. Solution curves are drawn starting at the (potential list of) points x_0 and y_0, and will assume these happen at $t = 0$. The curves are drawn from t_min to t_max, and there will be a black dot plotted at t_min to indicate the direction of flow. Example implementation:

```
f = @(x,y) 2.*x - 3.* y;
g = @(x,y) -3.*x + y;
phasePortrait244(f, g, -3, 3, -3, 3, -2, 2, [1, 0, -1, 1, 0, -1],
↪  [1,1,1,-1,-1,-1]);
```

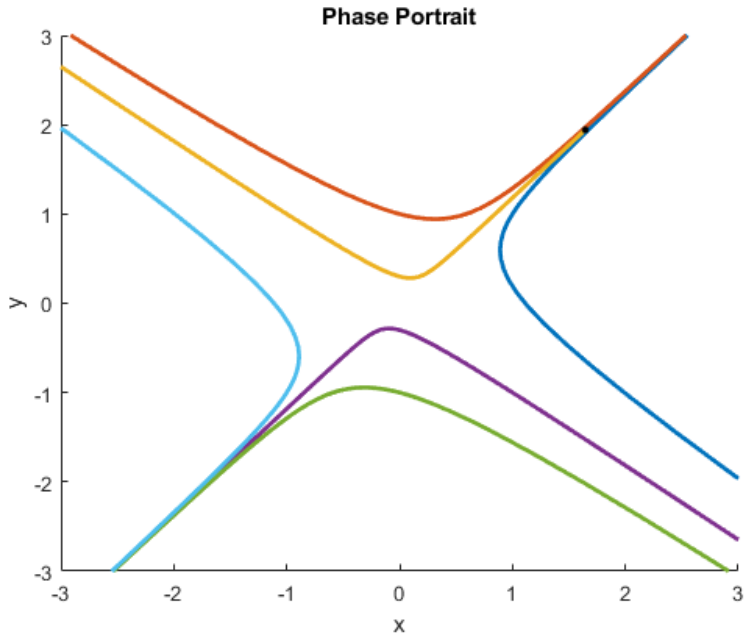


Figure A.11: Sample output from the `phasePortrait` *function.*

```
function [t, y] = rungeKuttaMethod(f, dt, Tf, T0, y0)
% This method solves the ODE dy/dt = f(t, y) using the Runge Kutta method
% from t=T0 to t = Tf with time step dt and initial condition y0 at t = T0.
% In this case, f should be a function of two variables, t
% (time) and y.
```

```
function [t,y] = rungeKuttaSystemMethod(f, T0, Tf, dt, y0)
% This method solves the ODE system dy/dt = f(t, y) using the Runge Kutta
↪  method
% from t=T0 to t = Tf with time step dt and initial condition y0 at t = T0.
% In this case, f should be a vector valued function of two variables, t
% (time) and y (n-dimensional vector of unknowns). The length of the vector
% y0 will determine the size of the system.
```

These two methods use the Runge-Kutta method to numerically solve the differential equation $\frac{dy}{dt} = f(t, y)$ or the system $\frac{d\vec{x}}{dt} = F(t, \vec{x})$. It will return the list of t and y values that are generated by this method.

```
function [S,I,R] = SIRModel_244(r, c, ICs, Tf)
% This code runs an SIR model for disease spread. The system of differential
↪  equations used here is
%   S' = -r*S*I
%   I' = r*S*I - cI
%   R' = c*I
%
% The solution is computed using the RungeKutta method, with the helper
% method rungeKuttaSystemMethod. The system is solved from t=0 to t=Tf,
% with initial conditions ICs given as a 3 component vector.
```

```
function [S,I,Q,R,D] = SIRQModel_244(alpha, beta, gamma, delta, eta, rho,
↪  ICs, Tf)
% This code runs a more complicated SIR model that adds in Q (a quarantined
% population) and D (a deceased population). The system of differential
↪  equations used here is
%   S' = -alpha*S*I
%   I' = alpha*S*I - (beta+gamma+delta)I
%   Q' = beta*I - (eta + rho)Q
%   R' = gamma*I + eta*Q
%   D' = delta*I + rho*Q
%
% The solution is computed using the RungeKutta method, with the helper
% method rungeKuttaSystemMethod. The system is solved from t=0 to t=Tf,
% with initial conditions ICs given as a 5 component vector.
```

```
function [S,I,Q,R,D] = SIRQVModel_244(alpha, beta, gamma, delta, eta, rho,
↪  zeta, ICs, Tf)
% This code runs a more complicated SIR model that adds in Q (a quarantined
% population) and D (a deceased population). The V component adds
% vaccination into the picture, where members are moved from S to R
% directly. The system of differential equations used here is
%   S' = -alpha*S*I - zeta*S
%   I' = alpha*S*I - (beta+gamma+delta)I
%   Q' = beta*I - (eta + rho)Q
%   R' = gamma*I + eta*Q+zeta*S
%   D' = delta*I + rho*Q
%
% The solution is computed using the RungeKutta method, with the helper
% method rungeKuttaSystemMethod. The system is solved from t=0 to t=Tf,
% with initial conditions ICs given as a 5 component vector.
```

Each of these last three methods use the Runge Kutta method to numerical solve a disease modeling problem with their respective equations. The shared arguments are the initial conditions, which are a three or five component vector depending on the problem type, and the final time T_f. The step-size used is one day, and the method will return the list of time-stepped values for each population (every day) from $t = 0$ to $t = T_f$. For SIR, the equations are

$$\frac{dS}{dt} = -rSI \qquad \frac{dI}{dt} = rSI - cI \qquad \frac{dR}{dt} = cI.$$

For SIRQ, the equations are

$$\begin{aligned}
\frac{dS}{dt} &= -\alpha SI \\
\frac{dI}{dt} &= \alpha SI - \beta I - \gamma I - \delta I \\
\frac{dQ}{dt} &= \beta I - \eta Q - \rho Q \\
\frac{dR}{dt} &= \gamma I + \eta Q \\
\frac{dD}{dt} &= \delta I + \rho Q
\end{aligned}$$

and for SIRQV, it is

$$\begin{aligned}
\frac{dS}{dt} &= -\alpha SI - \zeta S \\
\frac{dI}{dt} &= \alpha SI - \beta I - \gamma I - \delta I \\
\frac{dQ}{dt} &= \beta I - \eta Q - \rho Q \\
\frac{dR}{dt} &= \gamma I + \eta Q + \zeta S \\
\frac{dD}{dt} &= \delta I + \rho Q
\end{aligned}$$

An example implementation is

```
[S,I,R] = SIRModel_244(0.1, 0.2, [0.99; 0.01; 0], 400);
[S,I,Q,R,D] = SIRQModel_244(0.15, 0.08, 0.02, 0.03, 0.01, 0.04, [0.95; 0.05;
↪  0; 0; 0], 400);
[S,I,Q,R,D] = SIRQVModel_244(0.15, 0.08, 0.02, 0.03, 0.01, 0.04,0.2, [0.95;
↪  0.05; 0; 0; 0], 400);
```

Appendix B

Prerequisite Material

This chapter provides a review of some of the material from previous classes that may be a little rusty by the time one reaches differential equations. This can be used as a reference whenever these topics come up throughout the book. A lot of this material (or inspiration for it) is taken from the Precalculus book by Stitz and Zeager [SZ].

B.1 Polynomials and Factoring

Note: Attribution: [SZ], §A.8, A.9, 2.2-2.4

There are several components of differential equations, particularly higher order equations and systems, that involve dealing with and finding roots of polynomials, using these results to generate solutions to differential equations. This appendix will review some properties of and techniques related to polynomials.

B.1.1 Definitions and Operations

First we start with the definition of a polynomial. A *polynomial* is a sum of terms each of which is a real number or a real number multiplied by one or more variables to natural number powers. Some examples of polynomials are $x^2 + x\sqrt{3} + 4$, $27x^2y + \frac{7x}{2}$ and 6. Things like $3\sqrt{x}$, $4x - \frac{2}{x+1}$ and $13x^{2/3}y^2$ are *not* polynomials. Below, we review some terminology about polynomials.

Definition B.1.1

- Terms in polynomials without variables are called *constant* terms.
- In non-constant terms, the real number factor in the expression is called the *coefficient* of the term.
- The *degree* of a non-constant term is the sum of the exponents on the variables in the term; non-zero constant terms are defined to have degree 0. The degree of a polynomial is the highest degree of the nonzero terms.
- Terms in a polynomial are called *like* terms if they have the same variables each with the same corresponding exponents.
- A polynomial is said to be *simplified* if all arithmetic operations have been completed and there are no longer any like terms.
- A simplified polynomial is called a
  - *monomial* if it has exactly one nonzero term
  - *binomial* if it has exactly two nonzero terms
  - *trinomial* if it has exactly three nonzero terms

For example, $x^2 + x\sqrt{3} + 4$ is a trinomial of degree 2. The coefficient of x^2 is 1 and the constant term is 4. The polynomial $27x^2y + \frac{7x}{2}$ is a binomial of degree 3 ($x^2y = x^2y^1$) with constant term 0.

The concept of 'like' terms really amounts to finding terms which can be combined using the Distributive Property. For example, in the polynomial $17x^2y - 3xy^2 + 7xy^2$, $-3xy^2$ and $7xy^2$ are like terms, since they have the same variables with the same corresponding

exponents. This allows us to combine these two terms as follows:

$$17x^2y - 3xy^2 + 7xy^2 = 17x^2y + (-3)xy^2 + 7xy^2 + 17x^2y + (-3+7)xy^2 = 17x^2y + 4xy^2$$

Note that even though $17x^2y$ and $4xy^2$ have the same variables, they are not like terms since in the first term we have x^2 and $y = y^1$ but in the second we have $x = x^1$ and $y = y^2$ so the corresponding exponents aren't the same. Hence, $17x^2y + 4xy^2$ is the simplified form of the polynomial.

There are four basic operations we can perform with polynomials: addition, subtraction, multiplication and division. Addition, subtraction, and multiplication follow the standard properties of real numbers after distributing or expanding all terms (for multiplication) and then collecting like terms again. Division, on the other hand, is a bit more complicated and will be discussed next.

Polynomial Long Division

We now turn our attention to polynomial long division. Dividing two polynomials follows the same algorithm, in principle, as dividing two natural numbers so we review that process first. Suppose we wished to divide 2585 by 79. The standard division tableau is given below.

$$\begin{array}{r} 32 \\ 79\overline{)\,2585} \\ -\underline{237\downarrow} \\ 215 \\ -\underline{158} \\ 57 \end{array}$$

In this case, 79 is called the *divisor*, 2585 is called the *dividend*, 32 is called the *quotient* and 57 is called the *remainder*. We can check our answer by showing:

$$\text{dividend} = (\text{divisor})(\text{quotient}) + \text{remainder}$$

or in this case, $2585 = (79)(32) + 57\checkmark$. We hope that the long division tableau evokes warm, fuzzy memories of your formative years as opposed to feelings of hopelessness and frustration. If you experience the latter, keep in mind that the Division Algorithm essentially is a two-step process, iterated over and over again. First, we guess the number of times the divisor goes into the dividend and then we subtract off our guess. We repeat those steps with what's left over until what's left over (the remainder) is less than what we started with (the divisor). That's all there is to it!

The division algorithm for polynomials has the same basic two steps but when we subtract polynomials, we must take care to subtract *like terms* only. As a transition to polynomial division, let's write out our previous division tableau in expanded form.

$$\begin{array}{r|l} & 3\cdot 10+2 \\ \hline 7\cdot 10+9 & 2\cdot 10^3+5\cdot 10^2+8\cdot 10+5 \\ & \underline{-(2\cdot 10^3+3\cdot 10^2+7\cdot 10)} \quad \downarrow \\ & 2\cdot 10^2+1\cdot 10+5 \\ & \underline{-(1\cdot 10^2+5\cdot 10+8)} \\ & 5\cdot 10+7 \end{array}$$

Written this way, we see that when we line up the digits we are really lining up the coefficients of the corresponding powers of 10 - much like how we'll have to keep the powers of x lined up in the same columns. The big difference between polynomial division and the division of natural numbers is that the value of x is an unknown quantity. So unlike using the known value of 10, when we subtract there can be no regrouping of coefficients as in our previous example. (The subtraction $215 - 158$ requires us to 'regroup' or 'borrow' from the tens digit, then the hundreds digit.) This actually makes polynomial division easier.* Before we dive into examples, we first state a theorem telling us when we can divide two polynomials, and what to expect when we do so.

Theorem B.1.1 (Polynomial Division)

Let d and p be nonzero polynomials where the degree of p is greater than or equal to the degree of d. There exist two unique polynomials, q and r, such that $p = d \cdot q + r$, where either $r = 0$ or the degree of r is strictly less than the degree of d.

Essentially, Theorem B.1.1 tells us that we can divide polynomials whenever the degree of the divisor is less than or equal to the degree of the dividend. We know we're done with the division when the polynomial left over (the remainder) has a degree strictly less than the divisor. It's time to walk through a few examples to refresh your memory.

Example B.1.1: Perform the indicated division. Check your answer by showing

$$\text{dividend} = (\text{divisor})(\text{quotient}) + \text{remainder}$$

1. $(x^3 + 4x^2 - 5x - 14) \div (x - 2)$
2. $(2t + 7) \div (3t - 4)$
3. $(6y^2 - 1) \div (2y + 5)$
4. $(w^3) \div \left(w^2 - \sqrt{2}\right)$.

Solution:

1. To begin $(x^3 + 4x^2 - 5x - 14) \div (x - 2)$, we divide the first term in the dividend, namely x^3, by the first term in the divisor, namely x, and get $\frac{x^3}{x} = x^2$. This then becomes the first term in the quotient. We proceed as in regular long division at this point: we multiply the entire divisor, $x - 2$, by this first term in the quotient to get

*In our opinion - you can judge for yourself.

$x^2(x-2) = x^3 - 2x^2$. We then subtract this result from the dividend.

$$\begin{array}{r|l} & x^2 \\ x-2 & x^3+4x^2-5x-14 \\ & \underline{-(x^3-2x^2)} \quad \downarrow \\ & \quad 6x^2-5x \end{array}$$

Now we 'bring down' the next term of the quotient, namely $-5x$, and repeat the process. We divide $\frac{6x^2}{x} = 6x$, and add this to the quotient polynomial, multiply it by the divisor (which yields $6x(x-2) = 6x^2 - 12x$) and subtract.

$$\begin{array}{r|l} & x^2 + 6x \\ x-2 & x^3+4x^2-5x-14 \\ & \underline{-(x^3-2x^2)} \quad \downarrow \\ & \quad 6x^2-5x \quad \downarrow \\ & \underline{-(6x^2-12x)} \quad \downarrow \\ & \quad 7x-14 \end{array}$$

Finally, we 'bring down' the last term of the dividend, namely -14, and repeat the process. We divide $\frac{7x}{x} = 7$, add this to the quotient, multiply it by the divisor (which yields $7(x-2) = 7x - 14$) and subtract.

$$\begin{array}{r|l} & x^2 + 6x + 7 \\ x-2 & x^3+4x^2-5x-14 \\ & \underline{-(x^3-2x^2)} \\ & \quad 6x^2-5x \\ & \underline{-(6x^2-12x)} \\ & \quad 7x-14 \\ & \underline{-(7x-14)} \\ & \quad 0 \end{array}$$

In this case, we get a quotient of $x^2 + 6x + 7$ with a remainder of 0. To check our answer, we compute

$$(x-2)\left(x^2+6x+7\right) + 0 = x^3 + 6x^2 + 7x - 2x^2 - 12x - 14 = x^3 + 4x^2 - 5x - 14 \checkmark$$

2. To compute $(2t+7) \div (3t-4)$, we start as before. We find $\frac{2t}{3t} = \frac{2}{3}$, so that becomes the first (and only) term in the quotient. We multiply the divisor $(3t-4)$ by $\frac{2}{3}$ and get

$2t - \frac{8}{3}$. We subtract this from the divided and get $\frac{29}{3}$.

$$\begin{array}{r|l} & \frac{2}{3} \\ \hline 3t-4 & 2t + 7 \\ & -\left(2t - \frac{8}{3}\right) \\ \hline & \frac{29}{3} \end{array}$$

Our answer is $\frac{2}{3}$ with a remainder of $\frac{29}{3}$. To check our answer, we compute

$$(3t-4)\left(\frac{2}{3}\right) + \frac{29}{3} = 2t - \frac{8}{3} + \frac{29}{3} = 2t + \frac{21}{3} = 2t + 7 \checkmark$$

3. When we set-up the tableau for $(6y^2 - 1) \div (2y + 5)$, we must first issue a 'placeholder' for the 'missing' y-term in the dividend, $6y^2 - 1 = 6y^2 + 0y - 1$. We then proceed as before. Since $\frac{6y^2}{2y} = 3y$, $3y$ is the first term in our quotient. We multiply $(2y + 5)$ times $3y$ and subtract it from the dividend. We bring down the -1, and repeat.

$$\begin{array}{r|l} & 3y - \frac{15}{2} \\ \hline 2y+5 & 6y^2 + 0y - 1 \\ & -(6y^2 + 15y) \quad \downarrow \\ & -15y - 1 \\ & -\left(-15y - \frac{75}{2}\right) \\ \hline & \frac{73}{2} \end{array}$$

Our answer is $3y - \frac{15}{2}$ with a remainder of $\frac{73}{2}$. To check our answer, we compute:

$$(2y+5)\left(3y - \frac{15}{2}\right) + \frac{73}{2} = 6y^2 - 15y + 15y - \frac{75}{2} + \frac{73}{2} = 6y^2 - 1 \checkmark$$

4. For our last example, we need 'placeholders' for both the divisor $w^2 - \sqrt{2} = w^2 + 0w - \sqrt{2}$ and the dividend $w^3 = w^3 + 0w^2 + 0w + 0$. The first term in the quotient is $\frac{w^3}{w^2} = w$, and when we multiply and subtract this from the dividend, we're left with just $0w^2 + w\sqrt{2} + 0 = w\sqrt{2}$.

$$\begin{array}{r|l} & w \\ \hline w^2+0w-\sqrt{2} & w^3 + 0w^2 + 0w + 0 \\ & -\left(w^3 + 0w^2 - w\sqrt{2}\right) \quad \downarrow \\ & 0w^2 + w\sqrt{2} + 0 \end{array}$$

Since the degree of $w\sqrt{2}$ (which is 1) is less than the degree of the divisor (which is 2), we are done.* Our answer is w with a remainder of $w\sqrt{2}$. To check, we compute:

$$\left(w^2 - \sqrt{2}\right) w + w\sqrt{2} = w^3 - w\sqrt{2} + w\sqrt{2} = w^3 \checkmark$$

B.1.2 Synthetic Division

Usually, when we want to divide polynomials, it is because we are trying to find all roots of a polynomial. This comes from the idea that if we have a polynomial $p(x)$ and a value x_0 so that $p(x_0) = 0$, then x_0 is a root of the polynomial. This means that $(x - x_0)$ is a factor of $p(x)$, so that we can write

$$p(x) = (x - x_0)q(x)$$

where $q(x)$ is a polynomial with one lower degree than p. We can find this $q(x)$ by dividing

$$q(x) = \frac{p(x)}{x - x_0},$$

which is why we need division to sort this out.

This means that we need to find the roots (or at least a root) to know what to divide $p(x)$ by in order to start this process. The main theorem that can tell us where to start is the Rational Roots Theorem.

Theorem B.1.2 (Rational Zeros Theorem)

Suppose $f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$ is a polynomial of degree n with $n \geq 1$, and a_0, a_1, $\ldots a_n$ are integers. If r is a rational zero of f, then r is of the form $\pm\frac{p}{q}$, where p is a factor of the constant term a_0, and q is a factor of the leading coefficient a_n.

The Rational Zeros Theorem gives us a list of numbers to try in our synthetic division and that is a lot nicer than simply guessing. If none of the numbers in the list are zeros, then either the polynomial has no real zeros at all, or all of the real zeros are irrational numbers.

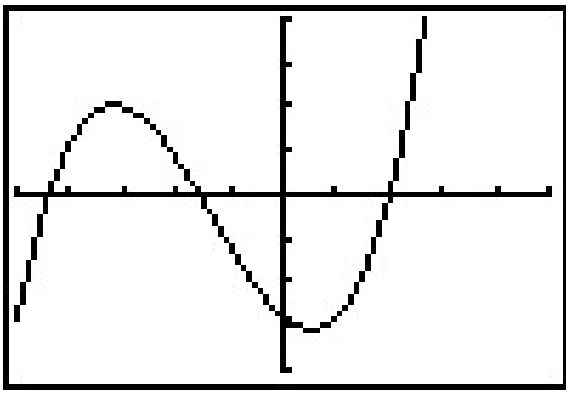
Example B.1.2: Let $f(x) = 2x^4 + 4x^3 - x^2 - 6x - 3$. Use the Rational Zeros Theorem to list all of the possible rational zeros of f.

Solution: To generate a complete list of rational zeros, we need to take each of the factors of constant term, $a_0 = -3$, and divide them by each of the factors of the leading coefficient $a_4 = 2$. The factors of -3 are ± 1 and ± 3. Since the Rational Zeros Theorem tacks on a $\pm$ anyway, for the moment, we consider only the positive factors 1 and 3. The factors of 2 are 1 and 2, so the Rational Zeros Theorem gives the list $\left\{\pm\frac{1}{1}, \pm\frac{1}{2}, \pm\frac{3}{1}, \pm\frac{3}{2}\right\}$ or $\left\{\pm\frac{1}{2}, \pm 1, \pm\frac{3}{2}, \pm 3\right\}$.

*Since $\frac{0w^2}{w^2} = 0$, we could proceed, write our quotient as $w + 0$, and move on... but even pedants have limits.

But this still doesn't make the process easy or straight-forward for finding the roots. How can we take this list of options and easily figure out where the roots are, and what the remaining polynomial $q(x)$ is?

We start by way of example: suppose we wish to determine the zeros of $f(x) = x^3 + 4x^2 - 5x - 14$. Setting $f(x) = 0$ results in the polynomial equation $x^3 + 4x^2 - 5x - 14 = 0$. Despite all of the factoring techniques we learned (and forgot!), this equation foils* us at every turn. Knowing that the zeros of f correspond to x-intercepts on the graph of $y = f(x)$, we use a graphing utility to produce the graph below on the left. The graph suggests that the function has three zeros, one of which appears to be $x = 2$ and two others for whom we are provided what we assume to be decimal approximations: $x \approx -4.414$ and $x \approx -1.586$. We can verify if these are zeros easily enough. We find $f(2) = (2)^2 + 4(2)^2 - 5(2) - 14 = 0$, but $f(-4.414) \approx 0.0039$ and $f(-1.586) \approx 0.0022$, While these last two values are probably by some measures, 'close' to 0, they are not *exactly* equal to 0. The question becomes: is there a way to use the fact that $x = 2$ is a zero to obtain the other two zeros? Based on our experience, if $x = 2$ is a zero, it seems that there should be a factor of $(x - 2)$ lurking around in the factorization of $f(x)$. In other words, we should expect that $x^3 + 4x^2 - 5x - 14 = (x - 2)\, q(x)$, where $q(x)$ is some other polynomial. How could we find such a $q(x)$, if it even exists? The answer comes from our old friend, polynomial division. Below on the right, we perform the long division: $(x^3 + 4x^2 - 5x - 14) \div (x - 2)$ and obtain $x^2 + 6x + 7$.



$$
\begin{array}{r|l}
 & x^2 + 6x + 7 \\
\hline
x-2 & x^3 + 4x^2 - 5x - 14 \\
 & \underline{-(x^3 - 2x^2)} \\
 & \quad 6x^2 - 5x \\
 & \quad \underline{-(6x^2 - 12x)} \\
 & \qquad 7x - 14 \\
 & \qquad \underline{-(7x - 14)} \\
 & \qquad\quad 0
\end{array}
$$

Said differently, $f(x) = x^3 + 4x^2 - 5x - 14 = (x - 2)(x^2 + 6x + 7)$. Using this form of $f(x)$, we find the zeros by solving $(x - 2)(x^2 + 6x + 7) = 0$. Setting each factor equal to 0, we get $x - 2 = 0$ (which gives us our known zero, $x = 2$) as well as $x^2 + 6x + 7 = 0$. The latter doesn't factor nicely, so we apply the Quadratic Formula to get $x = -3 \pm \sqrt{2}$. Sure enough, $-3 - \sqrt{2} \approx -4.414$ and $-3 + \sqrt{2} \approx -1.586$. We leave it to the reader to show $f(-3 - \sqrt{2}) = 0$ and $f(-3 + \sqrt{2}) = 0$.

The point of this section is to generalize the technique applied here. First up is a friendly reminder of what we can expect when we divide polynomials.

*pun intended

Theorem B.1.3

Suppose $d(x)$ and $p(x)$ are nonzero polynomial functions where the degree of p is greater than or equal to the degree of d. There exist two unique polynomial functions, $q(x)$ and $r(x)$, such that $p(x) = d(x)\,q(x) + r(x)$, where either $r(x) = 0$ or the degree of r is strictly less than the degree of d.

As you may recall, all of the polynomials in Theorem B.1.3 have special names. The polynomial p is called the *dividend*; d is the *divisor*; q is the *quotient*; r is the *remainder*. If $r(x) = 0$ then d is called a *factor* of p. The word 'unique' here is critical in that it guarantees there is only *one* quotient and remainder for each division problem.* The proof of Theorem B.1.3 is usually relegated to a course in Abstract Algebra, but we can still use the result to move forward with the rest of this section.

If we want to find all of the roots of a polynomial in a reasonable way, we had better find a more efficient way to divide polynomial functions by quantities of the form $x - c$. Fortunately, people like Ruffini and Horner have already blazed this trail. Let's take a closer look at the long division we performed at the beginning of the section and try to streamline it. First off, let's change all of the subtractions into additions by distributing through the -1s.

$$\begin{array}{r|l} & x^2 + 6x + 7 \\ \hline x-2 & x^3 + 4x^2 - 5x - 14 \\ & \underline{-x^3 + 2x^2} \\ & \quad 6x^2 - 5x \\ & \quad \underline{-6x^2 + 12x} \\ & \qquad 7x - 14 \\ & \qquad \underline{-7x + 14} \\ & \qquad\quad 0 \end{array}$$

Next, observe that the terms $-x^3$, $-6x^2$ and $-7x$ are the exact opposite of the terms above them. The algorithm we use ensures this is always the case, so we can omit them without losing any information. Also note that the terms we 'bring down' (namely the $-5x$ and -14) aren't really necessary to recopy, so we omit them, too.

$$\begin{array}{r|l} & x^2 + 6x + 7 \\ \hline x-2 & x^3 + 4x^2 - 5x - 14 \\ & \underline{\quad 2x^2} \\ & \quad 6x^2 \\ & \quad\ \underline{\quad 12x} \\ & \qquad 7x \\ & \qquad\ \underline{\quad 14} \\ & \qquad\quad 0 \end{array}$$

Let's move terms up a bit and copy the x^3 into the last row.

*Hence the use of the definite article 'the' when speaking of *the* quotient and *the* remainder.

$$\begin{array}{r|l} & x^2+6x+7 \\ \hline x-2 & x^3+4x^2-5x-14 \\ & \quad 2x^2 \;\; 12x \;\; 14 \\ \hline & x^3 \;\; 6x^2 \;\; 7x \;\; 0 \end{array}$$

Note that by arranging things in this manner, each term in the last row is obtained by adding the two terms above it. Notice also that the quotient polynomial can be obtained by dividing each of the first three terms in the last row by x and adding the results. If you take the time to work back through the original division problem, you will find that this is exactly the way we determined the quotient polynomial. This means that we no longer need to write the quotient polynomial down, nor the x in the divisor, to determine our answer.

$$\begin{array}{r|l} -2 & x^3+4x^2-5x-14 \\ & \quad 2x^2 \;\; 12x \;\; 14 \\ \hline & x^3 \;\; 6x^2 \;\; 7x \;\; 0 \end{array}$$

We've streamlined things quite a bit so far, but we can still do more. Let's take a moment to remind ourselves where the $2x^2$, $12x$ and 14 came from in the second row. Each of these terms was obtained by multiplying the terms in the quotient, x^2, $6x$ and 7, respectively, by the -2 in $x-2$, then by -1 when we changed the subtraction to addition. Multiplying by -2 then by -1 is the same as multiplying by 2, so we replace the -2 in the divisor by 2. Furthermore, the coefficients of the quotient polynomial match the coefficients of the first three terms in the last row, so we now take the plunge and write only the coefficients of the terms to get

$$\begin{array}{r|l} 2 & 1 \;\; 4 \;\; -5 \;\; -14 \\ & \quad 2 \;\; 12 \;\; 14 \\ \hline & 1 \;\; 6 \;\; 7 \;\; 0 \end{array}$$

We have constructed a *synthetic division tableau* for this polynomial division problem. Let's re-work our division problem using this tableau to see how it greatly streamlines the division process. To divide $x^3+4x^2-5x-14$ by $x-2$, we write 2 in the place of the divisor and the coefficients of $x^3+4x^2-5x-14$ in for the dividend. Then 'bring down' the first coefficient of the dividend.

$$\begin{array}{r|l} 2 & 1 \;\; 4 \;\; -5 \;\; -14 \\ \hline & \end{array} \qquad\qquad \begin{array}{r|l} 2 & 1 \;\; 4 \;\; -5 \;\; -14 \\ & \downarrow \\ \hline & 1 \end{array}$$

Next, take the 2 from the divisor and multiply by the 1 that was 'brought down' to get 2. Write this underneath the 4, then add to get 6.

$$\begin{array}{r|l} 2 & 1 \;\; 4 \;\; -5 \;\; -14 \\ & \downarrow \;\; 2 \\ \hline & 1 \end{array} \qquad\qquad \begin{array}{r|l} 2 & 1 \;\; 4 \;\; -5 \;\; -14 \\ & \downarrow \;\; 2 \\ \hline & 1 \;\; 6 \end{array}$$

Now take the 2 from the divisor times the 6 to get 12, and add it to the -5 to get 7.

$$\begin{array}{r|cccc} 2 & 1 & 4 & -5 & -14 \\ & \downarrow & 2 & 12 & \\ \hline & 1 & 6 & & \end{array} \qquad \begin{array}{r|cccc} 2 & 1 & 4 & -5 & -14 \\ & \downarrow & 2 & 12 & \\ \hline & 1 & 6 & 7 & \end{array}$$

Finally, take the 2 in the divisor times the 7 to get 14, and add it to the -14 to get 0.

$$\begin{array}{r|cccc} 2 & 1 & 4 & -5 & -14 \\ & \downarrow & 2 & 12 & 14 \\ \hline & 1 & 6 & 7 & \end{array} \qquad \begin{array}{r|cccc} 2 & 1 & 4 & -5 & -14 \\ & \downarrow & 2 & 12 & 14 \\ \hline & 1 & 6 & 7 & \boxed{0} \end{array}$$

The first three numbers in the last row of our tableau are the coefficients of the quotient polynomial. Remember, we started with a third degree polynomial and divided by a first degree polynomial, so the quotient is a second degree polynomial. Hence the quotient is $x^2 + 6x + 7$. The number in the box is the remainder. Synthetic division is our tool of choice for dividing polynomials by divisors of the form $x - c$. It is important to note that it works *only* for these kinds of divisors.* Also take note that when a polynomial (of degree at least 1) is divided by $x - c$, the result will be a polynomial of exactly one less degree. Finally, it is worth the time to trace each step in synthetic division back to its corresponding step in long division. While the authors have done their best to indicate where the algorithm comes from, there is no substitute for working through it yourself.

Example B.1.3: Use synthetic division to perform the following polynomial divisions. Identify the quotient and remainder.

1. $(5x^3 - 2x^2 + 1) \div (x - 3)$
2. $(t^3 + 8) \div (t + 2)$
3. $\dfrac{4 - 8z - 12z^2}{2z - 3}$

Solution:

1. When setting up the synthetic division tableau, the coefficients of even 'missing' terms need to be accounted for, so we enter 0 for the coefficient of x in the dividend.

$$\begin{array}{r|cccc} 3 & 5 & -2 & 0 & 1 \\ & \downarrow & 15 & 39 & 117 \\ \hline & 5 & 13 & 39 & \boxed{118} \end{array}$$

   Since the dividend was a third degree polynomial function, the quotient is a second degree (quadratic) polynomial function with coefficients 5, 13 and 39: $q(x) = 5x^2 + 13x + 39$. The remainder is $r(x) = 118$. According to Theorem B.1.3, we have $5x^3 - 2x^2 + 1 = (x - 3)(5x^2 + 13x + 39) + 118$, which we leave to the reader to check.

*You'll need to use good old-fashioned polynomial long division for divisors of degree larger than 1.

2. To use synthetic division here, we rewrite $t+2$ as $t-(-2)$ and proceed as before

$$\begin{array}{r|cccc} -2 & 1 & 0 & 0 & 8 \\ & \downarrow & -2 & 4 & -8 \\ \hline & 1 & -2 & 4 & \boxed{0} \end{array}$$

We get the quotient $q(t) = t^2 - 2t + 4$ and the remainder $r(t) = 0$. Relating the dividend, quotient and remainder gives: $t^3 + 8 = (t+2)(t^2 - 2t + 4)$, which is a specific instance of the 'sum of cubes' formula some of you may recall.

3. To divide $4 - 8z - 12z^2$ by $2z - 3$, two things must be done. First, we write the dividend in descending powers of z as $-12z^2 - 8z + 4$. Second, since synthetic division works only for factors of the form $z - c$, we factor $2z - 3$ as $2\left(z - \frac{3}{2}\right)$. Hence, we are dividing $-12z^2 - 8z + 4$ by two factors: 2 and $\left(z - \frac{3}{2}\right)$. Dividing first by 2, we obtain $-6z^2 - 4z + 2$. Next, we divide $-6z^2 - 4z + 2$ by $\left(z - \frac{3}{2}\right)$:

$$\begin{array}{r|ccc} \frac{3}{2} & -6 & -4 & 2 \\ & \downarrow & -9 & -\frac{39}{2} \\ \hline & -6 & -13 & \boxed{-\frac{35}{2}} \end{array}$$

Hence, $-6z^2 - 4z + 2 = \left(z - \frac{3}{2}\right)(-6z - 13) - \frac{35}{2}$. However when it comes to writing the dividend, quotient and remainder in the form given in Theorem B.1.3, we need to find $q(z)$ and $r(z)$ so that $-12z^2 - 8z + 4 = (2z-3)q(z) + r(z)$. Hence, starting with $-6z^2 - 4z + 2 = \left(z - \frac{3}{2}\right)(-6z - 13) - \frac{35}{2}$, we multiply 2 back on both sides:

$$\begin{array}{rcl} -6z^2 - 4z + 2 & = & \left(z - \frac{3}{2}\right)(-6z - 13) - \frac{35}{2} \\ 2\left(-6z^2 - 4z + 2\right) & = & 2\left[\left(z - \frac{3}{2}\right)(-6z - 13) - \frac{35}{2}\right] \\ -12z^2 - 8z + 4 & = & 2\left(z - \frac{3}{2}\right)(-6z - 13) - 2\left(\frac{35}{2}\right) \\ -12z^2 - 8z + 4 & = & (2z-3)(-6z - 13) - 35 \end{array}$$

At this stage, we have written $-12z^2 - 8z + 4$ in the *form* $(2z-3)q(z) + r(z)$, so we identify the quotient as $q(z) = -6z - 13$ and the remainder is $r(z) = -35$. But how can we be sure these are the same quotient and remainder polynomial functions we would have obtained if we had taken the time to do the long division in the first place? Because of the word 'unique' in Theorem B.1.3. The theorem states that there is only *one* way to decompose $-12z^2 - 8z + 4$ as $(2z-3)q(z) + r(z)$. Since we have found such a way, we can be sure it is the only way.*

The next example pulls together all of the concepts discussed in this section.

Example B.1.4: Let $p(x) = 2x^3 - 5x + 3$.

1. Find $p(-2)$ using The Remainder Theorem. Check your answer by substitution.

*But it wouldn't hurt to check, just this once.

2. Verify $x = 1$ is a zero of p and use this information to all the real zeros of p.

Solution:

1. The Remainder Theorem states $p(-2)$ is the remainder when $p(x)$ is divided by $x-(-2)$. We set up our synthetic division tableau below. We are careful to record the coefficient of x^2 as 0:

$$\begin{array}{r|rrrr} -2 & 2 & 0 & -5 & 3 \\ & \downarrow & -4 & 8 & -6 \\ \hline & 2 & -4 & 3 & \boxed{-3} \end{array}$$

   According to the Remainder Theorem, $p(-2) = -3$. We can check this by direct substitution into the formula for $p(x)$: $p(-2) = 2(-2)^3 - 5(-2) + 3 = -16 + 10 + 3 = -3$.

2. We verify $x = 1$ is a zero of p by evaluating $p(1) = 2(1)^3 - 5(1) + 3 = 0$. To see if there are any more real zeros, we need to solve $p(x) = 2x^3 - 5x + 3 = 0$. From the Factor Theorem, we know since $p(1) = 0$, we can factor $p(x)$ as $(x-1)q(x)$. To find $q(x)$, we use synthetic division:

$$\begin{array}{r|rrrr} 1 & 2 & 0 & -5 & 3 \\ & \downarrow & 2 & 2 & -3 \\ \hline & 2 & 2 & -3 & \boxed{0} \end{array}$$

   As promised, our remainder is 0, and we get $p(x) = (x-1)(2x^2 + 2x - 3)$. Setting this form of $p(x)$ equal to 0 we get $(x-1)(2x^2 + 2x - 3) = 0$. We recover $x = 1$ from setting $x - 1 = 0$ but we also obtain $x = \frac{-1 \pm \sqrt{7}}{2}$ from $2x^2 + 2x - 3 = 0$, courtesy of the Quadratic Formula.

□

Our next example demonstrates how we can extend the synthetic division tableau to accommodate zeros of multiplicity greater than 1.

Example B.1.5: Let $p(x) = 4x^4 - 4x^3 - 11x^2 + 12x - 3$. Show $x = \frac{1}{2}$ is a zero of multiplicity 2 and find all of the remaining real zeros of p.

Solution: While computing $p\left(\frac{1}{2}\right) = 0$ shows $x = \frac{1}{2}$ is a zero of p, to prove it has multiplicity 2, we need to factor $p(x) = \left(x - \frac{1}{2}\right)^2 q(x)$ with $q\left(\frac{1}{2}\right) \neq 0$,. We set up for synthetic division, but instead of stopping after the first division, we continue the tableau downwards and divide $\left(x - \frac{1}{2}\right)$ directly into the quotient we obtained from the first division as follows:

$$\begin{array}{r|rrrrr} \frac{1}{2} & 4 & -4 & -11 & 12 & -3 \\ & \downarrow & 2 & -1 & -6 & 3 \\ \hline \frac{1}{2} & 4 & -2 & -12 & 6 & \boxed{0} \\ & \downarrow & 2 & 0 & -6 & \\ \hline & 4 & 0 & -12 & \boxed{0} & \end{array}$$

We get:* $4x^4 - 4x^3 - 11x^2 + 12x - 3 = \left(x - \frac{1}{2}\right)^2 (4x^2 - 12)$. Note if we let $q(x) = 4x^2 - 12$, then $q\left(\frac{1}{2}\right) = 4\left(\frac{1}{2}\right)^2 - 12 = -11 \neq 0$ which proves $x = \frac{1}{2}$ is a zero of p of multiplicity 2. To find the remaining zeros of p, we set the quotient $4x^2 - 12 = 0$, so $x^2 = 3$ and extract square roots to get $x = \pm\sqrt{3}$. ⌟

One last wrinkle in this process is complex roots, since it is possible for a polynomial (particularly a quadratic polynomial) to have complex numbers as roots. For a reminder of some more properties of complex numbers see § B.2. For this section in particular, we only need a few basic facts.

For us, it suffices to review the basic vocabulary.

Definition B.1.2

- The imaginary unit $i = \sqrt{-1}$ satisfies the two following properties
  1. $i^2 = -1$
  2. If c is a real number with $c \geq 0$ then $\sqrt{-c} = i\sqrt{c}$
- The *complex numbers* are the set of numbers $\mathbb{C} = \{a + bi \,|\, a, b \in \mathbb{R}\}$
- Given a complex number $z = a+bi$, the *complex conjugate* of z, $\overline{z} = \overline{a + bi} = a-bi$.

Note that every real number is a complex number, that is $\mathbb{R} \subseteq \mathbb{C}$. To see this, take your favorite real number, say 117. We may write $117 = 117 + 0i$ which puts in the form $a + bi$. Hence, we we speak of the 'complex zeros' of a polynomial function, we are talking about not just the non-real, but also the real zeros.

Complex numbers, by their very definition, are two dimensional creatures. To see this, we may identify a complex number $z = a + bi$ with the point in the Cartesian plane (a, b). The horizontal axis is called the 'real' axis since points here have the form $(a, 0)$ which corresponds to numbers of the form $z = a + 0i = a$ which are the real numbers. The vertical axis is called the 'imaginary' axis since points here are of the form $(0, b)$ which correspond to numbers of the form $z = 0 + bi = bi$, the so-called 'purely imaginary' numbers. Below we plot some complex numbers on this so-called 'Complex Plane.' Plotting a set of complex numbers this way is called an Argand Diagram, and opens up a wealth of opportunities to explore many algebraic properties of complex numbers geometrically. For example, complex conjugation amounts to a reflection about the real axis, and multiplication by i amounts to a 90° rotation. While we won't have much use for the Complex Plane in this section, it is worth introducing this concept now, if, for no other reason, it gives the reader a sense of the vastness of the complex number system and the role of the real numbers in it.

Returning to zeros of polynomials, suppose we wish to find the zeros of $f(x) = x^2 - 2x + 5$. To solve the equation $x^2 - 2x + 5 = 0$, we note that the quadratic doesn't factor nicely, so we

*For those wanting more detail: the first division gives: $4x^4 - 4x^3 - 11x^2 + 12x - 3 = \left(x - \frac{1}{2}\right)\left(4x^3 - 2x^2 - 12x + 6\right)$. The second division gives: $4x^3 - 2x^2 - 12x + 6 = \left(x - \frac{1}{2}\right)\left(4x^2 - 12\right)$.

resort to the Quadratic Formula and obtain

$$x = \frac{-(-2) \pm \sqrt{(-2)^2 - 4(1)(5)}}{2(1)} = \frac{2 \pm \sqrt{-16}}{2} = \frac{2 \pm 4i}{2} = 1 \pm 2i.$$

Two things are important to note. First, the zeros $1 + 2i$ and $1 - 2i$ are complex conjugates. If ever we obtain non-real zeros to a quadratic function with *real number* coefficients, the zeros will be a complex conjugate pair. (Do you see why?)

We could ask if all of the theory of polynomial division holds for non-real zeros, in particular the division algorithm and the Remainder and Factor Theorems. The answer is 'yes.'

$$\begin{array}{r|rrr} 1+2i & 1 & -2 & 5 \\ & \downarrow & 1+2i & -5 \\ \hline & 1 & -1+2i & \boxed{0} \end{array}$$

Indeed, the above shows $x^2 - 2x + 5 = (x - [1 + 2i])(x - 1 + 2i) = (x - [1 + 2i])(x - [1 - 2i])$ which demonstrates both $(x - [1 + 2i])$ and $(x - [1 - 2i])$ are factors of $x^2 - 2x + 5$.*

But how do we know if a general polynomial has any complex zeros at all? We have many examples of polynomials with no real zeros. Can there be polynomials with no zeros whatsoever? The answer to that last question is "No." and the theorem which provides that answer is The Fundamental Theorem of Algebra.

Theorem B.1.4 (The Fundamental Theorem of Algebra)

Suppose f is a polynomial function with complex number coefficients of degree $n \geq 1$, then f has at least one complex zero.

The Fundamental Theorem of Algebra is an example of an 'existence' theorem in Mathematics. Like the Intermediate Value Theorem, the Fundamental Theorem of Algebra guarantees the existence of at least one zero, but gives us no algorithm to use in finding it. In fact, as we mentioned previously, there are polynomials whose real zeros, though they exist, cannot be expressed using the 'usual' combinations of arithmetic symbols, and must be approximated. It took mathematicians literally hundreds of years to prove the theorem in its full generality,† and some of that history is recorded . Note that the Fundamental Theorem of Algebra applies to not only polynomial functions with real coefficients, but to those with complex number coefficients as well.

Suppose f is a polynomial function of degree $n \geq 1$. The Fundamental Theorem of Algebra guarantees us at least one complex zero, z_1. The Factor Theorem guarantees that $f(x)$ factors as $f(x) = (x - z_1)\, q_1(x)$ for a polynomial function q_1, which has degree $n - 1$. If $n - 1 \geq 1$, then the Fundamental Theorem of Algebra guarantees a complex zero of q_1 as well, say z_2, so then the Factor Theorem gives us $q_1(x) = (x - z_2)\, q_2(x)$, and hence $f(x) = (x - z_1)\,(x - z_2)\, q_2(x)$. We can continue this process exactly n times, at which point

*It is a good review of the algebra of complex numbers to start with $(x - [1 + 2i])(x - [1 - 2i])$, perform the indicated operations, and simplify the result to $x^2 - 2x + 5$. See part 6 of Example B.2.1.

†So if its profound nature and beautiful subtlety escape you, no worries!

our quotient polynomial q_n has degree 0 so it's a constant. This constant is none-other than the leading coefficient of f which is carried down line by line each time we divide by factors of the form $x - c$.

Theorem B.1.5 (Complex Factorization Theorem)

Suppose f is a polynomial function with complex number coefficients. If the degree of f is n and $n \geq 1$, then f has exactly n complex zeros, counting multiplicity. If z_1, z_2, ..., z_k are the distinct zeros of f, with multiplicities m_1, m_2, ..., m_k, respectively, then $f(x) = a\,(x - z_1)^{m_1}\,(x - z_2)^{m_2} \cdots (x - z_k)^{m_k}$.

Theorem B.1.5 says two important things: first, every polynomial is a product of linear factors; second, every polynomial function is completely determined by its zeros, their multiplicities, and its leading coefficient. We put this theorem to good use in the next example.

Example B.1.6: Let $f(x) = 12x^5 - 20x^4 + 19x^3 - 6x^2 - 2x + 1$.

1. Find all of the complex zeros of f and state their multiplicities.
2. Factor $f(x)$ using Theorem B.1.5

Solution:

1. Since f is a fifth degree polynomial, we know that we need to perform at least three successful divisions to get the quotient down to a quadratic function. At that point, we can find the remaining zeros using the Quadratic Formula, if necessary. Using the techniques of synthetic division:

$$\begin{array}{r|rrrrrr} \frac{1}{2} & 12 & -20 & 19 & -6 & -2 & 1 \\ & \downarrow & 6 & -7 & 6 & 0 & -1 \\ \hline \frac{1}{2} & 12 & -14 & 12 & 0 & -2 & \boxed{0} \\ & \downarrow & 6 & -4 & 4 & 2 & \\ \hline -\frac{1}{3} & 12 & -8 & 8 & 4 & \boxed{0} & \\ & \downarrow & -4 & 4 & -4 & & \\ \hline & 12 & -12 & 12 & \boxed{0} & & \end{array}$$

   Our quotient is $12x^2 - 12x + 12$, whose zeros we find to be $\frac{1 \pm i\sqrt{3}}{2}$. From Theorem B.1.5, we know f has exactly 5 zeros, counting multiplicities, and as such we have the zero $\frac{1}{2}$ with multiplicity 2, and the zeros $-\frac{1}{3}$, $\frac{1+i\sqrt{3}}{2}$ and $\frac{1-i\sqrt{3}}{2}$, each of multiplicity 1.

2. Applying Theorem B.1.5, we are guaranteed that f factors as

$$f(x) = 12\left(x - \frac{1}{2}\right)^2\left(x + \frac{1}{3}\right)\left(x - \left[\frac{1 + i\sqrt{3}}{2}\right]\right)\left(x - \left[\frac{1 - i\sqrt{3}}{2}\right]\right)$$

A true test of Theorem B.1.5 would be to take the factored form of $f(x)$ in the previous example and multiply it out* to see that it really does reduce to $f(x) = 12x^5 - 20x^4 + 19x^3 - 6x^2 - 2x + 1$. When factoring a polynomial using Theorem B.1.5, we say that it is *factored completely over the complex numbers*, meaning that it is impossible to factor the polynomial any further using complex numbers. If we wanted to completely factor $f(x)$ over the *real numbers* then we would have stopped short of finding the nonreal zeros of f and factored f using our work from the synthetic division to write $f(x) = \left(x - \frac{1}{2}\right)^2 \left(x + \frac{1}{3}\right)(12x^2 - 12x + 12)$, or $f(x) = 12\left(x - \frac{1}{2}\right)^2 \left(x + \frac{1}{3}\right)(x^2 - x + 1)$. Since the zeros of $x^2 - x + 1$ are nonreal, we call $x^2 - x + 1$ an *irreducible quadratic* meaning it is impossible to break it down any further using *real* numbers.

The last two results of the section show us that, theoretically, the non-real zeros of polynomial functions with real number coefficients come exclusively from irreducible quadratics.

Theorem B.1.6 (Conjugate Pairs Theorem)

If f is a polynomial function with real number coefficients and z is a complex zero of f, then so is $\overline{z}$.

To prove the theorem, let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_2 x^2 + a_1 x + a_0$ be a polynomial function with real number coefficients. If z is a zero of f, then $f(z) = 0$, which means $a_n z^n + a_{n-1}z^{n-1} + \ldots + a_2 z^2 + a_1 z + a_0 = 0$. Next, we consider $f\left(\overline{z}\right)$ and apply Theorem B.2.1 below.

$$
\begin{aligned}
f\left(\overline{z}\right) &= a_n\left(\overline{z}\right)^n + a_{n-1}\left(\overline{z}\right)^{n-1} + \ldots + a_2\left(\overline{z}\right)^2 + a_1\overline{z} + a_0 && \\
&= a_n\overline{z^n} + a_{n-1}\overline{z^{n-1}} + \ldots + a_2\overline{z^2} + a_1\overline{z} + a_0 && \text{since } \left(\overline{z}\right)^n = \overline{z^n} \\
&= \overline{a_n}\,\overline{z^n} + \overline{a_{n-1}}\,\overline{z^{n-1}} + \ldots + \overline{a_2}\,\overline{z^2} + \overline{a_1}\,\overline{z} + \overline{a_0} && \text{since the coefficients are real} \\
&= \overline{a_n z^n} + \overline{a_{n-1}z^{n-1}} + \ldots + \overline{a_2 z^2} + \overline{a_1 z} + \overline{a_0} && \text{since } \overline{z}\,\overline{w} = \overline{zw} \\
&= \overline{a_n z^n + a_{n-1}z^{n-1} + \ldots + a_2 z^2 + a_1 z + a_0} && \text{since } \overline{z} + \overline{w} = \overline{z+w} \\
&= \overline{f(z)} && \\
&= \overline{0} && \\
&= 0 &&
\end{aligned}
$$

This shows that $\overline{z}$ is a zero of f. So, if f is a polynomial function with real number coefficients, Theorem B.1.6 tells us that if $a + bi$ is a nonreal zero of f, then so is $a - bi$. In other words, nonreal zeros of f come in conjugate pairs. The Factor Theorem kicks in to give us both $(x-[a+bi])$ and $(x-[a-bi])$ as factors of $f(x)$ which means $(x-[a+bi])(x-[a-bi]) = x^2 + 2ax + (a^2 + b^2)$ is an irreducible quadratic factor of f. As a result, we have our last theorem of the section.

*This is a good chance to test your algebraic mettle and see that all of this does actually work.

Theorem B.1.7 (Real Factorization Theorem)

Suppose f is a polynomial function with real number coefficients. Then $f(x)$ can be factored into a product of linear factors corresponding to the real zeros of f and irreducible quadratic factors which give the nonreal zeros of f.

Example B.1.7: Let $f(x) = x^4 + 64$.

1. Use synthetic division to show that $x = 2 + 2i$ is a zero of f.
2. Find the remaining complex zeros of f.
3. Completely factor $f(x)$ over the complex numbers.
4. Completely factor $f(x)$ over the real numbers.

Solution:

1. Remembering to insert the 0's in the synthetic division tableau we have

$$\begin{array}{r|cccc} 2+2i & 1 & 0 & 0 & 0 & 64 \\ & \downarrow & 2+2i & 8i & -16+16i & -64 \\ \hline & 1 & 2+2i & 8i & -16+16i & \boxed{0} \end{array}$$

2. Since f is a fourth degree polynomial, we need to make two successful divisions to get a quadratic quotient. Since $2 + 2i$ is a zero, we know from Theorem B.1.6 that $2 - 2i$ is also a zero. We continue our synthetic division tableau.

$$\begin{array}{r|cccc} 2+2i & 1 & 0 & 0 & 0 & 64 \\ & \downarrow & 2+2i & 8i & -16+16i & -64 \\ \hline 2-2i & 1 & 2+2i & 8i & -16+16i & \boxed{0} \\ & \downarrow & 2-2i & 8-8i & 16-16i & \\ \hline & 1 & 4 & 8 & \boxed{0} & \end{array}$$

   Our quotient polynomial is $x^2 + 4x + 8$. Using the quadratic formula, we solve $x^2 + 4x + 8 = 0$ and find the remaining zeros are $-2 + 2i$ and $-2 - 2i$.

3. Using Theorem B.1.5, we get $f(x) = (x-[2-2i])(x-[2+2i])(x-[-2+2i])(x-[-2-2i])$.
4. To find the irreducible quadratic factors of $f(x)$, we multiply the factors together which correspond to the conjugate pairs. We find $(x - [2 - 2i])(x - [2 + 2i]) = x^2 - 4x + 8$, and $(x - [-2 + 2i])(x - [-2 - 2i]) = x^2 + 4x + 8$, so $f(x) = (x^2 - 4x + 8)(x^2 + 4x + 8)$.

We close this section with an example where we are asked to manufacture a polynomial function with certain characteristics.

Example B.1.8: Find a polynomial function p of lowest degree that has integer coefficients and satisfies all of the following criteria:

- the graph of $y = p(x)$ touches and rebounds from the x-axis at $\left(\frac{1}{3}, 0\right)$
- $x = 3i$ is a zero of p.
- as $x \to -\infty$, $p(x) \to -\infty$
- as $x \to \infty$, $p(x) \to -\infty$

Solution:

To solve this problem, we will need a good understanding of the relationship between the x-intercepts of the graph of a function and the zeros of a function, the Factor Theorem, the role of multiplicity, complex conjugates, the Complex Factorization Theorem, and end behavior of polynomial functions. (In short, you'll need most of the major concepts of this chapter.) Since the graph of p touches the x-axis at $\left(\frac{1}{3}, 0\right)$, we know $x = \frac{1}{3}$ is a zero of even multiplicity. Since we are after a polynomial of lowest degree, we need $x = \frac{1}{3}$ to have multiplicity exactly 2. The Factor Theorem now tells us $\left(x - \frac{1}{3}\right)^2$ is a factor of $p(x)$. Since $x = 3i$ is a zero and our final answer is to have integer (hence, real) coefficients, $x = -3i$ is also a zero. The Factor Theorem kicks in again to give us $(x - 3i)$ and $(x + 3i)$ as factors of $p(x)$. We are given no further information about zeros or intercepts so we conclude, by the Complex Factorization Theorem that $p(x) = a\left(x - \frac{1}{3}\right)^2 (x - 3i)(x + 3i)$ for some real number a. Expanding this, we get $p(x) = ax^4 - \frac{2a}{3}x^3 + \frac{82a}{9}x^2 - 6ax + a$. In order to obtain integer coefficients, we know a must be an integer multiple of 9. Our last concern is end behavior. Since the leading term of $p(x)$ is ax^4, we need $a < 0$ to get $p(x) \to -\infty$ as $x \to \pm\infty$. Hence, if we choose $x = -9$, we get $p(x) = -9x^4 + 6x^3 - 82x^2 + 54x - 9$. We can verify our handiwork using the techniques developed in this chapter. ⌟

B.2 Complex Numbers

Note: Attribution: [SZ], §A.11

The equation $x^2+1=0$ has no real number solutions. However, it *would* have solutions if we could make sense of $\sqrt{-1}$. The *Complex Numbers* do just that - they give us a mechanism for working with $\sqrt{-1}$. As such, the set of complex numbers fill in an algebraic gap left by the set of real numbers.

Here's the basic plan. There is no real number x with $x^2=-1$, since for any real number $x^2 \geq 0$. However, we could formally extract square roots and write $x=\pm\sqrt{-1}$. We build the complex numbers by relabeling the quantity $\sqrt{-1}$ as i, the unfortunately misnamed *imaginary unit*.* The number i, while not a real number, is defined so that it plays along well with real numbers and acts very much like any other radical expression. For instance, $3(2i)=6i$, $7i-3i=4i$, $(2-7i)+(3+4i)=5-3i$, and so forth. The key properties which distinguish i from the real numbers are listed below.

Definition B.2.1

The imaginary unit i satisfies the two following properties:

1. $i^2=-1$
2. If c is a real number with $c \geq 0$ then $\sqrt{-c}=i\sqrt{c}$

Property 1 in the previous definition establishes that i does act as a square root† of -1, and property 2 establishes what we mean by the 'principal square root' of a negative real number. In property 2, it is important to remember the restriction on c. For example, it is perfectly acceptable to say $\sqrt{-4}=i\sqrt{4}=i(2)=2i$. However, $\sqrt{-(-4)} \neq i\sqrt{-4}$, otherwise, we'd get

$$2=\sqrt{4}=\sqrt{-(-4)}=i\sqrt{-4}=i(2i)=2i^2=2(-1)=-2,$$

which is unacceptable. The moral of this story is that the general properties of radicals do not apply for even roots of negative quantities. With Definition B.2.1 in place, we can define the set of *complex numbers*.

A *complex number* is a number of the form $a+bi$, where a and b are real numbers and i is the imaginary unit. The set of complex numbers is denoted $\mathbb{C}$.

Complex numbers include things you'd normally expect, like $3+2i$ and $\frac{2}{5}-i\sqrt{3}$. However, don't forget that a or b could be zero, which means numbers like $3i$ and 6 are also complex numbers. In other words, don't forget that the complex numbers *include* the real numbers,‡ so 0 and $\pi-\sqrt{21}$ are both considered complex numbers. The arithmetic of complex numbers

*Some Technical Mathematics textbooks label it 'j'. While it carries the adjective 'imaginary', these numbers have essential real-world implications. For example, every electronic device owes its existence to the study of 'imaginary' numbers.

†Note the use of the indefinite article 'a'. Whatever beast is chosen to be i, $-i$ is the other square root of -1.

‡In the language of set notation, $\mathbb{R} \subseteq \mathbb{C}$.

is as you would expect. The only things you need to remember are the two properties above. The next example should help recall how these animals behave.

Example B.2.1: Perform the indicated operations.

1. $(1-2i)-(3+4i)$
2. $(1-2i)(3+4i)$
3. $\dfrac{1-2i}{3-4i}$
4. $\sqrt{-3}\sqrt{-12}$
5. $\sqrt{(-3)(-12)}$
6. $(x-[1+2i])(x-[1-2i])$

Solution:

1. As mentioned earlier, we treat expressions involving i as we would any other radical. We distribute and combine like terms:

$$\begin{array}{rcll}(1-2i)-(3+4i) & = & 1-2i-3-4i & \text{Distribute} \\ & = & -2-6i & \text{Gather like terms}\end{array}$$

   Technically, we'd have to rewrite our answer $-2-6i$ as $(-2)+(-6)i$ to be (in the strictest sense) 'in the form $a+bi$'. That being said, even pedants have their limits, so $-2-6i$ is good enough.

2. Using the Distributive Property (a.k.a. F.O.I.L.), we get

$$\begin{array}{rcll}(1-2i)(3+4i) & = & (1)(3)+(1)(4i)-(2i)(3)-(2i)(4i) & \text{F.O.I.L.} \\ & = & 3+4i-6i-8i^2 & \\ & = & 3-2i-8(-1) & i^2=-1 \\ & = & 3-2i+8 & \\ & = & 11-2i & \end{array}$$

3. How in the world are we supposed to simplify $\frac{1-2i}{3-4i}$? Well, we deal with the denominator $3-4i$ as we would any other denominator containing two terms, one of which is a square root. We multiply both numerator and denominator by $3+4i$, the (complex) conjugate of $3-4i$. Doing so produces

$$\begin{array}{rcll}\dfrac{1-2i}{3-4i} & = & \dfrac{(1-2i)(3+4i)}{(3-4i)(3+4i)} & \text{Equivalent Fractions} \\ & = & \dfrac{3+4i-6i-8i^2}{9-16i^2} & \text{F.O.I.L.} \\ & = & \dfrac{3-2i-8(-1)}{9-16(-1)} & i^2=-1 \\ & = & \dfrac{11-2i}{25} & \\ & = & \dfrac{11}{25}-\dfrac{2}{25}\,i & \end{array}$$

4. We use property 2 of Definition B.2.1 first, then apply the rules of radicals applicable to real numbers to get $\sqrt{-3}\sqrt{-12} = \left(i\sqrt{3}\right)\left(i\sqrt{12}\right) = i^2\sqrt{3\cdot 12} = -\sqrt{36} = -6$.

5. We adhere to the order of operations here and perform the multiplication before the radical to get $\sqrt{(-3)(-12)} = \sqrt{36} = 6$.

6. We brute force multiply using the distributive property and find that

$$\begin{aligned}(x-[1+2i])(x-[1-2i]) &= x^2 - x[1-2i] - x[1+2i] + [1-2i][1+2i] \\ &= x^2 - x + 2ix - x - 2ix + 1 - 2i + 2i - 4i^2 \\ &= x^2 - 2x + 1 - 4(-1) \\ &= x^2 - 2x + 5\end{aligned}$$

□

In the previous example, we used the 'conjugate' idea from simplifying radical equations to divide two complex numbers. More generally, the *complex conjugate* of a complex number $a+bi$ is the number $a-bi$. The notation commonly used for complex conjugation is a 'bar': $\overline{a+bi} = a-bi$. For example, $\overline{3+2i} = 3-2i$ and $\overline{3-2i} = 3+2i$. To find $\overline{6}$, we note that $\overline{6} = \overline{6+0i} = 6-0i = 6$, so $\overline{6} = 6$. Similarly, $\overline{4i} = -4i$, since $\overline{4i} = \overline{0+4i} = 0-4i = -4i$. Note that $\overline{3+\sqrt{5}} = 3+\sqrt{5}$, not $3-\sqrt{5}$, since $\overline{3+\sqrt{5}} = \overline{3+\sqrt{5}+0i} = 3+\sqrt{5}-0i = 3+\sqrt{5}$. Here, the conjugation specified by the 'bar' notation involves reversing the sign before $i = \sqrt{-1}$, not before $\sqrt{5}$. The properties of the conjugate are summarized in the following theorem.

Theorem B.2.1 (Properties of the Complex Conjugate)

Let z and w be complex numbers.

- $\overline{\overline{z}} = z$
- $\overline{z+w} = \overline{z} + \overline{w}$
- $\overline{zw} = \overline{z}\,\overline{w}$
- $\overline{z^n} = (\overline{z})^n$, for any natural number n
- z is a real number if and only if $\overline{z} = z$.

Theorem B.2.1 says in part that complex conjugation works well with addition, multiplication and powers. The proofs of these properties can best be achieved by writing out $z = a+bi$ and $w = c+di$ for real numbers a, b, c and d. Next, we compute the left and right sides of each equation and verify that they are the same.

The proof of the first property is a very quick exercise.* To prove the second property, we compare $\overline{z+w}$ with $\overline{z}+\overline{w}$. We have $\overline{z}+\overline{w} = \overline{a+bi} + \overline{c+di} = a-bi+c-di$. To find $\overline{z+w}$, we first compute

$$z+w = (a+bi)+(c+di) = (a+c)+(b+d)i$$

so

$$\overline{z+w} = \overline{(a+c)+(b+d)i} = (a+c)-(b+d)i = a+c-bi-di = a-bi+c-di = \overline{z}+\overline{w}$$

*Trust us on this.

As such, we have established $\overline{z+w} = \overline{z}+\overline{w}$. The proof for multiplication works similarly. The proof that the conjugate works well with powers can be viewed as a repeated application of the product rule, and is best proved using a technique called Mathematical Induction.The last property is a characterization of real numbers. If z is real, then $z = a+0i$, so $\overline{z} = a-0i = a = z$. On the other hand, if $z = \overline{z}$, then $a + bi = a - bi$ which means $b = -b$ so $b = 0$. Hence, $z = a + 0i = a$ and is real.

We now return to the business of solving quadratic equations. Consider $x^2 - 2x + 5 = 0$. The discriminant $b^2 - 4ac = -16$ is negative, so we know that there are no *real* solutions, since the Quadratic Formula would involve the term $\sqrt{-16}$. Complex numbers, however, are built just for such situations, so we can go ahead and apply the Quadratic Formula to get:

$$x = \frac{-(-2) \pm \sqrt{(-2)^2 - 4(1)(5)}}{2(1)} = \frac{2 \pm \sqrt{-16}}{2} = \frac{2 \pm 4i}{2} = 1 \pm 2i.$$

Example B.2.2: Find the complex solutions to the following equations.*

1. $\dfrac{2x}{x+1} = x + 3$
2. $2t^4 = 9t^2 + 5$
3. $z^3 + 1 = 0$

Solution:

1. Clearing fractions yields a quadratic equation so we then proceed via normal quadratic equation methods.

$$\begin{array}{rcll} \dfrac{2x}{x+1} & = & x+3 & \\ 2x & = & (x+3)(x+1) & \text{Multiply by } (x+1) \text{ to clear denominators} \\ 2x & = & x^2 + x + 3x + 3 & \text{F.O.I.L.} \\ 2x & = & x^2 + 4x + 3 & \text{Gather like terms} \\ 0 & = & x^2 + 2x + 3 & \text{Subtract } 2x \end{array}$$

From here, we apply the Quadratic Formula

$$\begin{array}{rcll} x & = & \dfrac{-2 \pm \sqrt{2^2 - 4(1)(3)}}{2(1)} & \text{Quadratic Formula} \\ & = & \dfrac{-2 \pm \sqrt{-8}}{2} & \text{Simplify} \\ & = & \dfrac{-2 \pm i\sqrt{8}}{2} & \text{Definition of } i \\ & = & \dfrac{-2 \pm i2\sqrt{2}}{2} & \text{Product Rule for Radicals} \\ & = & \dfrac{\cancel{2}(-1 \pm i\sqrt{2})}{\cancel{2}} & \text{Factor and reduce} \\ & = & -1 \pm i\sqrt{2} & \end{array}$$

We get two answers: $x = -1 + i\sqrt{2}$ and its conjugate $x = -1 - i\sqrt{2}$. Checking both of these answers reviews all of the salient points about complex number arithmetic and is therefore strongly encouraged.

*Remember, all real numbers are complex numbers, so 'complex solutions' means both real and non-real answers.

2. Since we have three terms, and the exponent on one term ('4' on t^4) is exactly twice the exponent on the other ('2' on t^2), we have a Quadratic in Disguise. We proceed accordingly.

$$\begin{array}{rcll} 2t^4 & = & 9t^2+5 & \\ 2t^4-9t^2-5 & = & 0 & \text{Subtract } 9t^2 \text{ and } 5 \\ (2t^2+1)(t^2-5) & = & 0 & \text{Factor} \\ 2t^2+1=0 & \text{or} & t^2=5 & \text{Zero Product Property} \end{array}$$

From $2t^2+1=0$ we get $2t^2=-1$, or $t^2=-\frac{1}{2}$. We extract square roots as follows:

$$t=\pm\sqrt{-\frac{1}{2}}=\pm i\sqrt{\frac{1}{2}}=\pm i\frac{\sqrt{1}}{\sqrt{2}}=\pm i\frac{1}{\sqrt{2}}=\pm\frac{i\sqrt{2}}{2},$$

where we have rationalized the denominator per convention. From $t^2=5$, we get $t=\pm\sqrt{5}$. In total, we have four complex solutions - two real: $t=\pm\sqrt{5}$ and two non-real: $t=\pm\frac{i\sqrt{2}}{2}$.

3. To find the *real* solutions to $z^3+1=0$, we can subtract the 1 from both sides and extract cube roots: $z^3=-1$, so $z=\sqrt[3]{-1}=-1$. It turns out there are two more non-real complex number solutions to this equation. To get at these, we factor:

$$\begin{array}{rcll} z^3+1 & = & 0 & \\ (z+1)(z^2-z+1) & = & 0 & \text{Factor (Sum of Two Cubes)} \\ z+1=0 & \text{or} & z^2-z+1=0 & \end{array}$$

From $z+1=0$, we get our real solution $z=-1$. From $z^2-z+1=0$, we apply the Quadratic Formula to get:

$$z=\frac{-(-1)\pm\sqrt{(-1)^2-4(1)(1)}}{2(1)}=\frac{1\pm\sqrt{-3}}{2}=\frac{1\pm i\sqrt{3}}{2}$$

Thus we get *three* solutions to $z^3+1=0$ - one real: $z=-1$ and two non-real: $z=\frac{1\pm i\sqrt{3}}{2}$. As always, the reader is encouraged to test their algebraic mettle and check these solutions.

⌟

It is no coincidence that the non-real solutions to the equations in Example B.2.2 appear in complex conjugate pairs. Any time we use the Quadratic Formula to solve an equation with real coefficients, the answers will form a complex conjugate pair owing to the $\pm$ in the Quadratic Formula.

Theorem B.2.2 (Discriminant Theorem)

Given a Quadratic Equation $ax^2 + bx + c = 0$, where a, b and c are real numbers, let $D = b^2 - 4ac$ be the discriminant.

- If $D > 0$, there are two distinct real number solutions to the equation.
- If $D = 0$, there is one (repeated) real number solution.
  'Repeated' here comes from the fact that 'both' solutions $\frac{-b \pm 0}{2a}$ reduce to $-\frac{b}{2a}$.
- If $D < 0$, there are two non-real solutions which form a complex conjugate pair.

B.3 Differentiation and Integration Techniques

In this section, we will cover some of the basic derivative and integral formulas that will be necessary for success in Differential Equations. In order to be able to deal with equations that involve derivatives, we need to be able to take derivatives as well as remove them.

B.3.1 Derivative and Integral Formulas

The following is a table of some of the basic derivative formulas covered in a Calculus 1 course.

Function $f(x)$	**Derivative** $f'(x)$
x^n any n	nx^{n-1}
$\ln(x)$	$\frac{1}{x} = x^{-1}$
C constant	0
e^x	e^x
e^{ax}	ae^{ax}
$\sin(x)$	$\cos(x)$
$\cos(x)$	$-\sin(x)$
$\tan(x)$	$\sec^2(x)$
$\arctan(x) = \tan^{-1}(x)$	$\frac{1}{x^2+1}$

Similarly, we have a table for some basic integral formulas. As integration is the inverse operation to differentiation, this table will look like the reverse version of the previous table.

Function $f(x)$	**Integral** $\int f(x)\ dx$
x^n any $n \neq -1$	$\frac{1}{n+1}x^{n+1} + C$
$\frac{1}{x}$	$\ln(\lvert x\rvert) + C$
e^x	$e^x + C$
e^{ax}	$\frac{1}{a}e^{ax} + C$
$\sin(x)$	$-\cos(x) + C$
$\cos(x)$	$\sin(x) + C$
$\frac{1}{x^2+1}$	$\arctan(x) + C$ or $\tan^{-1}(x) + C$

B.3.2 Derivative Rules

The tables above only list a few simple functions for which we know how to compute the derivative and integral. However, there are some nice properties of derivatives and integrals that make this enough for our needs.

Linearity of the Derivative and Integral

The derivative and integral are both linear operators. This means that if we have two functions $f(x)$ and $g(x)$, and two constants a and b, then

$$\frac{d}{dx}\left(af(x) + bg(x)\right) = a\frac{df}{dx} + b\frac{dg}{dx}.$$

That is, we can move constants and addition and subtractions out of the differentiation, reducing a complicated function down to simpler functions that we know how to differentiate.

The same is true for integration or antidifferentiation; if we have functions $f(x)$ and $g(x)$ and constants a and b, then

$$\int af(x) + bg(x)\ dx = a\int f(x)\ dx + b\int g(x)\ dx.$$

Example B.3.1: Compute the following derivatives and integrals using linearity and the table of known formulas.

1. $\frac{d}{dx}\left(x^3 + \frac{4}{x^2} + 3e^x\right)$
2. $\frac{d}{dx}\left(\sin(x) - 2\cos(x) + 5\ln(x)\right)$
3. $\int \frac{2x^3+4x}{x^2}\ dx$
4. $\int 2\cos(x) - \frac{3}{x^2+1}\ dx$

Solution:

1. For this, we can use linearity and our formulas to write

$$\begin{aligned}\frac{d}{dx}\left(x^3 + \frac{4}{x^2} + 3e^x\right) &= \frac{d}{dx}\left(x^3\right) + \frac{d}{dx}\left(\frac{4}{x^2}\right) + \frac{d}{dx}\left(3e^x\right)\\ &= \frac{d}{dx}\left(x^3\right) + 4\frac{d}{dx}\left(x^{-2}\right) + 3\frac{d}{dx}\left(e^x\right)\\ &= 3x^2 - 8x^{-3} + 3e^x.\end{aligned}$$

2. This one gives

$$\begin{aligned}\frac{d}{dx}\left(\sin(x) - 2\cos(x) + 5\ln(x)\right) &= \frac{d}{dx}\left(\sin(x)\right) - \frac{d}{dx}\left(2\cos(x)\right) + \frac{d}{dx}\left(5\ln(x)\right)\\ &= \frac{d}{dx}\left(\sin(x)\right) - 2\frac{d}{dx}\left(\cos(x)\right) + 5\frac{d}{dx}\left(\ln(x)\right)\\ &= \cos(x) + 2\sin(x) + \frac{5}{x}.\end{aligned}$$

3. For this problem, we first want to simplify the expression algebraically, then integrate each term using linearity.

$$\begin{aligned}\int \frac{2x^3 + 4x}{x^2}\ dx &= \int \frac{2x^3}{x^2} + \frac{4x}{x^2}\ dx\\ &= \int 2x + \frac{4}{x}\ dx\\ &= 2\int x\ dx + 4\int \frac{1}{x}\ dx\\ &= x^2 + 4\ln(|x|) + C.\end{aligned}$$

4. This problem uses standard linearity to get to the final answer.

$$\begin{aligned}\int 2\cos(x) - \frac{3}{x^2+1}\,dx &= 2\int \cos(x)\,dx - 3\int \frac{1}{x^2+1}\,dx \\ &= 2\sin(x) - 3\arctan(x) + C.\end{aligned}$$

Product and Quotient Rule

Linearity gives us a way to handle sums and differences of derivatives. What about products? It turns out that doesn't work as simply, but there is still as nice formula to work it out. This gives us the Product Rule. If we have two functions $f(x)$ and $g(x)$, then

$$\frac{d}{dx}\left(f(x)g(x)\right) = f(x)\frac{dg}{dx} + \frac{df}{dx}g(x).$$

That is, the derivative has two terms, the first function times the derivative of the second, and the derivative of the first function times the second function. The product rule can also be used to compute the product of more than two functions; the general formula is that only one function is differentiated at a time and each function should be differentiated once. That is, for three functions, the formula is

$$\frac{d}{dx}\left(f(x)g(x)h(x)\right) = \frac{df}{dx}g(x)h(x) + f(x)\frac{dg}{dx}h(x) + f(x)g(x)\frac{dh}{dx}.$$

The Quotient Rule gives us a way to do the same thing, but with quotients. The formula here is that

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{g(x)\frac{df}{dx} - f(x)\frac{dg}{dx}}{(g(x))^2}.$$

This can also be derived using the product rule and the chain rule. It is important to get the order of the numerator correct, as there is a subtraction on top. For the product rule, the addition means that the order doesn't matter, but if the order for the quotient rule is incorrect, there will be an additional minus sign in the answer.

Example B.3.2: Compute the following derivatives.

1. $\frac{d}{dx}\left(e^x\cos(x)\right)$
2. $\frac{d}{dx}\left(\frac{\sin(x)}{x^2}\right)$
3. $\frac{d}{dx}\left(\frac{x^3e^x}{\tan(x)}\right)$.

Solution:

1. This is a direct application of the product rule.

$$\begin{aligned}\frac{d}{dx}\left(e^x\cos(x)\right) &= e^x\frac{d}{dx}\left(\cos(x)\right) + \frac{d}{dx}\left(e^x\right)\cos(x) \\ &= e^x(-\sin(x)) + (e^x)\cos(x) \\ &= e^x(\cos(x) - \sin(x)).\end{aligned}$$

2. This is a direct application of the quotient rule.

$$\begin{aligned}\frac{d}{dx}\left(\frac{\sin(x)}{x^2}\right) &= \frac{x^2\frac{d}{dx}(\sin(x)) - \sin(x)\frac{d}{dx}(x^2)}{(x^2)^2} \\ &= \frac{x^2\cos(x) - \sin(x)(2x)}{x^4} \\ &= \frac{x\cos(x) - 2\sin(x)}{x^3}.\end{aligned}$$

3. For this problem, we need to apply both the product rule and the quotient rule. Since the quotient rule is on the outside, we apply it first.

$$\begin{aligned}\frac{d}{dx}\left(\frac{x^3e^x}{\tan(x)}\right) &= \frac{\tan(x)\frac{d}{dx}(x^3e^x) - x^3e^x\frac{d}{dx}(\tan(x))}{(\tan(x))^2} \\ &= \frac{\tan(x)\left(x^3\frac{d}{dx}(e^x) + \frac{d}{dx}(x^3)e^x\right) - x^3e^x\sec^2(x)}{\tan^2(x)} \\ &= \frac{\tan(x)(x^3e^x + 3x^2e^x) - x^3e^x\sec^2(x)}{\tan^2(x)} \\ &= \frac{e^x(x^3+3x^2)}{\tan(x)} - \frac{x^3e^x}{\sin^2(x)}.\end{aligned}$$

Chain Rule

The only type of function we haven't discussed yet for differentiation is composite functions, and that is handled by the Chain Rule. For example, we don't have a direct way (yet) to differentiate functions like $\sin(3x)$ or $\frac{1}{x^3+4x+1}$, and the Chain Rule lets us to do that. This rule tells us that, for functions $f(x)$ and $g(x)$, we can compute the derivative of the composition $(f \circ g)(x)$ or $f(g(x))$ is

$$\frac{d}{dx}(f(g(x)) = f'(g(x))g'(x).$$

This means that we differentiate the "outside" function f, plug in the inside function, and then multiply this by the derivative of the "inside" function g. It requires us to identify what the "inner" and "outer" functions are, and then the formula gives what the derivative should be. This can be done in a few different ways, either moving from outside in, or moving from inside out. These problems are conventionally written with $u(x)$ as the inside function, but any letter can be used.

Example B.3.3: Compute the derivative of each of the following functions.

1. $f_1(x) = (x^3 + 5x + 1)^5$
2. $f_2(x) = \cos(3x^2 + 1)$
3. $f_3(x) = (1 + \sin(3x))^4$

Solution:

1. For this problem, we take $f(u) = u^5$ and $u(x) = x^3 + 5x + 1$, which gives that composing these functions gives the f_1 that we started with. Therefore, since $f'(u) = 5u^4$ and $u'(x) = 3x^2 + 5$, we have that

$$f_1'(x) = f'(u)u'(x) = 5u^4(3x^2 + 5) = 5(x^3 + 5x + 1)^4(3x^2 + 5).$$

2. For this case, the outside function is $\cos(u)$ and the inner function is $u(x) = 3x^2 + 1$. Using the same process, we get that

$$f_2'(x) = -\sin(u)(6x) = -6x\sin(3x^2 + 1).$$

3. Starting from the outside, we see that we can take $f(u) = u^4$. This makes $u(x) = 1 + \sin(3x)$, but we can't differentiate this directly; it requires another iteration of the Chain Rule. Taking $u(x) = 1 + \sin(v)$ for $v(x) = 3x$, we can then compute the derivative of each of these functions, and our original function $f_3(x) = f(u(v(x)))$. We can extend the Chain Rule to apply to three functions by taking it one step at a time. The result of this process is that

$$\frac{d}{dx}(f(u(v(x)))) = f'(u(v(x))\frac{d}{dx}(u(v(x))) = f'(u(v(x)))u'(v(x))v'(x),$$

so you need to pull off one derivative at time to get to the correct computation. Thus, for this problem, we get that

$$f_3'(x) = 4u^3(\cos(v))(3) = 12u^3\cos(v) = 12(1 + \sin(3x))^3\cos(3x).$$

□

B.3.3 Integration Techniques

Another main topic that will be needed throughout study of differential equations is various integration techniques. When trying to solve questions that involve derivatives, integration will be a very important step in that process.

Substitution

The substitution method for integration serves as the inverse operation to the Chain Rule for differentiation. Since

$$\frac{d}{dx}(f(u(x))) = f'(u(x))u'(x),$$

the definition of the integral as an antiderivative gives that

$$\int f'(u(x))u'(x)\ dx = f(u(x)) + C.$$

Integrals of this form can be computed using this formula, but it is often easier to think of this process in terms of "changing variables." This means the following: If we have an integral that looks like

$$\int f'(u(x))u'(x)\ dx$$

then we can define the variable u to represent the entire function $u(x)$. Then the differential du is defined by

$$du = u'(x)dx.$$

Then we can substitute both u and du into the original expression to get that

$$\int f'(u(x))u'(x)\ dx = \int f'(u)\ du = f(u) + C = f(u(x)) + C.$$

The last component of this process is changing the limits of integration if a definite integral is being computed. The idea is that an integral in x (denoted by dx) has its limits also in terms of x, where as the du integral has endpoints given in terms of u. The main way this comes up in problems is that

$$\int_a^b f(u(x))u'(x)\ dx = \int_{u(a)}^{u(b)} f(u)\ du$$

because we know that u is written in terms of x as $u = u(x)$. Thus if we plug the x endpoints into this function, we will be the new u endpoints.

Example B.3.4: Compute the following integrals using substitution.

1. $\int \cos(4x)\ dx$
2. $\int x\sin(3x^2+1)\ dx$
3. $\int_0^2 \dfrac{3x^2}{x^3+4}\ dx$

Solution:

1. For this situation we want to set $u = 4x$, because then the integrand, once we make the change of variables, will be $\cos(u)$, which we know how to integrate. With this, we have $du = 4\ dx$, which we can rewrite as $dx = \frac{1}{4}\ du$. Plugging all of this in gives that

$$\int \cos(4x)\ dx = \int \cos(u)\frac{1}{4}\ du = \frac{1}{4}\int cos(u)\ du = \frac{1}{4}\sin(u) + C = \frac{1}{4}\sin(4x) + C.$$

2. For the same reason, we want to set $u = 3x^2+1$ to make the resulting integral $\sin(u)\ du$. In this case, we have $du = 6x\ dx$ or $x\ dx = \frac{1}{6}\ du$. Plugging all of this in, we get

$$\int x\sin(3x^2+1)\ dx = \int \sin(u)\frac{1}{6}\ du = \frac{1}{6}\int \sin(u)\ du = -\frac{1}{6}\cos(u)+C = -\frac{1}{6}\cos(3x^2+1)+C.$$

3. We can follow the same logic here as for the previous examples, but since we have a definite integral, we also need to switch the limits of integration. In this case, we want to pick $u = x^3 + 4$, which gives $du = 3x^2\ dx$. This gives the resulting integral as

$$\int \frac{3x^2}{x^3+4}\ dx = \int \frac{1}{u}\ du.$$

For the limits of integration, we take the function $u(x) = x^3 + 4$ and plug in the original values of 0 and 2. This gives the value 4 and $x = 0$ and the value 12 at $x = 2$. Therefore, the result of this computation is

$$\int_0^2 \frac{3x^2}{x^3+4}\ dx = \int_4^{12} \frac{1}{u}\ du = \ln(|u|)\ |_4^{12} = \ln(12) - \ln(4) = \ln(3).$$

There can be some cases where these techniques will not work, because the u' term that you are looking for doesn't quite appear in the expression you are trying to integrate. In cases like this, you may need to use some more complicated methods (like trigonometric substitution) or connect to inverse trigonometric integrals or other known formulas.

Integration by Parts

Integration by parts is the method used to handle integrals of a product of functions. Like the substitution method is the inverse of the Chain Rule, integration by parts is the inverse of the product rule. There are two main formulas that are used for this process. For two differentiable functions $f(x)$ and $g(x)$, we have

$$\int f(x)g'(x)\ dx = f(x)g(x) - \int g(x)f'(x)dx.$$

The other form is

$$\int u\ dv = uv - \int v\ du,$$

which matches the original form after setting $u = f(x)$ and $v = g(x)$.

The most important part of this process is picking the appropriate functions for u and v in this formula. The general rule is given by the following list

- Logarithmic functions
- Inverse Functions
- Algebraic or Polynomial Functions
- Trigonometric Functions (sine and cosine)
- Exponential Functions

and you want to make u, the function that you are differentiating, the one that is higher on the list. The main reason for this list is that integration is much harder than differentiation, and so we generally want to integrate the part of the product that we have a formula for. This is why logarithms and inverses are on the top; we know how to differentiate them, but integration is difficult or impossible. Polynomials are good for both differentiation and integrals, but the benefit of differentiating them is that they eventually disappear, leaving us with an integral that we know how to solve. For example, x^2 becomes $2x$, and then differentiating a second time gives 2, which is just a constant and can be removed from the integral. Trigonometric and Exponential functions are interchangeable, they are easy to differentiate and integrate, and they don't go away if we keep applying either operation.

This method can also be performed mutliple times by redefining u and v and applying the same process to the integral that remains on the right-hand side. When doing this, it is important not to reverse the roles of u and v, because then the process will just undo what was done in the first step. There are also some cases where circular reasoning is used, integrating by parts twice to get to the same expression on both sides of the equal sign, which can then be solved for. One of those will be shown in the examples below.

Example B.3.5: Compute the following integrals.

1. $\int x \sin(2x)\ dx$
2. $\int 3x^2 e^{4x}\ dx$
3. $\int e^{2x} \cos(3x)\ dx$

Solution:

1. Based on our list, we should choose $u = x$, as it is a polynomial function. This means that $dv = \sin(2x)\ dx$. From this, we get that $du = dx$ and we compute v by integrating $\sin(2x)\ dx$, which requires a substitution. This results in $v = -\frac{1}{2}\cos(2x)$. Thus, the integration by parts formula gives

$$\int x \sin(2x)\ dx = x\left(-\frac{1}{2}\cos(2x)\right) - \int \left(-\frac{1}{2}\cos(2x)\right) dx.$$

   This last integral we can compute directly, again requiring a substitution. Thus, the final answer is

$$\int x \sin(2x)\ dx = -\frac{x}{2}\cos(2x) + \frac{1}{4}\sin(2x) + C.$$

2. By the same argument as the first example, we want to pick $u = 3x^2$ so then $dv = e^{4x}dx$. We can then compute that $du = 6x\ dx$ and $v = \frac{1}{4}e^{4x}$. Thus, integration by parts gives

$$\int 3x^2 e^{4x}\ dx = 3x^2\left(\frac{1}{4}e^{4x}\right) - \int \left(\frac{1}{4}e^{4x}\right)(6x\ dx) = \frac{3}{4}x^2 e^{4x} - \int \frac{3}{2}xe^{4x}\ dx.$$

   This last integral is not something that we know how to compute. However, it looks like a product, so we should be able to work it out using integration by parts. We can

set $u = \frac{3}{2}x$ and $dv = e^{4x}\,dx$. This is the same dv as before, which is good. If we had picked $dv = \frac{3}{2}x\,dx$, we would have just gotten back to where we started. From these choices, we get that $du = \frac{3}{2}\,dx$ and $v = \frac{1}{4}e^{4x}$. Integration by parts then gives that

$$\int \frac{3}{2}xe^{4x}\,dx = \frac{3}{8}xe^{4x} - \int \frac{3}{8}e^{4x}\,dx.$$

Now we can compute this last integral, which will give another factor of $\frac{1}{4}$, resulting in

$$\int \frac{3}{2}xe^{4x}\,dx = \frac{3}{8}xe^{4x} - \frac{3}{32}e^{4x} + C.$$

Finally, we can combine this with our first integration by parts step to get that

$$\int 3x^2e^{4x}\,dx = \frac{3}{4}x^2e^{4x} - \frac{3}{8}xe^{4x} + \frac{3}{32}e^{4x} + C.$$

3. For this example, we have both an exponential and a trigonometric function. We can pick either one to be u and dv, and as long as we are consistent with that choice, we will get to the correct answer. For this, we will choose $u = e^{2x}$ and $dv = \cos(3x)\,dx$. From these, we can compute that $du = 2e^{2x}\,dx$ and $v = \frac{1}{3}\sin(3x)$. Thus, integration by parts tells us that

$$\int e^{2x}\cos(3x)\,dx = \frac{1}{3}e^{2x}\sin(3x) - \int \frac{2}{3}e^{2x}\sin(3x)\,dx.$$

This new integral is again a product, so we need to handle it using integration by parts. To do this, we are going to pick $u = \frac{2}{3}e^{2x}$ and $dv = \sin(3x)\,dx$. *Note:* If you pick $u = \sin(3x)$ and $dv = \frac{2}{3}e^{2x}\,dx$, the second integration by parts will just give that

$$\int e^{2x}\cos(3x)\,dx = \int e^{2x}\cos(3x)\,dx$$

which does not help in solving the problem. With the correct choice of u and dv, $u = \frac{2}{3}e^{2x}$ and $dv = \sin(3x)\,dx$, we have that $du = \frac{4}{3}e^{2x}\,dx$ and $v = -\frac{1}{3}\cos(3x)$, so that integration by parts tells us that

$$\int \frac{2}{3}e^{2x}\sin(3x)\,dx = -\frac{2}{9}e^{2x}\cos(3x) - \int -\frac{4}{9}e^{2x}\cos(3x)\,dx.$$

Combining this with our first integration by parts gives

$$\begin{aligned}\int e^{2x}\cos(3x)\,dx &= \frac{1}{3}e^{2x}\sin(3x) - \int \frac{2}{3}e^{2x}\sin(3x)\,dx \\ &= \frac{1}{3}e^{2x}\sin(3x) + \frac{2}{9}e^{2x}\cos(3x) - \int \frac{4}{9}e^{2x}\cos(3x)\,dx \\ \int e^{2x}\cos(3x)\,dx &= \frac{1}{3}e^{2x}\sin(3x) + \frac{2}{9}e^{2x}\cos(3x) - \frac{4}{9}\int e^{2x}\cos(3x)\,dx.\end{aligned}$$

In this case, we can see that the integral on the left matches the integral on the right. If we combine these on the left side, we get

$$\frac{13}{9}\int e^{2x}\cos(3x)\,dx = \frac{1}{3}e^{2x}\sin(3x) + \frac{2}{9}e^{2x}\cos(3x)$$

which then allows us to solve for the answers as

$$\int e^{2x}\cos(3x)\,dx = \frac{3}{13}e^{2x}\sin(3x) + \frac{2}{13}e^{2x}\cos(3x).$$

□

Partial Fractions

Another integration technique that shows up frequently when dealing with rational functions is the method of partial fractions. This method works around decomposing a rational function into forms that we are able to integrate. For example, we do not have a formula or method to compute the integral

$$\int \frac{3}{x^2 - x - 2}\,dx.$$

since there is no simple function whose derivative is $\frac{3}{x^2-x-2}$. What functions like this can we integrate?

Example B.3.6: Compute the following antiderivatives

$$\text{(a)} \int \frac{1}{x-2}\,dx \qquad \text{(b)} \int \frac{1}{x^2+4}\,dx \qquad \text{(c)} \int \frac{x}{x^2+9}\,dx.$$

Solution:

(a) This integral can be computed by a substitution $u = x - 2$,

$$\int \frac{1}{x-2}\,dx = \int \frac{1}{u}\,du = \ln|x-2| + C.$$

(b) This integral is another substitution, but the goal here is arctangent, not a logarithm. We let $u = x/2$, so that $du = 1/2\,dx$, and then

$$\int \frac{1}{x^2+4}\,dx = \int \frac{1}{4u^2+4}2\,du = \frac{1}{2}\int \frac{1}{u^2+1}\,du = \frac{1}{2}\arctan\left(\frac{x}{2}\right) + C.$$

(c) With an x on top of the expression, we can now use a substitution $u = x^2 + 9$ to solve the integral.

$$\int \frac{x}{x^2+9}\,dx = \frac{1}{2}\int \frac{1}{u}\,du = \frac{1}{2}\ln|x^2+9| + C.$$

So, we can handle these types of integrals, but that doesn't necessarily help us with the initial one. Let's take a look at another example.

Example B.3.7: Compute

$$\int \frac{1}{x-2} - \frac{1}{x+1}\, dx.$$

Solution: This integral can be computed by splitting it into the two terms present. Each of those we know how to evaluate using the previous example. Thus, we have that

$$\int \frac{1}{x-2} - \frac{1}{x+1}\, dx = \ln(|x-2|) - \ln(|x+1|) + C = \ln\left(\frac{|x-2|}{|x+1|}\right) + C.$$

This is great! However, we can compute that, by adding fractions

$$\frac{1}{x-2} - \frac{1}{x+1} = \frac{(x+1)-(x-2)}{(x-2)(x+1)} = \frac{3}{x^2-x-2}.$$

So this gives us a way to compute the original integral of this section, and we now know that

$$\int \frac{3}{x^2-x-2}\, dx = \ln\left(\frac{|x-2|}{|x+1|}\right) + C.$$

This gives an idea for how we may be able to evaluate integrals of rational functions. In the case of the integral above, we would need to figure out a way to convert between

$$\frac{3}{x^2-x-2} \text{ and } \frac{1}{x-2} - \frac{1}{x+1},$$

that is, we need to split the complicated fraction into the smaller, simpler partial fractions that we can integrate. Based on our work in Example B.3.6, we know that we can integrate functions that have a linear term in the denominator and a quadratic term in the denominator, and the process of putting these fractions together into a single term involves multiplying the individual denominators together. This gives the motivation for the method of partial fractions for integrating rational functions:

1. Factor the denominator of the function we need to integrate. Any polynomial can be factored into linear terms (terms like $x-a$) or irreducible quadratic terms (terms like x^2+4 or x^2+2x+5).

2. Write an expression with unknown coefficients for each factor in the expression. If it is a linear term, it will need just a single constant, but if there is a quadratic term, it needs a numerator of the form $Ax+B$.

3. Solve for the necessary constants (more on this later).

4. Integrate each of the resulting expressions, which are all forms where we know the antiderivative.

5. Combine the terms into a single expression.

The process is best shown through an example.

Example B.3.8: Compute

$$\int \frac{3x+1}{x^3-x^2-12x}\,dx.$$

Solution: We start by factoring the denominator. We can factor an x out of each term, and then the resulting quadratic can be factored. Since

$$x^3-x^2-12x = x(x+3)(x-4)$$

we want to figure out coefficients A, B, and C so that

$$\frac{3x+1}{x^3-x^2-12x} = \frac{A}{x} + \frac{B}{x+3} + \frac{C}{x-4}$$

where we have one term per factor of the denominator. In order to find these constants, our first trick is to multiply both sides of this equation by the entire denominator on the left. This gives

$$3x+1 = A(x+3)(x-4) + B(x)(x-4) + C(x)(x+3) \tag{B.1}$$

where we have cancelled the appropriate terms from the top and bottom of each expression. One way to go from here to the constants is to expand out the right-hand side and recognize that for these two sides to be equal for all x, the coefficient of x^2, x, and the constant term must match. This will result in solve a system of 3 equations.

An easier approach to doing this is to plug values for x into each side, and to pick those values cleverly. One clever choice for (B.1) is to set $x = 0$. If we do that, both the B and C terms will go away, because they are multiplied by zero. Thus, if we plug in zero, we get

$$1 = -12A + 0 + 0$$

which implies that $A = -1/12$. For the next term, we can plug in -3 to make the $x+3$ terms go away, resulting in

$$-8 = B(-3)(-7)$$

so that $B = -8/21$. Plugging in $x = 4$ gives

$$13 = C(4)(7)$$

so that $C = 13/28$. Therefore, we can write that

$$\frac{3x+1}{x^3-x^2-12x} = \frac{-1/12}{x} + \frac{-8/21}{x+3} + \frac{13/28}{x-4}$$

Then, we can integrate both sides to get that

$$\int \frac{3x+1}{x^3-x^2-12x}\,dx = -\frac{1}{12}\ln(|x|) - \frac{8}{21}\ln(|x+3|) + \frac{13}{28}\ln(|x-4|) + C.$$

The same type of approach applies if there are irreducible quadratics in the expression.

Example B.3.9: Compute

$$\int \frac{2x^2 - 6}{x^3 - x^2 + 4x - 4}\, dx.$$

Solution: The denominator can be factored as $(x^2 + 4)(x - 1)$, which can be determined by grouping. This means that to do the partial fraction decomposition, we need to find coefficients A, B, and C so that

$$\frac{2x^2 - 6}{x^3 - x^2 + 4x - 4} = \frac{Ax + B}{x^2 + 4} + \frac{C}{x - 1}.$$

Note that the x^2+4 term has $Ax+B$ on top instead of just A. This is because the term on the bottom is a quadratic, and there will always be a number of coefficients on top that matches the degree of the term in the denominator. By multiplying both sides by the denominator will give the equation

$$2x^2 - 6 = (Ax + B)(x - 1) + C(x^2 + 4)$$

where we need to find the appropriate constants. In this case, we can plug in $x = 1$ to determine that $-4 = 5C$ or $C = -4/5$. However, there is no value we can plug in to make $x^2 + 4 = 0$. We could use complex numbers here, but assuming we don't want to do that, we can plug in any two numbers and go from there. Plugging in $x = 0$ is nice because it makes the A term go away, resulting in

$$-6 = B(-1) + C(4) = -B - \frac{16}{5}$$

which we can solve to get $B = 14/5$. Finally, we can plug in any other number for x to get an equation to solve for A. Let's use -1 to give that

$$-4 = \left(-A + \frac{14}{5}\right)(-2) + \left(-\frac{4}{5}\right)(5) = 2A - \frac{28}{5} - 4$$

which gives that $A = -14/5$. Therefore, we can write

$$\frac{2x^2 - 6}{x^3 - x^2 + 4x - 4} = \frac{-14/5 x + 14/5}{x^2 + 4} + \frac{-4/5}{x - 1}.$$

Therefore, we can write the integral we want to compute as

$$\begin{aligned}
\int \frac{2x^2 - 6}{x^3 - x^2 + 4x - 4}\, dx &= \int \frac{-14/5 x + 14/5}{x^2 + 4} + \frac{-4/5}{x - 1}\, dx \\
&= -\frac{14}{5}\int \frac{x}{x^2 + 4}\, dx + \frac{14}{5}\int \frac{1}{x^2 + 4}\, dx - \frac{4}{5}\int \frac{1}{x - 1}\, dx \\
&= -\frac{7}{5}\ln(|x^2 + 4|) + \frac{7}{5}\arctan\left(\frac{x}{2}\right) - \frac{4}{5}\ln(|x - 1|) + C
\end{aligned}$$

There are a few extra complications that can result from using this method.

1. If there is an irreducible quadratic like $x^2 + 2x + 5$ in the denominator, we will want to separate that out and complete the square before integrating. In this case, we have $x^2 + 2x + 5 = (x + 1)^2 + 4$, so we will want to use $A(x + 1) + B$ when solving for coefficients (to make the u-substitution work better), and will get a slightly more complicated result.

2. If there are repeated factors, like $(x - 1)^2$ in the denominator, we need to include one term in the partial fraction expansion for every power of that factor. For instance, the expansion should look like

$$\frac{1}{(x+1)(x-3)^3} = \frac{A}{x+1} + \frac{B}{x-3} + \frac{C}{(x-3)^2} + \frac{D}{(x-3)^3}.$$

3. If the rational function has an equal or higher degree in the numerator than in the denominator, we will need to do long division to remove a standard polynomial (which we know how to integrate) and a proper rational function that can be integrated using partial fractions.

Combining all of these techniques together will allow us to integrate pretty much any rational function that we need for a given application.

Further Reading

[BM] Paul W. Berg and James L. McGregor, *Elementary Partial Differential Equations*, Holden-Day, San Francisco, CA, 1966.

[BD] William E. Boyce and Richard C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, 11th edition, John Wiley & Sons Inc., New York, NY, 2017.

[EP] C.H. Edwards and D.E. Penney, *Differential Equations and Boundary Value Problems: Computing and Modeling*, 5th edition, Pearson, 2014.

[F] Stanley J. Farlow, *An Introduction to Differential Equations and Their Applications*, McGraw-Hill, Inc., Princeton, NJ, 1994. (Published also by Dover Publications, 2006.)

[I] E.L. Ince, *Ordinary Differential Equations*, Dover Publications, Inc., New York, NY, 1956.

[JL] Jiří Lebl, *Notes on Diffy Qs*, Open-source publication, https://www.jirka.org/diffyqs.

[SZ] Carl Stitz and Jeff Zeager, *Precalculus*. Version 4. 2017. https://www.stitz-zeager.com/

[T] William F. Trench, *Elementary Differential Equations with Boundary Value Problems*. Books and Monographs. Book 9. 2013. https://digitalcommons.trinity.edu/mono/9

[ZW] Dennis Zill and Warren Wright, *Advanced Engineering Mathematics*, 6th Edition, Jones & Bartlett Learning, 2016.

Answers to Selected Exercises

0.1.5: Compute $x' = -2e^{-2t}$ and $x'' = 4e^{-2t}$. Then $(4e^{-2t}) + 4(-2e^{-2t}) + 4(e^{-2t}) = 0$.

0.1.8: Yes.

0.1.10: $y = x^r$ is a solution for $r = 0$ and $r = 2$.

0.1.13: $C_1 = 100$, $C_2 = -90$

0.1.15: $\varphi = -9e^{8s}$

0.1.17: a) $x = 9e^{-4t}$ b) $x = \cos(2t) + \sin(2t)$ c) $p = 4e^{3q}$ d) $T = 3\sinh(2x)$

0.2.2: a) PDE, equation, second order, linear, nonhomogeneous, constant coefficient.
b) ODE, equation, first order, linear, nonhomogeneous, not constant coefficient, not autonomous.
c) ODE, equation, seventh order, linear, homogeneous, constant coefficient, autonomous.
d) ODE, equation, second order, linear, nonhomogeneous, constant coefficient, autonomous.
e) ODE, system, second order, nonlinear.
f) PDE, equation, second order, nonlinear.

0.2.6: equation: $a(x)y = b(x)$, solution: $y = \frac{b(x)}{a(x)}$.

0.2.7: $k = 0$ or $k = 1$

0.2.9: b) First order with three components.
c) Third order with one component.
d) The product is three in both cases. $(1 \times 3 = 3 \times 1)$.

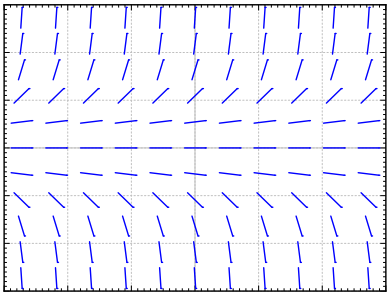
1.1.4: $y = e^x + \frac{x^2}{2} + 9$

1.1.11: 170

1.1.15: The equation is $r' = -C$ for some constant C. The snowball will be completely melted in 25 minutes from time $t = 0$.

1.1.16: $y = Ax^3 + Bx^2 + Cx + D$, so 4 constants.

1.2.2:



$y = 0$ is a solution such that $y(0) = 0$.

1.2.9: a) $y' = \cos y$, b) $y' = y\cos(x)$, c) $y' = \sin x$. Justification left to reader.

1.3.2: $x = (3t-2)^{1/3}$

1.3.4: $x = \sin^{-1}\big(t+1\big)$

1.3.8: a) $\frac{y^2}{2} = x^2 + C$ b) $y = 2\sqrt{x^2+3}$ c) $y = -2\sqrt{x^2+1}$

1.3.9: If $n \neq 1$, then $y = \big((1-n)x+1\big)^{1/(1-n)}$. If $n = 1$, then $y = e^x$.

1.3.12: $y = Ce^{x^2}$

1.3.15: $x = e^{t^3} + 1$

1.3.18: $x^3 + x = t + 2$

1.3.21: $\sin(y) = -\cos(x) + C$

1.3.24: $y = \frac{1}{1-\ln x}$

1.3.33: The range is approximately 7.45 to 12.15 minutes.

1.3.34: a) $x = \frac{1000e^t}{e^t+24}$. b) 102 rabbits after one month, 861 after 5 months, 999 after 10 months, 1000 after 15 months.

1.4.13: $y = Ce^{-x^3} + 1/3$

1.4.19: $y = 2e^{\cos(2x)+1} + 1$

1.5.16: Yes a solution exists. $y' = f(x,y)$ where $f(x,y) = xy$. The function $f(x,y)$ is continuous and $\frac{\partial f}{\partial y} = x$, which is also continuous near $(0,0)$. So a solution exists and is unique. (In fact $y = 0$ is the solution).

1.5.17: No, the equation is not defined at $(x,y) = (1,0)$.

1.5.18: Picard does not apply as f is not continuous at $y = 0$. The equation does not have a continuously differentiable solution. Suppose it did. Notice that $y'(0) = 1$. By the first derivative test, $y(x) > 0$ for small positive x. But then for those x we would have $y'(x) = 0$, so clearly the derivative cannot be continuous.

1.5.19: The solution is $y(x) = \int_{x_0}^{x} f(s)\,ds + y_0$, and this does indeed exist for every x.

1.6.7: Approximately: 1.0000, 1.2397, 1.3829

1.6.9: a) 0, 8, 12 b) $x(4) = 16$, so errors are: 16, 8, 4. c) Factors are 0.5, 0.5, 0.5.

1.6.10: a) 0, 0, 0 b) $x = 0$ is a solution so errors are: 0, 0, 0.

1.6.12: a) Improved Euler: $y(1) \approx 3.3897$ for $h = 1/4$, $y(1) \approx 3.4237$ for $h = 1/8$, b) Standard Euler: $y(1) \approx 2.8828$ for $h = 1/4$, $y(1) \approx 3.1316$ for $h = 1/8$, c) $y = 2e^x - x - 1$, so $y(2)$ is approximately 3.4366. d) Approximate errors for improved Euler: 0.046852 for $h = 1/4$, and 0.012881 for $h = 1/8$. For standard Euler: 0.55375 for $h = 1/4$, and 0.30499 for $h = 1/8$. Factor is approximately 0.27 for improved Euler, and 0.55 for standard Euler.

1.7.4: a) 0, 1, 2 are critical points. b) $x = 0$ is unstable (semistable), $x = 1$ is asymptotically stable, and $x = 2$ is unstable. c) 1

1.7.9: a) There are no critical points. b) ∞

1.7.11: a) α is a stable critical point, β is an unstable one. b) α, c) α, d) ∞ or DNE.

1.8.3: a) $\frac{dx}{dt} = kx(M-x) + A$ b) $\frac{kM+\sqrt{(kM)^2+4Ak}}{2k}$

1.9.3: a) $e^{xy} + \sin(x) = C$ b) $x^2 + xy - 2y^2 = C$ c) $e^x + e^y = C$ d) $x^3 + 3xy + y^3 = C$

1.9.10: a) Integrating factor is y, equation becomes $dx + 3y^2\,dy = 0$. b) Integrating factor is e^x, equation becomes $e^x\,dx - e^{-y}\,dy = 0$. c) Integrating factor is y^2, equation becomes $(\cos(x) + y)\,dx + x\,dy = 0$. d) Integrating factor is x, equation becomes $(2xy + y^2)\,dx + (x^2 + 2xy)\,dy = 0$.

1.9.15: a) The equation is $-f(x)\,dx + \frac{1}{g(y)}\,dy$, and this is exact because $M = -f(x)$, $N = \frac{1}{g(y)}$, so $M_y = 0 = N_x$. b) $-x\,dx + \frac{1}{y}\,dy = 0$, leads to potential function $F(x, y) = -\frac{x^2}{2} + \ln|y|$, solving $F(x, y) = C$ leads to the same solution as the example.

1.10.5: 250 grams

1.10.9: $P(5) = 1000e^{2\times 5 - 0.05\times 5^2} = 1000e^{8.75} \approx 6.31 \times 10^6$

1.10.10: $Ah' = I - kh$, where k is a constant with units $\mathrm{m}^2 s$.

1.11.4: $\alpha = .123$. The alpha value used before noise was added to the data is 0.124, so very close, but not identically the same.

1.11.5: a) $\alpha = 4.03 \times 10^{-5}$, so $\alpha \approx 0$. c) $K = 324.07$ and $\alpha = 5.061$.

1.12.2: $y = \frac{2}{3x-2}$

1.12.4: $y = \frac{3-x^2}{2x}$

1.12.9: $y = \left(7e^{3x} + 3x + 1\right)^{1/3}$

1.12.13: $y = \sqrt{x^2 - \ln(C - x)}$

2.1.5: Yes. To justify try to find a constant A such that $\sin(x) = Ae^x$ for all x.

2.1.6: No. $e^{x+2} = e^2 e^x$.

2.1.7: $y = 5$

2.1.13: $y = C_1 \ln(x) + C_2$

2.1.21: $y = C_1 e^{(-2+\sqrt{2})x} + C_2 e^{(-2-\sqrt{2})x}$

2.1.22: $y = \frac{2(a-b)}{5}\,e^{-3x/2} + \frac{3a+2b}{5}\,e^x$

2.1.23: $y = \frac{a\beta - b}{\beta - \alpha} e^{\alpha x} + \frac{b - a\alpha}{\beta - \alpha} e^{\beta x}$

2.1.24: $y'' - 3y' + 2y = 0$

2.1.25: $y'' - y' - 6y = 0$

2.2.5: $3\sqrt{2}\cos\left(2x - \frac{\pi}{4}\right)$

2.2.13: $y = e^{-x/4}\cos\left((\sqrt{7}/4)x\right) - \sqrt{7}e^{-x/4}\sin\left((\sqrt{7}/4)x\right)$

2.2.14: $z(t) = 2e^{-t}\cos(t)$

2.2.17: There is no such equation. The two roots will always be complex conjugates, which means the exponential parts will match, and the trigonometric functions will have the same argument.

2.3.4: $y = C_1 e^{3x} + C_2 x e^{3x}$

2.3.10: c) $y(x) = C_1 x + C_2 \frac{1}{x^3}$

2.3.11: c) $y(x) = C_1 \frac{1}{x} + C_2 \frac{1}{x^2}$

2.3.12: c) $y(x) = C_1x^2 + C_2x^5$

2.4.5: $k = 8/9$ (and larger)

2.4.8: a) $k = 500000$ b) $\frac{1}{5\sqrt{2}} \approx 0.141$ c) 45000 kg d) 11250 kg

2.4.10: $m_0 = \frac{1}{3}$. If $m < m_0$, then the system is overdamped and will not oscillate.

2.4.11: a) $0.05I'' + 0.1I' + (1/5)I = 0$ b) $I = Ce^{-t}\cos(\sqrt{3}\,t - \gamma)$ or $I = C_1e^{-t}\cos(\sqrt{3}t) + C_2e^{-t}\sin(\sqrt{3}t)$ c) $I = 10e^{-t}\cos(\sqrt{3}\,t) + \frac{10}{\sqrt{3}}e^{-t}\sin(\sqrt{3}\,t)$

2.5.5: $y = \frac{-16\sin(3x)+6\cos(3x)}{73}$

2.5.9: $y(x) = x^2 - 4x + 6 + e^{-x}(x - 5)$

2.5.12: a) $y = \frac{2e^x+3x^3-9x}{6}$ b) $y = C_1\cos(\sqrt{2}x) + C_2\sin(\sqrt{2}x) + \frac{2e^x+3x^3-9x}{6}$

2.5.23: $y = \frac{2xe^x-(e^x+e^{-x})\log(e^{2x}+1)}{4}$

2.5.25: $y = \frac{-\sin(x+c)}{3} + C_1e^{\sqrt{2}\,x} + C_2e^{-\sqrt{2}\,x}$

2.6.6: $x_{sp} = \frac{(\omega_0^2-\omega^2)F_0}{m(2\omega p)^2+m(\omega_0^2-\omega^2)^2}\cos(\omega t) + \frac{2\omega pF_0}{m(2\omega p)^2+m(\omega_0^2-\omega^2)^2}\sin(\omega t) + \frac{A}{k}$, where $p = \frac{\gamma}{2m}$ and $\omega_0 = \sqrt{\frac{k}{m}}$.

2.6.9: $\omega = \frac{\sqrt{31}}{4\sqrt{2}} \approx 0.984$ $C(\omega) = \frac{16}{3\sqrt{7}} \approx 2.016$

2.6.12: a) $\omega = 2$ b) 25

2.7.3: $y = C_1e^x + C_2x^3 + C_3x^2 + C_4x + C_5$

2.7.8: a) $r^3 - 3r^2 + 4r - 12 = 0$ b) $y''' - 3y'' + 4y' - 12y = 0$ c) $y = C_1e^{3x} + C_2\sin(2x) + C_3\cos(2x)$

2.7.10: $y(x) = C_1e^{4x} + C_2e^{-x} + C_3e^{-x}\cos(2x) + C_4e^{-x}\sin(2x)$

2.7.19: No. $e^1e^x - e^{x+1} = 0$.

2.7.22: Yes. (Hint: First note that $\sin(x)$ is bounded. Then note that x and $x\sin(x)$ cannot be multiples of each other.)

2.7.24: $y = 0$

2.7.28: $y''' - y'' + y' - y = 0$

3.1.5: a) $\sqrt{10}$ b) $\sqrt{14}$ c) 3

3.1.7: a) $\begin{bmatrix}9\\-2\end{bmatrix}$ b) $\begin{bmatrix}-3\\3\end{bmatrix}$ c) $\begin{bmatrix}5\\-3\end{bmatrix}$ d) $\begin{bmatrix}-4\\8\end{bmatrix}$ e) $\begin{bmatrix}3\\7\end{bmatrix}$ f) $\begin{bmatrix}-8\\3\end{bmatrix}$

3.1.9: a) $\begin{bmatrix}\frac{-1}{\sqrt{2}}\\0\\\frac{1}{\sqrt{2}}\end{bmatrix}$ b) $\begin{bmatrix}\frac{1}{\sqrt{6}}\\\frac{-1}{\sqrt{6}}\\\frac{2}{\sqrt{6}}\end{bmatrix}$ c) $\left(\frac{2}{\sqrt{33}}, \frac{-5}{\sqrt{33}}, \frac{2}{\sqrt{33}}\right)$

3.1.14: a) 20 b) 10 c) 20

3.1.18: a) $(3, -1)$ b) $(4, 0)$ c) $(-1, -1)$

3.2.2: a) $\begin{bmatrix}7&4&4\\2&3&4\end{bmatrix}$ b) $\begin{bmatrix}5&-3&0\\13&10&6\\-1&3&1\end{bmatrix}$

3.2.4: a) $\begin{bmatrix} -1 & 13 \\ 9 & 14 \end{bmatrix}$ b) $\begin{bmatrix} 2 & -5 \\ 5 & 5 \end{bmatrix}$

3.2.6: a) $\begin{bmatrix} 22 & 31 \\ 42 & 44 \end{bmatrix}$ b) $\begin{bmatrix} 18 & 18 & 12 \\ 6 & 0 & 8 \\ 34 & 48 & -2 \end{bmatrix}$ c) $\begin{bmatrix} 11 & 12 & 36 & 14 \\ -2 & 4 & 5 & -2 \\ 13 & 38 & 20 & 28 \end{bmatrix}$ d) $\begin{bmatrix} -2 & -12 \\ 3 & 24 \\ 1 & 9 \end{bmatrix}$

3.2.11: a) $\begin{bmatrix} 1/2 \end{bmatrix}$ b) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ c) $\begin{bmatrix} -5 & 2 \\ 3 & -1 \end{bmatrix}$ d) $\begin{bmatrix} 1/2 & -1/4 \\ -1/2 & 1/2 \end{bmatrix}$

3.2.13: a) $\begin{bmatrix} 1/2 & 0 \\ 0 & 1/3 \end{bmatrix}$ b) $\begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & -1 \end{bmatrix}$ c) $\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix}$

3.3.2: a) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ b) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ c) $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$ d) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1/3 \\ 0 & 0 & 0 \end{bmatrix}$ e) $\begin{bmatrix} 1 & 0 & 0 & 77/15 \\ 0 & 1 & 0 & -2/15 \\ 0 & 0 & 1 & -8/5 \end{bmatrix}$

f) $\begin{bmatrix} 1 & 0 & -1/2 & 0 \\ 0 & 1 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ g) $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ h) $\begin{bmatrix} 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

3.3.4: a) $x_1 = -2$, $x_2 = 7/3$ b) no solution c) $a = -3$, $b = 10$, $c = -8$ d) x_3 is free, $x_1 = -1 + 3x_3$, $x_2 = 2 - x_3$

3.3.5: $x_1 = 4$, $x_2 = -3$, $x_3 = -2$, $x_4 = 3$

3.3.6: $x_1 = -4$, $x_2 = -1$, $x_3 = 1$, $x_4 = 2$

3.3.7: No solution exists.

3.3.8: Infinitely many solutions of the form $x_1 = 19t - 37$, $x_2 = 37 - 20t$, $x_3 = 15t - 37$, $x_4 = t$ for any real number t.

3.3.9: There is no solution.

3.3.10: There are infinitely many solutions of the form $x_1 = 2 - t$, $x_2 = 4 - 2t$, $x_3 = t$ for any real number t.

3.3.12: The work is not correct. It looks like the author used row 1 to try to cancel the second column from rows 2 and 3, which we can not do. The correct method would be to use row 2 to cancel row 3, resulting in a solution $x_1 = 9$, $x_2 = -25$, and $x_3 = 10$.

3.4.2: a) 3 b) 1 c) 2

3.4.5: a) $\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$, $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$ b) $\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$ c) $\begin{bmatrix} 1 & 0 & 1/3 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 & -1/3 \end{bmatrix}$

3.4.6: a) $\begin{bmatrix} 7 \\ 7 \\ 7 \end{bmatrix}, \begin{bmatrix} -1 \\ 7 \\ 6 \end{bmatrix}, \begin{bmatrix} 7 \\ 6 \\ 2 \end{bmatrix}$ b) $\begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$ c) $\begin{bmatrix} 0 \\ 6 \\ 4 \end{bmatrix}, \begin{bmatrix} 3 \\ 3 \\ 7 \end{bmatrix}$

3.4.7: 3

3.4.8: 4

3.4.10: $\begin{bmatrix} 3 \\ 1 \\ -5 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \\ -1 \end{bmatrix}$

3.4.12: a) $\begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}$ dimension 2, b) $\begin{bmatrix}1\\1\\1\end{bmatrix}, \begin{bmatrix}1\\1\\2\end{bmatrix}$ dimension 2, c) $\begin{bmatrix}5\\3\\1\end{bmatrix}, \begin{bmatrix}5\\-1\\5\end{bmatrix}, \begin{bmatrix}-1\\3\\-4\end{bmatrix}$ dimension 3, d) $\begin{bmatrix}2\\2\\4\end{bmatrix}, \begin{bmatrix}2\\2\\3\end{bmatrix}$ dimension 2, e) $\begin{bmatrix}1\\1\end{bmatrix}$ dimension 1, f) $\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\2\end{bmatrix}$ dimension 2

3.4.14:

a) Put the vectors as the columns of a matrix and row reduce. If there are any non-pivot columns, the vectors are linearly dependent.

b) No, there can be at most three pivot columns, so with four columns, one must be non-pivot.

c) Yes, there is no reason you can't have all of the two columns being pivot columns.

d) Put the vectors as the columns of a matrix, and look for solutions to $A\vec{x} = \vec{b}$. We need the rank of this matrix to be at least 3.

e) Yes, the matrix with four columns can have rank three.

f) No, it is impossible for a matrix with only two columns to have rank three.

3.4.15:

a) The rank is 2.

b) No, it is not in the span.

c) Yes, it is in the span, because the first vector is exactly $\vec{b}$.

d) This says that these two spans are not the same. We can not use the row-reduced matrix in order to figure out if something is in the span. We need to use the pivot columns to go back to the original vectors to simplify the span.

e)

$$D_2 = \begin{bmatrix}1 & -1 & -1\\0 & 1 & 1/2\\0 & 0 & 0\end{bmatrix}$$

f) No, it is not. If we add the two rows together, we get $[1\ 0\ -1/2]$ and we have no way to cancel out that last term. This suggests that we can use either the rows of the original matrix or the rows of the row-reduced form in order to work out the span of the rows.

3.5.3: a) -2 b) 8 c) 0 d) -6 e) -3 f) 28 g) 16 h) -24

3.5.5: a) 3 b) 9 c) 3 d) 1/4

3.5.6: -10

3.5.7: 6

3.5.8: 6

3.5.10: Rank is 3. Therefore A is not invertible (since the rank is not 4), and there are non-zero solutions to $A\vec{x} = \vec{0}$.

3.5.11: Rank is 3. Therefore A is invertible, and there is exactly one solution to $A\vec{x} = \begin{bmatrix}1\\1\\1\end{bmatrix}$, namely $A^{-1}\begin{bmatrix}1\\1\\1\end{bmatrix}$.

3.5.12: -1. The only solution is $\vec{x} = 0$.

3.5.13: 2. The colums are linearly independent.

3.5.14: 8. There is exactly one solution, found by row reduction or multiplying by A^{-1}.

3.5.16: 1/12

3.5.19: 1 and 3

3.6.1: $\lambda_1 = -2$, $\vec{v}_1 = \begin{bmatrix}-3\\1\end{bmatrix}$, $\lambda_2 = 4$, $\vec{v}_2 = \begin{bmatrix}3\\-2\end{bmatrix}$.

3.6.2: $\lambda_1 = -2$, $\vec{v}_1 = \begin{bmatrix}1\\4\end{bmatrix}$, $\lambda_2 = -4$, $\vec{v}_2 = \begin{bmatrix}0\\1\end{bmatrix}$.

3.6.3: $\lambda_1 = -4$, $\vec{v}_1 = \begin{bmatrix}1\\3\end{bmatrix}$, $\lambda_2 = -3$, $\vec{v}_2 = \begin{bmatrix}1\\4\end{bmatrix}$.

3.6.4: $\lambda_1 = 3 + 2i$, $\vec{v}_1 = \begin{bmatrix}3-i\\4\end{bmatrix}$, $\lambda_2 = 3 - 2i$, $\vec{v}_2 = \begin{bmatrix}3+i\\4\end{bmatrix}$.

3.6.5: $\lambda_1 = -1 + i$, $\vec{v}_1 = \begin{bmatrix}2\\-1+i\end{bmatrix}$, $\lambda_2 = -1 - i$, $\vec{v}_2 = \begin{bmatrix}2\\-1-i\end{bmatrix}$.

3.6.6: $\lambda_1 = -2 + 2i$, $\vec{v}_1 = \begin{bmatrix}1-i\\4\end{bmatrix}$, $\lambda_2 = -2 - 2i$, $\vec{v}_2 = \begin{bmatrix}1+i\\4\end{bmatrix}$.

3.6.7: $\lambda_1 = 4$, $\vec{v}_1 = \begin{bmatrix}-2\\1\end{bmatrix}$

3.6.8: $\lambda_1 = -3$, $\vec{v}_1 = \begin{bmatrix}1\\2\end{bmatrix}$

3.6.9: $\lambda_1 = 2$, $\vec{v}_1 = \begin{bmatrix}0\\1\\0\end{bmatrix}$, $\lambda_2 = 1$, $\vec{v}_2 = \begin{bmatrix}0\\-1\\1\end{bmatrix}$, $\lambda_3 = 4$, $\vec{v}_3 = \begin{bmatrix}1\\-3\\-2\end{bmatrix}$

3.6.10: $\lambda_1 = -4$, $\vec{v}_1 = \begin{bmatrix}1\\3\\-3\end{bmatrix}$, $\lambda_2 = -3$, $\vec{v}_2 = \begin{bmatrix}0\\-1\\1\end{bmatrix}$, $\lambda_3 = -1$, $\vec{v}_3 = \begin{bmatrix}3\\0\\1\end{bmatrix}$

3.6.11: $\lambda_1 = 1+3i$, $\vec{v}_1 = \begin{bmatrix} 0 \\ 2 \\ -1+i \end{bmatrix}$, $\lambda_2 = 1-3i$, $\vec{v}_2 = \begin{bmatrix} 0 \\ 2 \\ -1-i \end{bmatrix}$, $\lambda_3 = -2$, $\vec{v}_3 = \begin{bmatrix} 1 \\ 2 \\ -2 \end{bmatrix}$

3.6.12: $\lambda_1 = 2$, $\vec{v}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$, $\lambda_2 = 1$, $\vec{v}_2 = \begin{bmatrix} 0 \\ -2 \\ 1 \end{bmatrix}$ (double root)

3.6.13: $\lambda_1 = -2$, $\vec{v}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$, $\lambda_2 = 1$, $\vec{v}_2 = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$.

3.6.14: $\lambda_1 = 2$, $\vec{v}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, Generalized eigenvector $\vec{w} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

3.6.15:

a) $(-2-\lambda)\left[(5-\lambda)(1-\lambda)+12\right]+36 = (2-\lambda)(5-\lambda)(1-\lambda)-12(2+\lambda)-12(-3)$ which can be regrouped as $(2-\lambda)(5-\lambda)(1-\lambda)-12(2+\lambda)-12(-3) = (2-\lambda)(5-\lambda)(1-\lambda)+12(1-\lambda)$ and can then be factored as $(1-\lambda)(\lambda^2-5\lambda+2\lambda-10+12) = (1-\lambda)(\lambda-1)(\lambda-2)$.

b) $r_1 = 1$ with algebraic multiplicity 2, and $r_2 = 2$ with algebraic multiplicity 1.

c) [3 4 −4]. Geometric multiplicity is 1.

d) [1 0 −1]. Geometric multiplicity is 1.

e) [−1/3 1 0]. There are many answers here, and they will satisfy $v_2 = 1$ and $v_1+v_3 = -1/3$.

3.6.19: $\begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 1 \\ 0 & 0 & -2 \end{bmatrix}$

3.7.3: a) $\begin{bmatrix} 3 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \\ 3 \\ -1 \end{bmatrix}$ b) $\begin{bmatrix} -1 \\ -1 \\ 0 \end{bmatrix}$ c) $\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$ d) $\begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$

3.7.7: a) 3 b) 2 c) 3 d) 2 e) 3

3.7.9: a) $\begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ b) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & -1 & 0 \end{bmatrix}$ c) $\begin{bmatrix} 5/2 & 1 & -3 \\ -1 & -1/2 & 3/2 \\ -1 & 0 & 1 \end{bmatrix}$

3.7.11: a) $\begin{bmatrix} -1 \\ 3 \end{bmatrix}$ b) $\begin{bmatrix} -3 \\ 1 \end{bmatrix}$

3.7.12: (i) Trace is 1, determinant is -2. Eigenvalues are -1 and 2.
(ii) Trace is -2, determinant is 10. Eigenvalues are $-1 \pm 3i$.
(iii) Trace is -2, determinant is -8. Eigenvalues are -4 and 2.
(iv) Trace is -8, determinant is 16. Eigenvalue is -4 repeated.

3.7.13: (i) Trace is 6, determinant is 6. Eigenvalues are 1, 2, and 3.
(ii) Trace is -9, determinant is -39. Eigenvalues are -3 and $-3 \pm 2i$.
(iii) Trace is 1, determinant is -24. Eigenvalues are 2, 3, -4.

4.1.6: $y_1 = C_1 e^{3x}$, $y_2 = y(x) = C_2 e^x + \frac{C_1}{2} e^{3x}$, $y_3 = y(x) = C_3 e^x + \frac{C_1}{2} e^{3x}$

4.1.7: $x = \frac{5}{3} e^{2t} - \frac{2}{3} e^{-t}$, $y = \frac{5}{3} e^{2t} + \frac{4}{3} e^{-t}$

4.1.10: $x_1' = x_2$, $x_2' = x_3$, $x_3' = x_1 + t$

4.1.11: $y_3' + y_1 + y_2 = t$, $y_4' + y_1 - y_2 = t^2$, $y_1' = y_3$, $y_2' = y_4$

4.1.17: $x_1 = x_2 = at$. Explanation of the intuition is left to reader.

4.1.19: a) Left to reader. b) $x_1' = \frac{r}{V}(x_2 - x_1)$, $x_2' = \frac{r}{V} x_1 - \frac{r-s}{V} x_2$. c) As t goes to infinity, both x_1 and x_2 go to zero, explanation is left to reader.

4.1.20: a) (i), b) (iii), c) (ii) Justification left to reader.

4.1.21: a) (iii), b) (ii), c) (i) Justification left to reader.

4.2.7: -15

4.2.11: -2

4.2.13: $x_1 = 3$, $x_2 = 4$, $x_3 = -3$.

4.2.14: Infinitely many solutions of the form $x_1 = \frac{19}{15} + \frac{2}{15} t$, $x_2 = \frac{7}{15} t - \frac{46}{15}$, $x_3 = t$ for any real number t.

4.2.15: No solution.

4.2.16: $x_1 = -2$, $x_2 = 1$, $x_3 = -4$.

4.2.19: $\vec{x} = \left[\begin{smallmatrix} 15 \\ -5 \end{smallmatrix}\right]$

4.2.22: a) $\left[\begin{smallmatrix} 1/a & 0 \\ 0 & 1/b \end{smallmatrix}\right]$ b) $\left[\begin{smallmatrix} 1/a & 0 & 0 \\ 0 & 1/b & 0 \\ 0 & 0 & 1/c \end{smallmatrix}\right]$

4.2.24: $\lambda_1 = 1$, $\vec{v}_1 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$, $\lambda_2 = 2$, $\vec{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

4.2.25: $\lambda_1 = -2 + 2i$, $\vec{v}_1 = \begin{bmatrix} -3 + i \\ 4 \end{bmatrix}$, $\lambda_2 = -2 - 2i$, $\vec{v}_2 = \begin{bmatrix} -3 - i \\ 4 \end{bmatrix}$.

4.2.26: $\lambda_1 = 4$, $\vec{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$, $\lambda_2 = -2$, $\vec{v}_2 = \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix}$, $\lambda_3 = -3$, $\vec{v}_3 = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$.

4.3.2: $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]' = \left[\begin{smallmatrix} 3 & -1 \\ t & 0 \end{smallmatrix}\right] \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] + \left[\begin{smallmatrix} e^t \\ 0 \end{smallmatrix}\right]$

4.3.6: Yes.

4.3.8: No. $2 \left[\begin{smallmatrix} \cosh(t) \\ 1 \end{smallmatrix}\right] - \left[\begin{smallmatrix} e^t \\ 1 \end{smallmatrix}\right] - \left[\begin{smallmatrix} e^{-t} \\ 1 \end{smallmatrix}\right] = \vec{0}$

4.3.11: a) $\vec{x}\,' = \left[\begin{smallmatrix} 0 & 2t \\ 0 & 2t \end{smallmatrix}\right] \vec{x}$ b) $\vec{x} = \left[\begin{smallmatrix} C_2 e^{t^2} + C_1 \\ C_2 e^{t^2} \end{smallmatrix}\right]$

4.4.4: $\vec{x} = C_1 \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right] e^t + C_2 \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right] e^{-t}$

4.4.10: a) Eigenvalues: $4, 0, -1$ Eigenvectors: $\left[\begin{smallmatrix} 1 \\ 0 \\ 1 \end{smallmatrix}\right]$, $\left[\begin{smallmatrix} 0 \\ 1 \\ 0 \end{smallmatrix}\right]$, $\left[\begin{smallmatrix} 3 \\ 5 \\ -2 \end{smallmatrix}\right]$

b) $\vec{x} = C_1 \left[\begin{smallmatrix} 1 \\ 0 \\ 1 \end{smallmatrix}\right] e^{4t} + C_2 \left[\begin{smallmatrix} 0 \\ 1 \\ 0 \end{smallmatrix}\right] + C_3 \left[\begin{smallmatrix} 3 \\ 5 \\ -2 \end{smallmatrix}\right] e^{-t}$

4.4.12: $\vec{x}(t) = C_1 \begin{bmatrix} 1 \\ 3 \end{bmatrix} e^{-4t} + C_2 \begin{bmatrix} 1 \\ 4 \end{bmatrix} e^{-3t}$

4.4.13: $\vec{x}(t) = C_1 \begin{bmatrix} -4 \\ 3 \end{bmatrix} e^{-4t} + C_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-t}$

4.4.14: $\vec{x}(t) = C_1 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} e^{3t} + C_2 \begin{bmatrix} -1 \\ 1 \\ -3 \end{bmatrix} e^{4t} + C_3 \begin{bmatrix} 3 \\ -2 \\ -2 \end{bmatrix} e^{2t}$

4.4.15: $\vec{x}(t) = C_1 \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} e^{-4t} + C_2 \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} e^{-t} + C_3 \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} e^{-2t}$

4.5.4: $\vec{x} = C_1 \begin{bmatrix} \cos(t) \\ -\sin(t) \end{bmatrix} + C_2 \begin{bmatrix} \sin(t) \\ \cos(t) \end{bmatrix}$

4.5.6: a) Eigenvalues: $\frac{1+\sqrt{3}i}{2}, \frac{1-\sqrt{3}i}{2}$, Eigenvectors: $\begin{bmatrix} -2 \\ 1-\sqrt{3}i \end{bmatrix}, \begin{bmatrix} -2 \\ 1+\sqrt{3}i \end{bmatrix}$

b) $\vec{x} = C_1 e^{t/2} \begin{bmatrix} -2\cos\left(\frac{\sqrt{3}t}{2}\right) \\ \cos\left(\frac{\sqrt{3}t}{2}\right) + \sqrt{3}\sin\left(\frac{\sqrt{3}t}{2}\right) \end{bmatrix} + C_2 e^{t/2} \begin{bmatrix} -2\sin\left(\frac{\sqrt{3}t}{2}\right) \\ \sin\left(\frac{\sqrt{3}t}{2}\right) - \sqrt{3}\cos\left(\frac{\sqrt{3}t}{2}\right) \end{bmatrix}$

4.6.11: a) $3, 0, 0$ b) No defects. c) $\vec{x} = C_1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} e^{3t} + C_2 \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} + C_3 \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$

4.6.13: a) $1, 1, 2$

b) Eigenvalue 1 has a defect of 1

c) $\vec{x} = C_1 \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} e^{t} + C_2 \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \right) e^{t} + C_3 \begin{bmatrix} 3 \\ 3 \\ -2 \end{bmatrix} e^{2t}$

4.6.15: a) $2, 2, 2$

b) Eigenvalue 2 has a defect of 2

c) $\vec{x} = C_1 \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix} e^{2t} + C_2 \left(\begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix} \right) e^{2t} + C_3 \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} + \frac{t^2}{2} \begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix} \right) e^{2t}$

4.6.19: $A = \begin{bmatrix} 5 & 5 \\ 0 & 5 \end{bmatrix}$

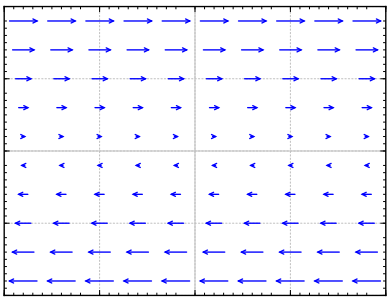
4.6.20: (a) Nodal Source, (b) Saddle, (c) Spiral Sink, (d) Center, (e) Spiral Source, (f) Saddle.

4.7.5: a) Two eigenvalues: $\pm\sqrt{2}$ so the behavior is a saddle. b) Two eigenvalues: 1 and 2, so the behavior is a source. c) Two eigenvalues: $\pm 2i$, so the behavior is a center (ellipses). d) Two eigenvalues: -1 and -2, so the behavior is a sink. e) Two eigenvalues: 5 and -3, so the behavior is a saddle.

4.7.7: Spiral source.

4.7.8: a) Nodal source c) Spiral source c) Saddle c) Nodal sink e) Spiral sink f) Improper nodal sink

4.7.13:



The solution does not move anywhere if $y = 0$. When y is positive, the solution moves (with constant speed) in the positive x direction. When y is negative, the solution moves (with constant speed) in the negative x direction. It is not one of the behaviors we have seen.

Note that the matrix has a double eigenvalue 0 and the general solution is $x = C_1 t + C_2$ and $y = C_1$, which agrees with the description above.

4.7.14: (i) $T = -6$, $D = 8$. Nodal sink. All points nearby are nodal sinks.
(ii) $T = 2$, $D = -3$. Saddle. All points nearby are saddles.
(iii) $T = 0$, $D = 1$. Center. Points nearby are all spirals, but they could be asymptotically stable, centers, or unstable. Stability is unknown.
(iv) $T = 6$, $D = 10$. Spiral source. All points nearby are spiral sources.
(v) $T = -8$, $D = 16$. Improper nodal sink. All points nearby will be asymptotically stable, but they could be nodal sinks, improper nodal sinks, or spiral sinks.
(vi) $T = 4$, $D = 3$. Nodal source. All points nearby are nodal sources.
(vii) $T = -4$, $D = 13$. Spiral sink. All points nearby are spiral sinks.
(viii) $T = 2$, $D = 1$. Improper nodal source. All points nearby will be unstable, but they may be spirals, nodal sources, or improper nodal sources.

4.8.6: The general solution is (particular solutions should agree with one of these):
$x(t) = C_1 e^{9t} + 4C_2 e^{4t} - t/3 - 5/54, \quad y(t) = C_1 e^{9t} - C_2 e^{4t} + t/6 + 7/216$

4.8.8: The general solution is (particular solutions should agree with one of these):
$x(t) = C_1 e^t + C_2 e^{-t} + te^t, \quad y(t) = C_1 e^t - C_2 e^{-t} + te^t$

4.8.10: $\vec{x} = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right] \left(\frac{5}{2}e^t - t - 1\right) + \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right] \frac{-1}{2} e^{-t}$

4.9.4: $\vec{x} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \left(a_1 \cos(\sqrt{3}\,t) + b_1 \sin(\sqrt{3}\,t)\right) + \begin{bmatrix} 0 \\ 1 \\ -2 \end{bmatrix} \left(a_2 \cos(\sqrt{2}\,t) + b_2 \sin(\sqrt{2}\,t)\right) +$
$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \left(a_3 \cos(t) + b_3 \sin(t)\right) + \begin{bmatrix} -1 \\ 1/2 \\ 2/3 \end{bmatrix} \cos(2t)$

4.9.8: $\begin{bmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & m \end{bmatrix} \vec{x}\,'' = \begin{bmatrix} -k & k & 0 \\ k & -2k & k \\ 0 & k & -k \end{bmatrix} \vec{x}$. Solution: $\vec{x} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \left(a_1 \cos(\sqrt{3k/m}\,t) + b_1 \sin(\sqrt{3k/m}\,t)\right)$
$+ \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \left(a_2 \cos(\sqrt{k/m}\,t) + b_2 \sin(\sqrt{k/m}\,t)\right) + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \left(a_3 t + b_3\right)$.

4.9.9: $x_2 = (2/5)\cos(\sqrt{1/6}\,t) - (2/5)\cos(t)$

4.9.12: $\vec{x} = \left[\begin{smallmatrix} 1 \\ 9 \end{smallmatrix}\right] \left(\left(\frac{1}{140} + \frac{1}{120\sqrt{6}}\right) e^{\sqrt{6}t} + \left(\frac{1}{140} + \frac{1}{120\sqrt{6}}\right) e^{-\sqrt{6}t} - \frac{t}{60} - \frac{\cos(t)}{70}\right)$
$+ \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right] \left(\frac{-9}{80}\sin(2t) + \frac{1}{30}\cos(2t) + \frac{9t}{40} - \frac{\cos(t)}{30}\right)$

4.10.4: $e^{tA} = \begin{bmatrix} \frac{e^{3t}+e^{-t}}{2} & \frac{e^{-t}-e^{3t}}{2} \\ \frac{e^{-t}-e^{3t}}{2} & \frac{e^{3t}+e^{-t}}{2} \end{bmatrix}$

4.10.5: $e^{tA} = \begin{bmatrix} 2e^{3t}-4e^{2t}+3e^t & \frac{3e^t}{2}-\frac{3e^{3t}}{2} & -e^{3t}+4e^{2t}-3e^t \\ 2e^t-2e^{2t} & e^t & 2e^{2t}-2e^t \\ 2e^{3t}-5e^{2t}+3e^t & \frac{3e^t}{2}-\frac{3e^{3t}}{2} & -e^{3t}+5e^{2t}-3e^t \end{bmatrix}$

4.10.6: a) $e^{tA} = \begin{bmatrix} (t+1)\,e^{2t} & -te^{2t} \\ te^{2t} & (1-t)\,e^{2t} \end{bmatrix}$ b) $\vec{x} = \begin{bmatrix} (1-t)\,e^{2t} \\ (2-t)\,e^{2t} \end{bmatrix}$

4.10.15: $\begin{bmatrix} 1+2t+5t^2 & 3t+6t^2 \\ 2t+4t^2 & 1+2t+5t^2 \end{bmatrix} \quad e^{0.1A} \approx \left[\begin{smallmatrix} 1.25 & 0.36 \\ 0.24 & 1.25 \end{smallmatrix}\right]$

4.10.17: a) $\begin{bmatrix} 5(3^n) - 2^{n+2} & 4(3^n) - 2^{n+2} \\ 5(2^n) - 5(3^n) & 5(2^n) - 4(3^n) \end{bmatrix}$ b) $\begin{bmatrix} 3 - 2(3^n) & 2(3^n) - 2 \\ 3 - 3^{n+1} & 3^{n+1} - 2 \end{bmatrix}$
c) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ if n is even, and $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ if n is odd.

5.1.3: (i) is c), (ii) is a), (iii) is b)

5.1.5: a) Critical points $(0,0)$ and $(0,1)$. At $(0,0)$ using $u = x$, $v = y$ the linearization is $u' = -2u - (1/\pi)v$, $v' = -v$. At $(0,1)$ using $u = x$, $v = y - 1$ the linearization is $u' = -2u + (1/\pi)v$, $v' = v$.
b) Critical point $(0,0)$. Using $u = x$, $v = y$ the linearization is $u' = u + v$, $v' = u$.
c) Critical point $(1/2, -1/4)$. Using $u = x - 1/2$, $v = y + 1/4$ the linearization is $u' = -u + v$, $v' = u + v$.

5.1.11: Critical points are $(0,0,0)$, and $(-1,1,-1)$. The linearization at the origin using variables $u = x$, $v = y$, $w = z$ is $u' = u$, $v' = -v$, $z' = w$. The linearization at the point $(-1,1,-1)$ using variables $u = x + 1$, $v = y - 1$, $w = z + 1$ is $u' = u - 2w$, $v' = -v - 2w$, $w' = w - 2u$.

5.1.12: $u' = f(u,v,w)$, $v' = g(u,v,w)$, $w' = 1$.

5.1.14: a) $(0,0)$: saddle (unstable), $(1,0)$: source (unstable), b) $(0,0)$: spiral sink (asymptotically stable), $(0,1)$: saddle (unstable), c) $(1,0)$: saddle (unstable), $(0,1)$: saddle (unstable)

5.1.21: A critical point x_0 is stable if $f'(x_0) < 0$ and unstable when $f'(x_0) > 0$.

5.2.2: a) $\frac{1}{2}y^2 + \frac{1}{3}x^3 - 4x = C$, critical points: $(-2,0)$, an unstable saddle, and $(2,0)$, a stable center. b) $\frac{1}{2}y^2 + e^x = C$, no critical points. c) $\frac{1}{2}y^2 + xe^x = C$, critical point at $(-1,0)$ is a stable center.

5.2.3: Critical point at $(0,0)$. Trajectories are $y = \pm\sqrt{2C - (1/2)x^4}$, for $C > 0$, these give closed curves around the origin, so the critical point is a stable center.

5.3.2: a) Critical points are $\omega = 0$, $\theta = k\pi$ for any integer k. When k is odd, we have a saddle point. When k is even we get a sink. b) The findings mean the pendulum will simply go to one of the sinks, for example $(0,0)$ and it will not swing back and forth. The friction is too high for it to oscillate, just like an overdamped mass-spring system.

5.3.4: a) Solving for the critical points we get $(0, -h/d)$ and $(\frac{bh+ad}{ac}, \frac{a}{b})$. The Jacobian matrix at $(0, -h/d)$ is $\begin{bmatrix} a+bh/d & 0 \\ -ch/d & -d \end{bmatrix}$ whose eigenvalues are $a + bh/d$ and $-d$. So the eigenvalues are always real of opposite signs and we get a saddle (In the application however we are only looking at the positive quadrant so this critical point is not relevant). At $(\frac{bh+ad}{ac}, \frac{a}{b})$ we get Jacobian matrix $\begin{bmatrix} 0 & -\frac{b(bh+ad)}{ac} \\ \frac{ac}{b} & \frac{bh+ad}{a} - d \end{bmatrix}$. b) For the specific numbers given, the second critical point is $(\frac{550}{3}, 40)$ the matrix is $\begin{bmatrix} 0 & -11/6 \\ 3/25 & 1/4 \end{bmatrix}$, which has eigenvalues $\frac{5 \pm i\sqrt{327}}{40}$. Therefore there is a spiral source. This means the solution spirals outwards. The solution will eventually hit one of the axes, $x = 0$ or $y = 0$, so something will die out in the forest.

5.3.5: The critical points are on the line $x = 0$. In the positive quadrant the y' is always positive and so the fox population always grows. The constant of motion is $C = y^a e^{-cx-by}$, for any C this curve must hit the y-axis (why?), so the trajectory will simply approach a point on the y axis somewhere and the number of hares will go to zero.

5.4.3: $(0,0)$, unstable, $r = \sqrt{3}$, asymptotically stable.

5.4.4: $(0,0)$, asymptotically stable, $r = \sqrt{2}$, unstable, $r = 2$, asymptotically stable.

5.4.7: Use Bendixson–Dulac Theorem. a) $f_x + g_y = 1 + 1 > 0$, so no closed trajectories. b) $f_x + g_y = -\sin^2(y) + 0 < 0$ for all x, y except the lines given by $y = k\pi$ (where we get zero), so no closed trajectories. c) $f_x + g_y = y + 0 > 0$ for all x, y except the line given by $y = 0$ (where we get zero), so no closed trajectories.

5.4.8: Using Poincaré–Bendixson Theorem, the system has a limit cycle, which is the unit circle centered at the origin as $x = \cos(t) + e^{-t}$, $y = \sin(t) + e^{-t}$ gets closer and closer to the unit circle. Thus we also have that $x = \cos(t)$, $y = \sin(t)$ is the periodic solution.

5.4.12: $f(x, y) = y$, $g(x, y) = \mu(1 - x^2)y - x$. So $f_x + g_y = \mu(1 - x^2)$. The Bendixson–Dulac Theorem says there is no closed trajectory lying entirely in the set $x^2 < 1$.

5.4.14: The closed trajectories are those where $\sin(r) = 0$, therefore, all the circles centered at the origin with radius that is a multiple of π are closed trajectories.

5.5.1: Critical points: $(0, 0, 0)$, $(3\sqrt{8}, 3\sqrt{8}, 27)$, $(-3\sqrt{8}, -3\sqrt{8}, 27)$. Linearization at $(0, 0, 0)$ using $u = x$, $v = y$, $w = z$ is $u' = -10u + 10v$, $v' = 28u - v$, $w' = -(8/3)w$. Linearization at $(3\sqrt{8}, 3\sqrt{8}, 27)$ using $u = x - 3\sqrt{8}$, $v = y - 3\sqrt{8}$, $w = z - 27$ is $u' = -10u + 10v$, $v' = u - v - 3\sqrt{8}w$, $w' = 3\sqrt{8}u + 3\sqrt{8}v - (8/3)w$. Linearization at $(-3\sqrt{8}, -3\sqrt{8}, 27)$ using $u = x + 3\sqrt{8}$, $v = y + 3\sqrt{8}$, $w = z - 27$ is $u' = -10u + 10v$, $v' = u - v + 3\sqrt{8}w$, $w' = -3\sqrt{8}u - 3\sqrt{8}v - (8/3)w$.